Clustering Based Feature Selection Algorithm

Mr. Chandrashekhar W Gandhare
B.E (Computer Technology)
RGCER,Nagpur

Mr. Gaurav V Wanjare
B.E (Computer Technology)
RGCER,Nagpur

Mr.Sandip P Wandhare
B.E (Computer Technology)
RGCER,Nagpur

Mr. Mohit M Deshpande
B.E(Computer Technology)
RGCER,Nagpur

***Abstract* - Feature selection is vital in the field of pattern classification due to accuracy and processing time considerations. The selection of proper features is of greater importance when the initial feature set is considerably large. Text classification is unsupervised machine learning method. It needs representation of objects and similarity measure, which compares distribution of features between objects. In this paper, we describe the hybrid method used for text clustering which is the combination of active feature selection, genetic algorithm and bisecting K-means. Internal quality measures compute the effectiveness of clustering. Our method is compared with K-means.**

Keywords: summarization, unsupervised, similarity Measures, classifier.

1. Introduction

The purpose of text classification is to use the contents of a text or document to assign it to one or more categories. It has applications in document organization and management, information retrieval, and certain machine learning algorithms. Common general techniques for text classification include both unsupervised and supervised pattern classification methods. Some common approaches use clustering instead of simple feature selection, linear discriminant methods, neural networks and support vector machines Feature selection forms an important subset within the much larger area of text classification. Correctly identifying the relevant features in a text is of vital importance to the task of text classification.

Internet contains large amount of unclassified data. The unstructured texts contains massive amount of information which cannot be used for further processing by computers.The purpose of feature selection is to determine which features are the most relevant to the current classification task. In text classification, features are typically words from a document. Choosing an appropriate feature selection method for text classification can be vital because of the large number of features usually present in text documents to extract useful patterns from documents from processing methods and algorithms need to be used. Text document clustering groups similar documents that to form a similar type of cluster, while documents that are different have separated apart into different type of clusters. Clustering is used in information retrieval and information extraction, by grouping similar types of information sources together.

Feature selection (FS) is a commonly used step in machine learning, especially when dealing with a high dimensional space of features. The main purpose of FS is to choose a subset of features from the original set of features forming patterns in a given dataset. Feature selection is extensive and it spreads throughout many fields, including machine learning, pattern recognition, text categorization, data mining , and signal processing .NLP organizes the text documents into meaningful cluster according to their content and to visualize the collection, providing an overview of the range of documents and of their relationships, so that they can be browsed more easily. Natural language processing approaches can be applied both to feature extraction and feature reduction phases of the text classification process.

Natural language processing can be used in ways that encompass both feature extraction and reduction, tools can be used to identify keywords from a text document or even create a semi structured summary of the text. Many clustering techniques have been proposed in the literature. Clustering algorithms are mainly categorized into Hierarchical and Partitioning methods. Hierarchical clustering method works by grouping data objects into a tree of clusters .

2. PROPOSED WORK

BASIC STEPS

Step 1: **Feature Extraction**

- Natural Language Processing (NLP) is used for extracting features like Parts of Speech (POS) from text document.

- Feature Extraction from Single Text Document Using NLP.
- Feature Extraction from Multiple Text Documents Using NLP.

Step 2: **Feature Filtering /Selection**

- Active feature selection *partitions the data and actively chooses* useful instances for feature selection.
- Genetic algorithm (GA) optimizes features to implement global searching.
- Good Individual (fitness is high)□ selection operator of an initial value. Bi-Sec ting k -means algorithm chooses operator .Selection, crossover & mutation, on the feature space search → optimal solution.

Step 3: **Classification Algorithm**

- Uses (clustering) k-means algorithm as selection operation to control the scope of the search and we have classified the documents into clusters.

Step 4: **Evaluating the Performance**

- Performance of result is compared with existing algorithm on the basis of parameters like precisions, recall and F-Measure.

