International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 24 (2015) pp 44504-44508

An Efficient Hybrid Information Retrieval Approach for Unstructured Document Classification

R. Umamaheswari[1],
Asso. Prof/CSE, Gnanamani College of Technology, Namakkal, T. N, INDIA. E-mail:umait1978@gmail.com

Dr. N. Shanthi[2]
Professor & Dean/CSE Nandha Engineering College, Erode, T. N, INDIA, E-Mail:shanthimoorthi@yahoo.com

ABSTRACT
Lot of analysis is going on for extracting information from unstructured data. Most of the web documents are not having proper structure. So information retrieval (IR) has lot of challenges while extracting information from the web. Clustering and classification are the most widely used techniques to group the documents in an efficient manner. Neural network pattern recognition tool has been recognized as efficient tool to construct the patterns for large datasets. K-Means algorithm is used to group the documents based on similarity measures. By default, K-means uses Euclidian distance for calculating the similarity. In this paper both the K-Means and neural network pattern recognition algorithm is combined to improve the efficiency of clustering. Different performance measures are used to measure classifier accuracy. This hybrid approach achieves good result for unstructured documents.

Keywords: Text Mining, K-Means Clustering, Neural Network, Similarity Measure.

1. INTRODUCTION

Web mining is the process of analyzing the data in order to bring useful patterns from the unstructured web data. Clustering and Classification play an important role in text mining. Guadalupe et al. [2], reflected in their research that the digital documents have been increasing dramatically over the years. The problems are information management, searching and retrieval of relevant documents, etc. It is necessary to develop the methods to organize large amount of web data into smaller number of meaningful clustering which will help to solve most of the problems. In [9], K-Means clustering algorithm produces good performance in many applications. K-Means uses Euclidian distance for clustering. By optimizing similarity measures the optimal clusters can be formed and thus performance is improved. Another important quality of K-Means algorithm is that it can be easily combined with other algorithms for the best results. Generally the problem of clustering can be thought as optimization process.

Neural network is used to classify inputs into set of target categories. The neural network pattern recognition tool helps us to select data, create and train a network and evaluate the performance using cross entropy and confusion matrices. In this pattern recognition networks are feed forward networks that can be trained to classify inputs according to target classes. Pratiksha Y et al. introduce hybrid approaches in order to produce good results. They combined decision tree with neural networks. They showed the improved accuracy in text classification task compared with previous results. So in our approach we combined the k-means with neural network pattern recognition to develop the efficient clustering method.

2. RELATED WORK

Document clustering or Text classification uses similarity measures. A similarity measure is a method which computes the degree of similarity between a pair of text objects. In [7] Nithya. P et al. studied various similarity measures. There is a large number of similarity calculations proposed in the related work, the best similarity measure is Similarity Measure for Text Processing (SMTP). The experimental result shows good performance for SMTP. The main problem is selecting feature for calculating similarity scores. The existing similarity and distance measures performance in text document clustering is still not clear. In [4], more number of similarity measures shown various methods and effectiveness in clustering text and attracted considerable research interests recently.

Clustering algorithm aims to find a natural structure or relationship in an unlabeled datasets [2]. They focused the performance of K-means algorithm with different initializations and more iteration. The experiment dataset used are IRIS and Portuguese. The main problem is applying clustering algorithms in real time applications and selecting efficient clustering algorithm. In [3], the performance of spherical K-Means algorithm uses cosine similarity algorithm for clustering.

Anna Huang [4] shows the performance by using K-means clustering algorithm for unordered text documents. They compared the performance results for seven datasets. The main problem is enhanced document representation. Now a days, many researchers using poor document representation. Documents are lacking in semantic representation. If the researchers use proper document representation the result will be improved. G. Saltonet et al. [5] proposed variety of similarity or distance measures such as cosine similarity, Jaccard correlation co-efficient. Similarity is mostly worked in terms of dissimilarity or distance.

Anil Kumar Pantidaret et al. [6] used Shared Nearest Neighbour (SNN) to produce small clusters and mainly focus on multidimensional databases. They showed the results based on Euclidean distance, Jaccard co-efficient, cosine and Pearson correlation function.

3. PROPOSED METHODOLOGY

The proposed approach is to classify the text documents by using hybrid method. Classification uses combination of K-means algorithm and Neural Network Pattern recognition.

