Hybrid Active Feature Selection For Text Classification

Mrs. Rashmi G. Dukhi[1,] Ms. Antara Bhattacharya[2]
[1,2]mtech(Cse),Ghrietw,Nagpur Assistant Professor,Ghrietw,Nagpur

ABSTRACT: Clustering is the most common form of unsupervised learning.In clustering, it is the distribution and makeup of the data that will determine cluster membership. It needs representation of objects and similarity measure. which compares distribution of features between objects. For the high dimensionality, feature extraction and feature selection improves the performance of clustering algorithms.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.External quality measures computes the effectiveness of clustering.Our hybrid method is compared with K-means.

Keywords : summarization; unsupervised; similarity measures; classifier.

I. INTRODUCTION

Internet contains vast amount of unstructured text. The unstructured texts contains massive amount of information which cannot be used for further processing by computers. To extract useful patterns from documents,specific processing methods and algorithms need to be used.Huge information lies in collections of documents in the form of digital libraries and repositories, and digitized personal information such as blog articles and emails. Clustering of text documents plays a vital role in efficient Document Organization, Summarization, Topic Extraction and Information Retrieval. Text document clustering groups similar documents that to form a coherent cluster, while documents that are different have separated apart into different clusters. Clustering is used in information retrieval and information extraction, by grouping similar types of information sources together.

In text categorization problems , feature space is determined by the vocabularies from the natural language documents whose size is commonly of hundreds of thousands of words. Meanwhile the collection of documents available for classification is typically large. For instance, numerous Web pages and online articles are available, but we are only interested in searching for those of a particular topic, which is a very small fraction of the whole. Feature selection that studies how to select informative (or discriminative) features and remove irrelevant, redundant or noisy ones from data, is an important and frequently used technique for data preprocessing in machine learning. By reducing the dimensionality of data, feature selection reduces the overall computational cost, improves the performance of learning algorithms and enhances the comprehensibility of the data models. With the help of feature selection, machine learning algorithms become more scalable, reliable and accurate.

NLP tools extracts novel knowledge out of very large unstructured collections of text documents (text data mining).

NLP organizes the documents into meaningful groups 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. Linguistic features can be extracted from texts and used as part of their feature vectors. Feature extraction from such condensed forms of the original documents reduces the dimensionality of the input vector without reducing the classification performance.

Feature selection is a process that chooses a subset from the original feature set according to some criterions. It improves efficiency, accuracy and comprehensibility of the models built by learning algorithms. Feature selection techniques have been widely employed in a variety of applications, such as genomic analysis, information retrieval, and text categorization The selected feature retains original physical meaning and provides a better understanding for the data and learning process.The process of text classification is shown in figure1.

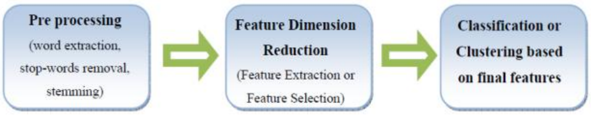
Pre processing
(word extraction, stop-words removal, stemming)
Feature Dimension Reduction
(Feature Extraction or Feature Selection)
Classification or Clustering based on final features

Fig. 1. Steps in Text Classification

The first phase consist of preprocessing the text documents.The second phase of text clustering is feature dimension reduction. Feature extraction and feature selection are two commonly used methods for reducing the dimension of corpus. Feature extraction is the process of extracting new features from the set of all features by means of some functional mapping. Feature selection methods on the other hand select some of the existing terms based on some measures and generate the final feature vector.

Clustering is the most common form of unsupervised learning and this is the major difference between clustering and classification. No super-vision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. Clustering is sometimes erroneously referred to as automatic classification; however, this is inaccurate, since the clusters found are not known prior to processing whereas in case of classification the classes are pre-defined. In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to the classification where the classifier learns the association between objects and classes from a so called training set, i.e. a set of data correctly labeled by hand, and then replicates the learnt behavior on unlabeled data.

Many clustering techniques have been proposed in the literature. Clustering algorithms are mainly categorized into Hierarchical and Partitioning methods [2, 3, 4, 5]. Hierarchical clustering method works by grouping data objects into a tree of clusters [6]. These methods can further be classified into agglomerative and divisive Hierarchical clustering depending on whether the Hierarchical decomposition is formed in a bottom-up or top-down fashion. K-means and its variants [7, 8, 9] are the most well-known partitioning methods [10]. Clustering is a technique which has no predefined class labels but using similarity measures between different objects, it put most similar object in one class and dissimilar in another class. Figure 2 descibes the general steps used in document clustering. Very first words are separated and then weights are applied to each of them.

