International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 01 June 2014, Page No. 1- 4
ISSN: 2278-2419

Scaling Down Dimensions and Feature Extraction in Document Repository Classification

Asha Kurian[1], M.S.Josephine[2], V.Jeyabalaraja[3]
[1]Research Scholar, Department of Computer Applications, Dr. M.G.R. Educational and Research Institute University, Chennai
[2] Professor, Department of Computer Applications, Dr. M.G.R. Educational and Research Institute University, Chennai
[3]Professor, Department of computer Science Engineering, Velammal Engineering College, Chennai
E-mail: ashk47@yahoo.com , josejbr@yahoo.com, jeyabalaraja@gmail.com

Abstract-In this study a comprehensive evaluation of two supervised feature selection methods for dimensionality reduction is performed - Latent Semantic Indexing (LSI) and Principal Component Analysis (PCA). This is gauged against unsupervised techniques like fuzzy feature clustering using hard fuzzy C-means (FCM) . The main objective of the study is to estimate the relative efficiency of two supervised techniques against unsupervised fuzzy techniques while reducing the feature space. It is found that clustering using FCM leads to better accuracy in classifying documents in the face of evolutionary algorithms like LSI and PCA. Results show that the clustering of features improves the accuracy of document classification.

Keywords: Feature extraction, text classification, categorization, dimensionality reduction

I. INTRODUCTION

Text Categorization or text classification attempts to assort documents in a repository into different class labels.A classifier learns from a training set of documents that are already classified and labeled. A general model is devised that correctly labels further incoming documents.A repository typically consists of thousands of documents. Retrieving a selection becomes a laborious task, unless the documents are indexed or categorized in some particular order. Document categorization is modeled along the lines of Information Retrieval [1] and Natural Language Processing [5] where a user query elicits documents of maximal significance in relation to the query. The sorting is done by grouping the document terms and phrases and identifying some association or correlation between them. Establishing relations among words is compounded due to polysemy and the presence of synonyms. Every document contains thousands of unique terms resulting in a highly dimensional feature space.

To reduce information retrieval time, the dimensionality of the document collection can be reduced by selecting only those terms which best describe the document. Dimensionality reduction techniques try to find out the context-meaning of words, disregarding those which are inconsequential. Feature selection algorithms reduce the feature space by selecting appropriate vectors, whereas feature extraction algorithms transform the vectors into a sub-space of scaled down dimension. Feature selection can be followed by supervised or unsupervised learning. Classification becomes supervised when a collection of labeled documents helps in the learning using a train-test set.

II. REVIEW OF LITERATURE

1. Text mining and Organization in Large Corpus, December 2005. Two dimensionality deduction methods: Singular Vector Decomposition (SVD) and Random Projection (RP) are compared, along with three selected clustering algorithms: K-means, Non-negative Matrix Factorization (NMF) and Frequent Itemset. These selected methods and algorithms are compared based on their performance and time consumption.

2. Improving Methods for Single-label Text Categorization, July 2007 – An evaluation of the evolutionary feature reduction algorithms is done. In this paper a comprehensive comparison of the performance of a number of text categorization methods in two different data sets is presented. In particular, the Vector and Latent Semantic Analysis (LSA) methods, a classifier based on Support Vector Machines (SVM) and the k-Nearest Neighbor variations of the Vector and LSA models is evaluated.

3. A Fuzzy based approach to text mining and document clustering, October 2013- In this paper, how to apply fuzzy logic in text mining in order to perform document clustering is shown. Fuzzy c-means (FCM) algorithm was used to cluster these documents into clusters.

4. Feature Clustering algorithms for text classification-Novel Techniques and Reviews, August 2010 - In this paper, some of the important techniques for text classification have been reviewed and novel parameters using fuzzy set approach have been discussed in detail.

5. Classification of text using fuzzy based incremental feature clustering algorithm, International Journal of Advanced Research in Computer Engineering and Technology Volume 1, Issue 5, July 2012 – A fuzzy based incremental feature clustering algorithm is proposed. Based on the similarity test the feature vector of a document set is classified and grouped into clusters following clustering properties and each cluster is characterized by a membership function with statistical mean and deviation.

III. FEATURE SELECTION METHODS FOR DIMENSIONALITY REDUCTION

Feature selection can either be supervised or unsupervised. A brief summary of the different feature extraction methods used

in this study are entailed. As a first step, document pre-processing removes stopwords, short words, numbers and alphanumeric characters. Noise removed, the text is metamorphosed into a term-weighted matrix whose rows indicate the terms and columns represent the documents they make an appearance in. Each cell entry holds the rate of occurrence of words in the document (also called weights). The term weighting factors are:
Local term frequency(tf) – frequency of a term in the document global document frequency(idf) – accounts for the proportionality of the term within a document as well as globally for the whole collection.The normalized tf / idf factor is the general weighting method used. The matrix thus obtained consists mostly of sparse elements.