The additional tool Neural Network Pattern Recognition tool is used to classify inputs into a set of target categories. A two-layer feed-forward network, with sigmoid hidden and output neurons, can classify vectors arbitrarily well, and 10 hidden neurons were used for classification process. Another important process is the neural network is trained with scaled conjugate gradient back propagation.
The proposed architecture as follow:

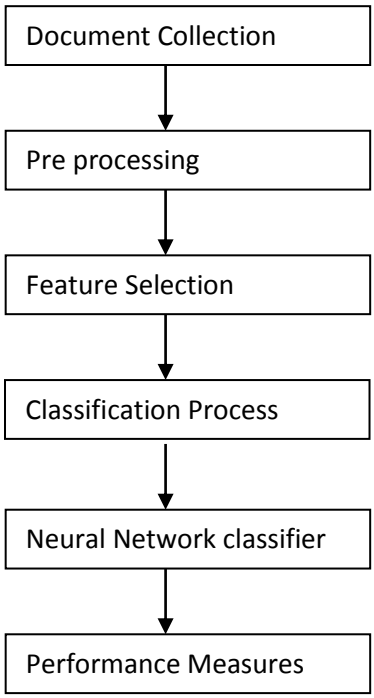

Figure 3. 1: System Methodology

i. Document Collection
In this proposed work, datasets are collected from UCI Repository. 10Newsgroup dataset is used for text classification process.

ii. Pre Processing
The main idea behind preprocessing is to get the proper data for analysis and also produces good performance.
The dataset contains the documents about news related information. There are ten categories of documents available. WEKA tool is used here for preprocessing.

Preprocessing Steps:
Step 1: Combine multiple text document files into singlearff file using WEKA tool.
Step 2: Covert the contents into tokens.
Step 3: Apply the string to word vector function to create the Term Vector Matrix.
Step 4: Remove the stop words.
Step 5: Apply the stemming algorithm that converts different words into similar canonical form.
Step 6: Covert the Term Vector Matrix into equivalent Matrix which is used as input for MATLAB for further analysis.

Document representation is very important in Text Mining because most of the document is in unstructured format. Machine Learning algorithm produced good results based on the proper document representation. The preprocessing is used to represent the documents in convenient way.
Another important process of preprocessing is to remove the stop words and stemming process.
Stop words-Stop words are useless in Information Retrieval (IR) and Text Mining process. These are the words frequently occur gives no meaning. The top words like is, are, the, they etc., are removed.
Stemming-The process of finding root or stem of a word. This will help to reduce the words to their stems. For Example the words connection, connects, connected, and connecting all can be stemmed to the word 'connect'. In this proposed work the predefined iterated Lovins stemmer algorithm is used.

Document Similarity
The document similarity is finds the similarity between documents by using various methods. In K-Means algorithm there are various measures used to find the distance between the documents. They are
Squared Euclidean distance
Cityblock
Cosine
Correlation
Hamming
Different method produces different performances.
In this proposed work the Squared Euclidean distance and Cosine method is used.

Document Representation
Document representation can be done in following ways.
1. Document Similarity with (Binary weighting)
2. Document Similarity with term frequency

1. Document Similarity with (Binary weighting)
This is the basic method which uses binary weighting to represent the document. Here the weights tij=1 if document i contains term j and zero, otherwise.

2. Document Similarity with term frequency
Term Frequency-Inverse Document Frequency (TF-IDF)
This method mainly used to captures the relevancy among the words. TF is defined as number of times a term occurs in a document. IDF is used to measure the importance of a term in text document collection.

$$tf(t,d) = \frac{f(t,d)}{maximum\ number\ of\ occurances\ of\ word}$$

$$idf(t,d) = \log(\frac{|D|}{number\ of\ document\ term\ t\ appears})$$

$$tf - idf(t,d,D) = tf(t,d) * idf(t,d)$$

iii. Feature Selection
The main use of feature selection is to select the subset of features from the original documents. Here documents are unstructured and all the words are considered as important and

all the tokens are arranged in ascending order. High dimensionality feature space decreases the efficiency and accuracy of the classifier. In this work feature selection is done through stop word removal and stemming process.

iv. Classification Process

In experiment the use of K-Means clustering algorithm results less computation and produce good results for large datasets [8][9][10].

Algorithm:

K-Means. The K-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

K: the number of clusters,
D: a data set containing n objects.
Output : A set of k clusters.

Method:

Arbitrarily chose k objects from D as the initial cluster centers;

1. **Repeat**
2. (re) assign each object to the cluster t which the object is the most similar, based on the mean value of the objects in the cluster;
3. Update the cluster means, that is calculate the mean value of the objects for each cluster;
4. **Until** no change;

Apply the k-means algorithm to cluster the input document with default random selection of centroid. The output of this algorithm is given as input to the neural network pattern recognition.

v. Neural Network Classifier

In this proposed work the output of k-means algorithm is fed into neural network. Based on the number of classes the target classes were created and assigned to the neural network classifier. It trains the input and also evaluates the performance of the clusters. By using this pattern it can able to classify the large datasets.
Creating target class algorithm K-Means output:

Algorithm Target Class:

Input: classarray
Output: targetarray

1. Initialize the targetarray(rowsize, colsize)
2. For i=1 :rowsize
3. Compare the classarray with class value and assign the corresponding classvalue to target array.
4. End For
5. End

The retrain network operation is repeated till minimizing Cross-Entropy results and gives good classification. Lower values represent better results and zero means no error.

vi. Performance Measures

The performance measures used here is confusion matrix, Receiver Operator Characteristic (ROC) and performance curve from Neural Network Pattern Recognition.
Confusion matrix and ROC are interrelated. In [12] clearly given the measure details like, given a classifier and an instance, there are four possible outcomes. If the instance is positive and it is classified as positive, it is counted as a true positive; if it is classified as negative, it is counted as a false negative. If the instance is negative and it is classified as negative, it is counted as a true negative; if it is classified as positive, it is counted as a false positive.

	True Class	
Hypothesized Class	True Positives	False Positives
	False Negatives	True Negatives
Column Total	P	N

Figure 3. 1 Confusion matrix format

$$False\ Positive\ Rate = \frac{FP}{N}$$

$$True\ Positive\ Rate = \frac{TP}{P}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{P}$$

$$Accuracy = \frac{TP + TN}{P + N}$$

The above measures used to evaluate the classifier model.

4. EXPERIMENTAL RESULTS

The datasets used in this paper are downloaded from UCI Repository. The datasets are unstructured, the large variation of data and that are structured to analysis to get optimal results. This repository contains twenty newsgroups dataset for text analysis. In this paper ten newsgroups dataset are taken for the initial analysis with k-means and neural network classification. In this paper the Experiments are performed on ten newsgroups, they are Computer graphics, IBM PC Hardware, Autos, Baseball, Hockey, Electronics, Medical, Space, Politics and Religion.

Data	Data Set for Experiments	Classes	Number of Documents	Number of Terms
20 Newsgroup	10 Newsgroup	Computer Graphics	486	20782
		IBM PC Hardware	491	32307
		Autos	495	24806
		Baseball	497	27271
		Hockey	499	29181
		Electronics	490	27943
		Medical	495	25622
		Space	493	33058
		Politics	387	23024
		Religion	314	26515

This proposed work classification problems involving constructing best model by using training, validation and test set. The targets can consist of ten clusters. The input is given to Neural Network Pattern Classifier the various output can be analyzed for Euclidean and Cosine distance.

Result for Binary weighting

Table 4. 1. Classifier accuracy

Algorithm	Neural network pattern Recognition Hidden neuron =10		
KMeans-Distance	Iteration 1	Iteration 2	Iteration 3
Euclidean distance	93. 8	90	93
Cosine distance	66. 2	96	93. 7

The performance of the proposed hybrid approach is analyzed in various ways. The performance curve, confusion matrix and ROC (Receiver Operator Characteristics) curves are used for analysis.

Table 4. 1 represents the classifier accuracy for various iterations. The best model can be used for identify the new data. In figure 4. 1 shows the performance curve which has minimum cross entropy error. It shows the result of various good classifier accuracy for hybrid approach.

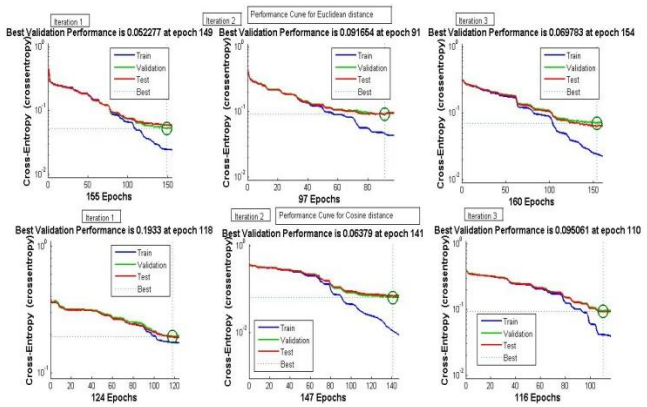

Figure 4. 1: Best Validation Performance for Euclidean distance and Cosine distance function