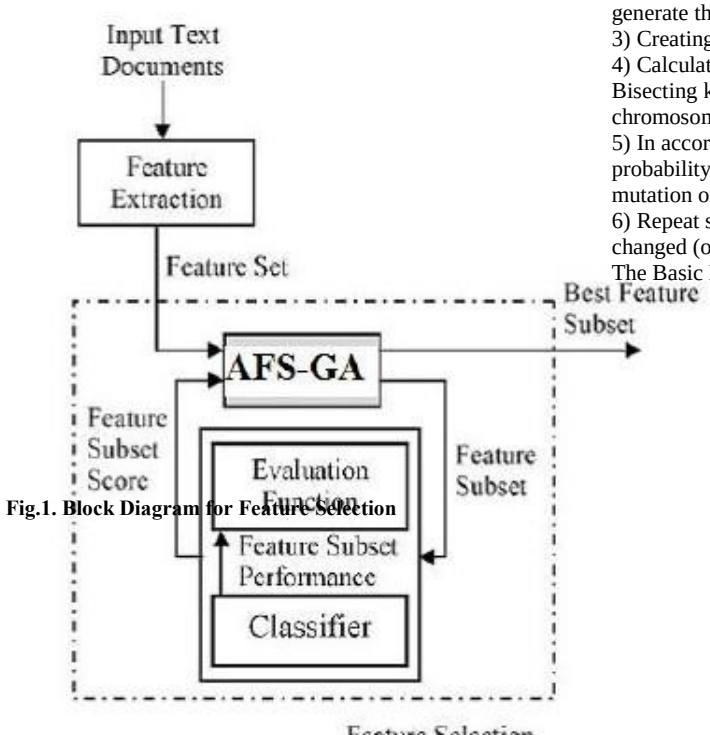

Fig.1. Block Diagram for Feature Selection

After preprocessing of text documents ,feature extraction is used to transform the input text documents into a feature set(feature vector).Feature Selection is applied to the feature set to reduce the dimensionality of it. We apply feature selection methods to text clustering task to improve the clustering performance. IN this project, we explore the possibility of active feature selection that can influence which instances are used for feature selection by exploiting some characteristics of the data. Our objective is to actively select instances with higher probabilities to be informative in determining feature relevance so as to improve the performance of feature selection without increasing the number of sampled instances. Active sampling used in active feature selection chooses instances in two steps: first, it partitions the data according to some homogeneity criterion; and second, it randomly selects instances from these partitions. In this project, we are applying a combination of NLP, Active feature selection and unsupervised method GA along with clustering thus we would get a better output for text classification with respect to the methods available. We will perform a comparative study on a variety of feature selection methods for text clustering, with other algorithms. Finally, we evaluate the performance of hybrid feature selection (AFS) method based on clustering. Nature of similarity measure plays a very important role in the success or failure of a clustering method. An important step in any clustering is to select a distance measure, which will determine how the similarity of two elements.

After extracting features, feature selection is performed and features are classified into clusters based on the similarity. The specific steps of feature selection and clustering are as follows:

Feature selection optimization based on combination GA and bisecting k-means using genetic algorithm to implement global searching, and using bisecting k-means algorithm is used as operator.

The specific algorithm is described as follows:

1) Determine the text features encoding scheme, using binary encoding, the chromosome length is L.

2) Initialization control parameters: N is population size, Pc is crossover probability, P_m is mutation probability, generate the hereditary algebra.

3) Creating an initial population with m individuals.

4) Calculating the fitness of each individual, using Bisecting k-means algorithm to select the parent chromosome of crossover and mutation.

5) In accordance with crossover probability Pc, mutation probability Pm, generate offspring via crossover and mutation on parent population.

6) Repeat steps 4) and 5) , until all the individual no longer changed (or reach the termination conditions).

The Basic Bisecting K-means Algorithm for finding

K clusters is:
1. Pick a leaf cluster C to split.
2. Use the basic K-means algorithm. (Bisecting step) to split into two sub-clusters C_1 & C_2.
3. Repeat the bisecting step and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of Cluster is obtained.