A. Latent Semantic Indexing (LSI)

Every document has an underlying semantic structure relating to a particular abstraction. Latent Semantic Indexing taps this theory to identify the relation between words and the context in which they used. Mathematical and statistical techniques are used for this inference [2]. Dimensions that are of no consequence to the text are to be eliminated. At the same time removal of these dimensions should not result in a loss of interpretation's starts by pre-processing the document of stop words, stemming etc. This followed by converting the text document to term weighted matrix subsisting of mostly sparse vectors. Cell entries are incremented corresponding to the frequency of words. LSI, accompanied by the powerful Singular Value Decomposition (SVD) is used for feature selection. Singular Value Decomposition (SVD) works with the sparse matrix of term-weights and transforms it into a product of three matrices. The Left Singular vector comprising of original row elements, Right Singular Vector consisting of original column elements both transformed orthogonally and the Singular value which is a diagonal matrix containing the scaling value [4]. The diagonal matrix is the component used for dimensionality reduction. The smallest value in this matrix indicates terms which are inconsequential and can be removed. LSI along with SVD identifies features that do not contribute to the semantic structure of the document.

B. Principal Component Analysis (PCA)

Principal Component Analysis uses the notion of eigenvalues and eigenvectors for feature variable reduction procedure. Given data with high dimensionality, PCA starts by subtracting the mean from each data point so that the averaged mean for each dimension is zero. The original data is projected along with the altered data. A covariance matrix is a square matrixC [n x n] = (cov ($dimension_x$,$dimension_y$)) C is a n x n matrix and each cell (x,y) shows the covariance between two different dimensions x and y.The covariance matrix represents the dependence of dimensions on each other. Positive values indicate that as one dimension increase, the dependent dimension also scales up. The eigenvalues and eigenvectors representation of the covariance matrix is used to plot the principal component axis, the best line along which the data can be laid out. Arranging the eigenvalues and eigenvectors in a decreasing order shows the components with higher relevance corresponding to higher eigenvalues [8]. The vectors with lower values can be disregarded as non significant. The higher second principal component plotted is perpendicular to the first principal component. All eigenvectors are orthogonal to each other irrespective of how many dimensions are present. Each principal component is analyzed to what extent they contribute to the variance of the data points[9] . As a final step the transpose of the chosen principal components is multiplied with the mean adjusted data to derive the final data set representing those dimensions to be retained.

IV. FUZZY CLUSTERING

Fuzzy C-Means algorithm is an iterative algorithm to group the data into different clusters [10]. Every cluster has a cluster centre. The data points are put into the clusters based on the probability with which they belong to the group. Points closer to the center are more closely integrated to the group when compared to those that are further away from the cluster center. A decision as to which cluster to put a data point into is taken based on the membership function of mean and standard deviation. A membership function determines how well the data fit into a particular cluster. On every iteration, the FCM algorithm upgrades the cluster centers and the membership functions. At the same time minimization of the objective function is also done.

The objective function of FCM is given by

$U_m = \sum_1^n \sum_1^c xij \, \|xi - cj\|^2$

Where *m* is the number of iterations

x*ij* denotes the membership of x*i* in the j^{th} cluster.

x*i* are the different dimensions of the data.

The iteration stops when the algorithm correctly pinpoints the cluster center and no change seen in the minimization. The algorithm consists of four steps .

Step1. Identify the initial cluster centers and a matrix X such that each x_{ij} element of X takes a value in [0,1] that denotes the extent to which the element belongs to the cluster.

Step 2.During each iteration update the matrix value [0,1] from the given cluster center.

Step 3.Calculate the objective function for that iteration.

Step 4. If the value has reduced from the previous iteration, continue with the iteration otherwise the procedure halts [11] and [12]. Similar features are now clustered, with features closer to the center having a stronger resemblance.

V. METHODOLOGY

The feature extraction methods try to interpret the underlying semantics of the text and identify the unique words that can be eliminated. The accuracy of classification depends on this optimal set. To assess the efficiency and performance of feature reduction, we first find the accuracy of classification, given the training and test dataset [7]. The evolutionary algorithms like LSI and PCA were mainly chosen because they exhibit good accuracy when reducing dimensions of a document. This is used as a measure find how much it scores against FCM clustering. Supervised feature reduction works on