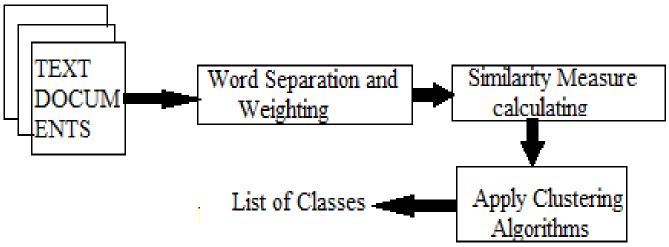
TEXT DOCUMENTS
Word Separation and Weighting
Similarity Measure calculating
Apply Clustering Algorithms
List of Classes

Fig 2. BasicClustering Steps

II. PROPOSED WORK

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(HAFS) 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. A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity, Jaccard coefficient, Euclidean distance and Pearson Correlation Coefficient.

A. Preprocessing

The set of documents is given as input to feature extraction which is preprocessed first.Text Preprocessing aims at transforming the text collection into a useful form for the learning algorithms, involving tasks as treatment, cleaning and reduction of the data.

1) Term Filtering – Remove stop words like the,and,of,a.

2) *Stemming*- Stemming is the process of reducing words to their stem or root form. For example 'cook', 'cooking', 'cooked' are all forms of the same word used in different constraint but for measuring similarity these should be considered same.
3) *Selection of Index Terms*-Term can be individual words or noun phrases.

B. Extraction of feature terms using the NLP

Natural language processing approaches is be applied both to feature extraction and feature reduction phases of the text classification process. Linguistic features can be extracted from texts and used as part of their feature vectors.For example parts of the text that are written in direct speech, use of different types of declinations, length of sentences, proportions of different parts of speech in sentences (such as noun phrases, preposition phrases or verb phrases) can all be detected and used as a feature vector or in addition to word frequency feature vector.

The important features i.e words from the document are extracted using NLP & GA.Parts of Speech(POS) tagging is used for extraction of features.POS tag the document via the standard ngram tagger.It takes a sentence as input,assigns a POS tag to each word in the sentence and produces the tagged text as output.

C. Feature Selection

A typical feature selection procedure (shown in Figure 3 consists of four basic steps:
1) subset generation;
2) subset evaluation;
3) stopping criterion and
4) result validation.

The process begins with subset generation that employs a certain search strategy to produce candidate feature subsets. Then each candidate subset is evaluated according to a certain evaluation criterion and compared with the previous best one. If it is better, then it replaces the previous best. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Finally the selected best feature subset is validated by prior knowledge or some test data. Search strategy and evaluation criterion are two key topics in the study of feature selection.

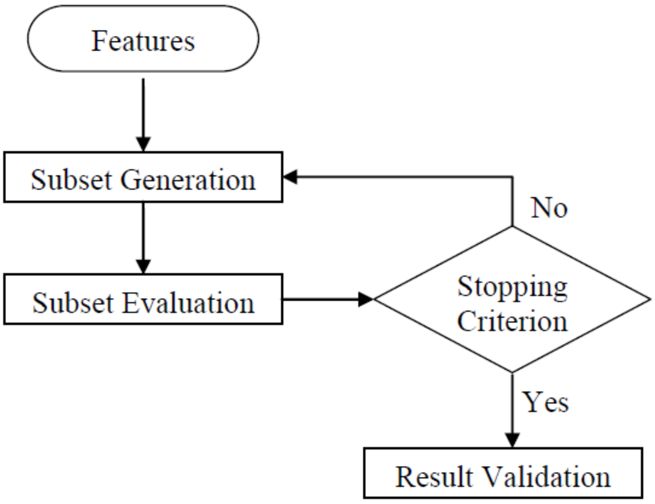

Fig 3. Feature Selection Procedure

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.

III. EVALUATION CRITERIA

One of the most important issues in clusters analysis is the evaluation of the clustering results. Evaluating clustering results is the analysis of the output to understand how well it reproduces the original structure of the data.
The ways of evaluation are divided in two parts:

1. Internal quality measure

2. External Quality measure

In internal quality measures, the overall similarity measure is used based on the pair wise similarity of documents and there is no external knowledge to be used.
For external quality measure some external knowledge for the data is required.