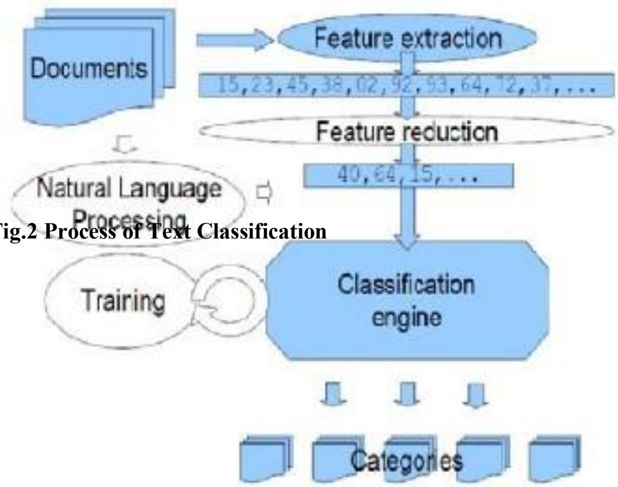

Fig.2 Process of Text Classification

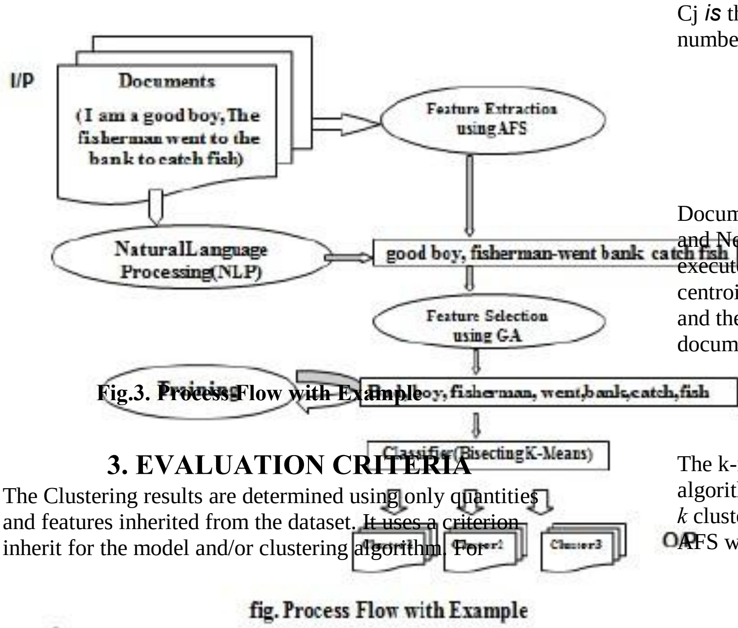

Fig.3. Process Flow with Example

3. EVALUATION CRITERIA

The Clustering results are determined using only quantities and features inherited from the dataset. It uses a criterion inherit for the model and/or clustering algorithm. For instance the average similarity of objects in the same cluster.

The main characteristics of the Improved K-means clustering algorithm can be summarized in two ways. First, compared with previous clustering algorithms in which the task of clustering is done without knowing the semantic nature of each cluster, the K-means clustering recognizes this semantic nature of clusters by using the key phrases extracted from the documents in the cluster. Second, the traditional K-means must specify the number of clusters *k* in advance by the user, which results in the change of clustering results as the value of *k* changes. Improved K-means solves this problem by automatically determining this number. Improved K-means clustering algorithm uses Cosine measure and the Euclidean distance measure to calculate feature values, simultaneously. Hence, the similarity of two documents is computed by the following Expression:

$$\sin(d1, d2) = \frac{\text{cosine}(d1, d2)}{\text{euclidean}(d1, d2)}$$

Also it calculates F-measure using precision and recall which gives better clusters. Precision and recall is calculated using following expression:

Recall (i ,j) = C_{ij} / C_j
Precision (I, j) = C_{ij} / C_i

And F-measure is calculated using following expression:

F(i, j) = $\frac{2 * Recall(i,j) * Precision(i,j)}{Precision(i,j) + Recall(i,j)}$

Where C_{ij} is the number of members of topic i in cluster j, Cj *is* the number of members of cluster j and Ci *is* the number of members of topic i.

4. DATASET

Documents are collected from standard datasets viz. Text and Newsgroup , and the clustering algorithms are executed for this collection 5 times. In each test, *k* centroids are randomly selected from the document space, and the classification accuracy is measured by assigning documents to the correct cluster.

5. RESULT ANALYSIS

The k-means algorithm is most important and very popular algorithm for solving the problem clustering a data set into *k* clusters. First, we compare the clustering accuracy of AFS with k-means, k-means with Active feature selection

methods. In supervised and unsupervised feature selection methods were evaluated in terms of improving the clustering performance by conducting experiments in the case that the class labels of documents are available for the feature selection.

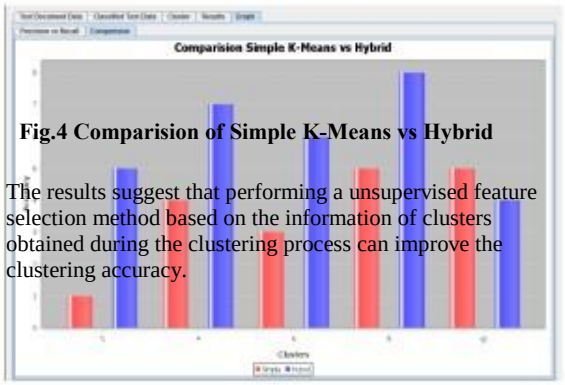

Fig.4 Comparision of Simple K-Means vs Hybrid

The results suggest that performing a unsupervised feature selection method based on the information of clusters obtained during the clustering process can improve the clustering accuracy.