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 01 June 2014, Page No. 1- 4
ISSN: 2278-2419

the principle that the number of clusters is determined beforehand. The fuzzy C-Means clustering requires no training or test samples to work on. The primary step in feature reduction is to preprocess a document. Pre-processing is an essential procedure that further reduces the complexity of dimensionality reduction and the subsequent classification or clustering process. The whole mass of text has to be transformed algebraically. The text data in the document is mapped to its vector representation, for the purpose of document indexing. The most common form of representation is a matrix format where words that appear frequently are weighted [2].The resultant matrix is a sparse matrix with most of elements being zeros. From the nature of text data, it can be seen that certain words do not contribute to the meaning of the text. The parser removes these words by stopword elimination. This is followed by lemmatization or stemming. Preprocessing also helps to eliminate words shorter than three letters, alphanumeric characters and numbers. With the weighted information, documents are ranked based on their similarity to the query. The cosine of the angle formed by these vectors is used as a similarity measure. Two common datasets have been used in this study. 20 News groups, R8 - the 8 most frequent classes from the Reuters-21578 collection datasets. The dataset is split as training and test documents. Out of the 18821 documents in 20NG, 60% are taken as a training set and the rest for testing. From the R8 group 70% documents are considered as training set. Using the training set for learning, the supervised techniques attempts to assign the new incoming, and unknown documents into their correct class labels [8].

VI. SUPERVISED VS UNSUPERVISED FEATURE SELECTION

All tests have been performed on the MATLAB computing environment. Before any analysis is undertaken, it is worthwhile to study the characteristics of the data collection used. R8, the eight most frequent classes taken from the Reuters collection is smaller in size compared to the 20Newsgroup collection. The general procedure followed is to train the classifier using the training data. The accuracy is judged by the number of test documents that are labeled correctly after the learning phase. In FCM since learning is evolving over iterations, bifurcation of data is unnecessary. In this study the techniques are assessed based on the accuracy, macro-averaged precision and recall, training and testing times [6]. The datasets are divided in a 60:40 and 70:30 ratios for R8 and 20NG as training and test documents. With the class labeling information from the training documents, the test documents can be classified. Initially the datasets are labeled by k-means after feature selection. This is used as a point of reference when clustering using Fuzzy C-Means algorithm is applied. A detailed comparison of the results shows that when features are clustered using FCM, it executes faster and is more accurate in classifying datasets. The savings are mainly due to foregoing the training and testing times.

Dataset	Collection	Classes	Train Docs	Test Docs	Total Docs
20NG	20Newsgroups	20	11293	7528	18821
R8	Reuters-21578	8	5485	2189	7674

Figure.1 shows the datasets used in the study. The number of classes in each collection along with the division of training and test documents is shown

VII. IMPLEMENTATION

All the statistics shown have been derived after data pre-processing followed by reduction of dimensions. The resultant dataset is further classified using the k-means classification method. The performance of the dimensionality reduction algorithms are graded by comparing the evaluation measures of recall, accuracy of classification and precision. Figure.2 and 3 depicts the result of classification using both the supervised and unsupervised techniques on the datasets under study.

	Recall	Accuracy	Precision
PCA	76.27	81.76	88.35
LSI	83.5	89.70	87.64
FCM	85.12	92.37	91.75

Figure.2 Collation of the performance measures for R8 dataset

FCM executes the best in terms of accuracy, precision and recall while classifying the R8 dataset. There is an approximate increase in accuracy of 4% when compared to LSI and 8% increment over PCA. FCM clusters 20Newsgroups with an accuracy of 3% over LSI and close to 10% over PCA.

	Recall	Accuracy	Precision
PCA	75.37	73.29	77.64
LSI	81.44	88.87	84.79
FCM	84.25	92.63	87.60

Figure.3 Collation of the performance measures of the 20Newsgroups data collection.

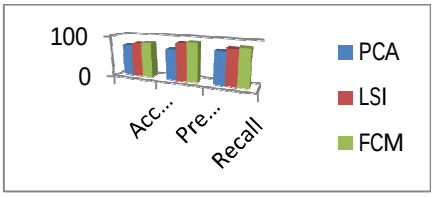

Figure.4 Bar chart showing the comparison of performance measures of the algorithms on R8 dataset.

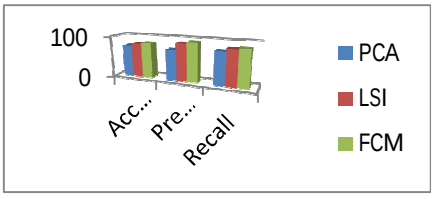

Figure.5 Bar chart showing the comparison of performance measures for 20Newsgroups dataset.