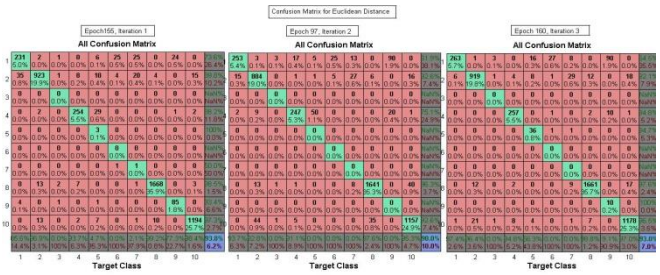

Figure 4. 2: Confusion Matrix for Euclidean Distance for three iterations

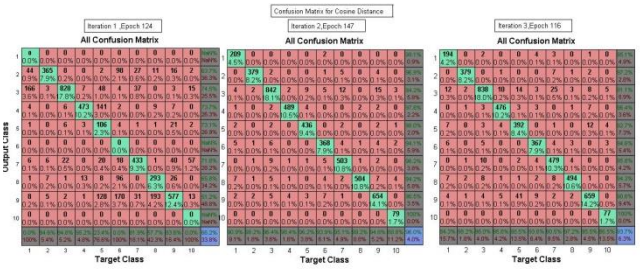

Figure 4. 3: Confusion Matrix for Cosine Distance for three iterations

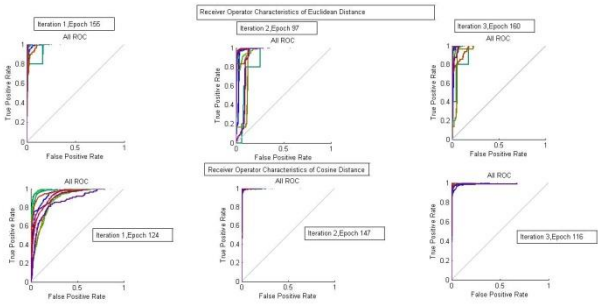

Figure 4. 4: ROC Curve for Euclidean Distance and Cosine Distance

4. 1 Result Discussion

The above figures show the results taken from various iterations of K-means algorithm and best epoch for Neural Network classifier. The best three iteration results are taken for analysis. In Figure 4. 1 shows the performance curve for training data, validation and test data. Here overall dataset is divided as three parts. First 70% is training data, second 15% is for validation and last 15% is for test data. Total instances are 4647, In that 3253 samples are used for training, 697 samples are used for validation, and 697 samples are used for test.

The Figure 4. 2 and 4. 3 shows the confusion matrix. Classifier performance showed for 10 clusters based on Euclidean and Cosine distance measures.

The green boxes represents instances are correctly classified and red boxes represents incorrectly classified instances. In figure 4. 4 shows the Receiver Operator Characteristic (ROC)

curves. ROC curve is one of the visualization tool for classification. ROC curves visualize true positive and false positive rates which also can be taken out of a confusion matrix. The steeper the curve (towards the upper left corner) is the better the classification.

5. CONCLUSION AND FUTURE WORK

Still lots of research are going on large amount of text documents retrieval. So many research proved the hybrid method provides good results in classification and clustering. Our approach produces good results for text classification with good pattern. We can apply this method for Information Retrieval (IR) and Searching to extract relevant information. The present research has two limitations. First, K-means algorithm selects default centroid for classification. Second, Semantic can be added for more relevant searching. Our future work concentrates on both of these limitations.

References

[1]. Pratiksha Y. Pawar and S. H. Gawande, Member, IACSIT, "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.

[2]. Guadalupe J. Torres, Ram B. Basnet, Andrew H. Sung, SrinivasMukkamala, Bernardete M. Ribeiro, "A Similarity Measure for Clustering and Its Applications", Manuscript submitted May 18, 2008. This work was supported in part by ICASA (Institute for Complex Additive Systems Analysis), a division of New Mexico Tech.

[3]. Sentence Similarity Measures For Essay Coherence Derrick Higgins, Jill Burstein, 2007

[4]. Anna Huang, Department of Computer Science, The University of Waikato, Hamilton, New Zealand, " Similarity Measures for Text Document Clustering", NZCSRSC 2008, April 2008, Christchurch, New Zealand.

[5]. Salton G. Automatic Text Processing. Addison-Wesley, New York, 1989.

[6]. Anil Kumar Patidar, Jitendra Agrawal, Nishchol Mishra, School of IT, Rajiv Gandhi, "Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach" International Journal of Computer Applications (0975-8887) Volume 40-No. 16, February 2012.

[7]. Nithya P., Umamaheswari R., Shanthi Dr. N., "An Enhanced Similarity Computation for Document Clustering Approaches "International Journal of Science and Engineering Research (IJSER), Vol 2 Issue 10 October-2014, ISSN-3221 5687, (P) 3221 568X.

[8]. Nithya P., Umamaheswari P., Shanthi Dr. N., "A Data Mining Objective Function with Future Selection Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 2, No. 10, pp. 1155-1158, 2015.

[9]. Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In In KDD Workshop on Text Mining, 2000.

[10]. Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '99, pages 16-22, New York, NY, USA, 1999. ACM.

[11]. Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '92, pages 318-329, New York, NY, USA, 1992. ACM.

[12]. Tom Fawcett, Institute for the Study of Learning and Expertise, "An introduction to ROC analysis", available on line at Pattern Recognition Letters 27 (2006) 861-874.