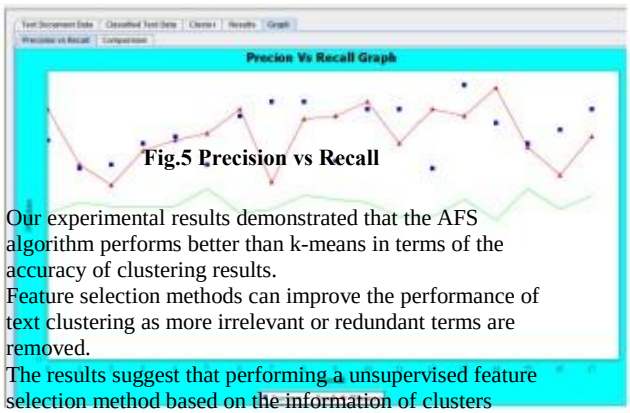

Fig.5 Precision vs Recall

Our experimental results demonstrated that the AFS algorithm performs better than k-means in terms of the accuracy of clustering results.
Feature selection methods can improve the performance of text clustering as more irrelevant or redundant terms are removed.
The results suggest that performing a unsupervised feature selection method based on the information of clusters obtained during the clustering process can improve the clustering accuracy.

6. CONCLUSION

Clustering is one of the most important tasks in the data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. In order to solve the high dimensionality and inherent data sparsity problems of feature space, feature selection methods are used. AFS has been compared with other clustering and feature selection algorithms, such as k means.also this project is useful for Business intelligence I to obtain some guided data mining methods by identifying the related services. The business intelligence data of large enterprises can help to generate a very powerful knowledgebase. In this project, Business intelligence is useful to obtain some guided data mining methods by identifying the related services.The core idea and information behind this architecture is to design a generic system that is flexible enough to suit to different requirement of the business. The business intelligence data of large enterprises can help to generate a very powerful knowledgebase.

REFERENCES;

[1] K.-Y. Whang, J. Jeon, K. Shim, J. Srivatava, Active Feature Selection Using Classes, PAKDD 2003, LNAI 2637, pp. 474–485, 2003.Springer-Verlag Berlin Heidelberg 2003.

[2] J. Li, E. Li, Y. Chen, L. Xu, and Y. Zhang,A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2013 ,pp.1-14.

[3]Li Wei-qiang,Wang Xiao-feng,Improved Method of Feature Selection based on Information Gain,IEEE 2012,pp. 1-4.

[4] Jialei Wang, Peilin Zhao, Steven C.H. Hoi, and Rong Jin, Online Feature Selection and Its Applications, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2013, Issue : 99,pp.1-14.

[5] Jihong Liu, Guoxiong Wang,2010,A Hybrid Feature Selection Method for Data Sets of thousands of Variables,IEEE 2010,pp. 288-291.

[6] Guoxiong Wang, Yafei Wang, Dan Li, A Dynamic Feature Selection Method Based on Combination of GA with K-means, 2010,Proceedings: 2nd International Conference on Industrial Mechatronics and Automation,pp.271-274.

[7]Wei Zhao,Yafei Wang,Dan Li, 2010,A New Feature Selection Algorithm in Text Categorization, Proceedings :International Symposium on Computer, Communication, Control and Automation,pp.146-149.

[8]Zi-jun Yu; Wei-gang Wu; Jing Xiao; Jun Zhang; Rui-Zhang Huang; Ou Liu. 2009. Keyword Combination

Extraction in Text Categorization Based on Ant Colony Optimization. Soft Computing and Pattern Recognition, 2009. SOCPAR '09. Inter- national Conference of, vol., no.,pp.430-435,4-7Dec.2009.DOI: 10.1109/SoCPaR.2009.90
[9]M. F. Zaiyadi and Baharudin. 2010. A proposed hybrid approach for feature selection in text document categorization. World academy of science, engineering and technology 72-2010,pp. 137-141.
[10]Wang Xiaoyue and Bai Rujiang. Applying RDF Ontologies to Im- prove Text Classification. Computational Intelligence and Natural Computing, 2009. CINC '09. International Conference on , vol.2, no., pp.118-121, 6-7 June 2009 doi: 10.1109/CINC.2009.115.