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 01 June 2014, Page No. 1- 4
ISSN: 2278-2419

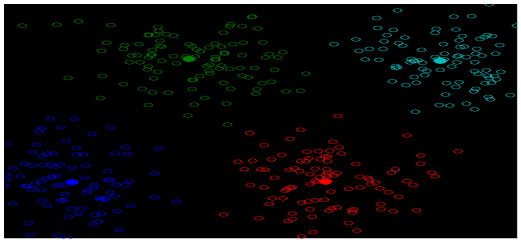

Figure.6 Clustering of R8 dataset using FCM algorithm

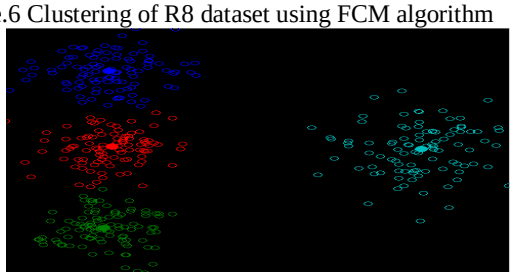

Figure.7 Clustering of 20NG dataset using FCM algorithm

Figures. 6 and 7 shows the clustering of data points using hard c-means clustering. Every point can be a member of only one cluster. Dual clustering is not allowed.FCM has clustered the datasets into four clusters. Cluster centers are marked by slightly larger darker circles. In R8 collection the distribution of terms are not very strongly bound to its cluster center. Features belonging to one cluster could share membership with another cluster as well. In the 20NG dataset, the members are closely tied to its cluster center.

VIII. CONCLUSION

This study evaluated the effectiveness of feature selection techniques on text categorization for dimensionality reduction. Both supervised and supervised techniques were experimented on. From the evolutionary techniques, Latent Semantic Indexing exhibits superior performance over reduction using Principal Component Analysis in terms of precision, accuracy and recall. While using unsupervised feature clustering using FCM, it is seen that there is an improvement over both LSI as well as PCA in its accuracy. Clustering with FCM shows higher accuracy as the training and testing times are minimized. In FCM it can be seen that at least 80% terms can be removed without degrading the resulting classification. The datasets under scanner show some fundamental differences. Classification results vary depending upon the features that are eliminated, the relation between the data, what factors are used to assess the similarity in collaboration with the classification method employed.Given a choice of optimally proven feature extraction methods, it is possible that the accuracy will improve, when two feature selection algorithms are used in conjunction with each other. Since the characteristics of every document collection is different, devising a classification algorithm that works best depending on the type of document content could also be introduced. The performance of the feature extraction algorithms can also be estimated using other efficacy measures.

REFERENCES

[1] Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley, Reading, Massachusetts, USA, 1999.

[2] Wei Xu, Xin Liu, Yihong Gong. Document Clustering Based On Nonnegative Matrix Factorization. In ACM. SIGIR, Toronto, Canada, 2003.

[3] Ian Soboroff. IR Models: The Vector Space Model. In Information Retrieval Lecture 7.

[4] http://www.csee.umbc.edu/_ian/irF02/lectures/07Models-VSM.pdf

[5] 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.

[6] Marko Grobelnik, Dunja Mladenic and J. Stefan. Institute, Slovenia Text-Mining Tutorial

[7] Patterns in Unstructured Data Discovery, Aggregation, and Visualization - A Presentation to the Andrew W. Mellon Foundation by Clara Yu,John Cuadrado ,Maciej Ceglowski, J. Scott Payne

[8] Yang, Y., and Jan Pedersen. "A Comparative Study on Feature Selection in Text Categorization." ICML 1997: 412-420.

[9] A tutorial on Principal Components Analysis Lindsay I Smith.

[10] Ana Cardoso-Cachopo, Improving Methods for Single-label Text Categorization, PhD Thesis, October, 2007.

[11] K.Sathiyakumari, V.Preamsudha, G.Manimekalai; "Unsupervised Approach for Document Clustering Using Modified Fuzzy C mean Algorithm"; International Journal of Computer & Organization Trends – Volume 11 Issue3-2011.

[12] R. Rajendra Prasath, Sudeshna Sarkar: Unsupervised Feature Generation using Knowledge Repositories for Effective Text Categorization. ECAI 2010: 1101-1102

[13] Nogueira ,T,M ;" On The Use of Fuzzy Rules to Text Document Classification "; 2010 10th International Conference on Hybrid Intelligent Systems (HIS),; 23-25 Aug 2010 Atlanta, US

Author Profile

Asha Kurian is a research scholar in the department of computer application from Dr. MGR University, Chennai. She did her Post Graduation in computer Application from Coimbatore Institute of Technology, Bharathiar University in 2003 . Her areas of interest include Data Mining and Artifical Intelligence.

M.S.Josephine is Working in Dept of Computer Applications, Dr.MGR University, Chennai. She graduated Masters Degree (MCA) from St.Joseph's College, Bharathidasan University, M.Phil (Computer Science) from Periyar University and Doctorate from Mother Teresa University in Computer Applications. Her research Interest includes Software Engineering , Expert System, Networks and Data Mining.