A. F-measure

This is an aggregation of precision and recall, here adopted to clustering evaluation purposes. *Precision* is the ratio of the number of relevant documents to the total number of documents retrieved for a query. *Recall* is the ratio of the number of relevant documents retrieved for a query to the total number of relevant documents in the entire collection. In terms of evaluating clustering, the f-measure of each single cluster c_i is:

$$F(c_i) = \max_{j=1\ldots m} 2 * \frac{P_j R_j}{P_j + R_j}$$

Where $P_j = \frac{| c_i \cap k_j |}{|k_j|}$

and

$$R_j = \frac{| c_i \cap k_j |}{|c_j|}$$

The final F-measure for the entire set is given as :

$$F = \frac{\sum_{i=1}^{m} F(c_i)}{N}$$

where N is the total number of documents.
Higher value of F-measure indicates better clustering.

IV. DATASET

We have used Reuters Transcibed Dataset .This is the most common dataset used for evaluation of document categorization and clustering.

This dataset is created by reading out 200 files from the 10 largest Reuters classes.There are 10 directories labeled by the topic name. Each contains 20 files of transcriptions.

Table 1: Characteristics of Reuter DataSet

DataSets	**No. of Documents**	**No. of classes**	**Class Size**	**Average length of documents**
Reuters Transcription(wheat)	20	02	185	19
Reuters Transcription(Trade)	21	03	195	21
Reuters Transcription(Ship)	20	02	185	19
Reuters Transcription(Money)	20	02	185	19
Reuters Transcription(Grain)	20	02	185	19
Reuters Transcription(Corn)	20	02	185	19

V. RESULT ANALYSIS

The k-means algorithm is very popular for solving the problem clustering a data set into *k* clusters. First, we compare the clustering accuracy of HAFS with k-means, k-means with Active feature selection methods. In [1], 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. As a preprocessing step of text clustering, the HAFS feature selection was reported as the best among the unsupervised feature selection methods evaluated in .

Table 2: FMeasure for various clusters

Sr.No	No.of Cluster(K)	Hybrid	Simple K-Means
1	3	0.72	0.46
2	4	0.74	0.40
3	6	0.64	0.32
4	8	0.78	0.50

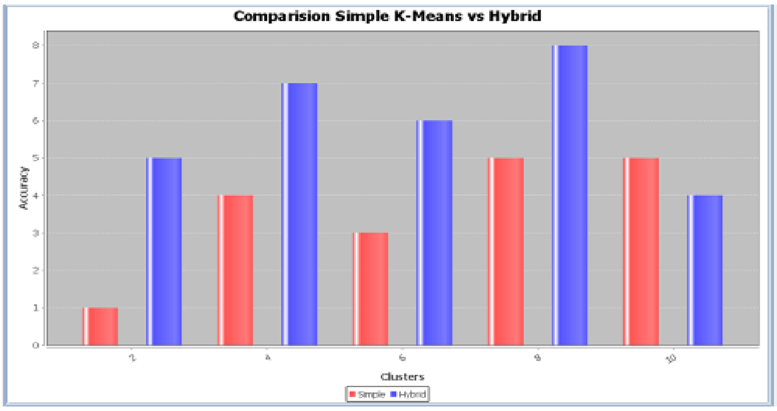

Fig 4. Comparison of hybrid method with K-Means on Reuter Transcribed Dataset

Our experimental results demonstrated that the HAFS 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.
- Result in Table 2 shows that FMeasure of HAFS performs better clustering than that of K-Means which suggest that it is better than regular K-Means.

VI. 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. In real case, the class information is unknown, so only unsupervised feature selection methods can be exploited. We have proposed a new text clustering algorithm HAFS that performs a unsupervised feature selection during the clustering process.The selected features improve the quality of clustering iteratively, and as the clustering process converges, the clustering result has higher accuracy. HAFS has been compared with other clustering and feature selection algorithms, such as k-means. Our experimental results show that HAFS performs better than other algorithms in terms of the clustering accuracy for different test data sets.

REFERENCES

[1]. T. Liu, S. Liu, Z. Chen, and W. Ma, “An Evaluation on Feature Selection for Text Clustering,” Proc. of Int’l Conf. on Machine Learning, 2003.

[2]. G. Kowalski, Information Retrieval Systems – Theory and Implementation, Kluwer Academic Publishers, 1997.

[3]. D.R. Cutting, D.R. Karger, J.O. Pedersen, and J.W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, SIGIR ‘92, Pages 318 – 329, 1992.

[4]. O. Zamir, O. Etzioni, O. Madani, R.M. Karp, Fast and Intuitive Clustering of Web Documents, KDD ’97, Pages 287-290, 1997.

[5]. D. Koller and M. Sahami, Hierarchically classifying documents using very few words, Proceedings of the 14th International Conference on Machine Learning (ML), pp. 170-178, 1997.

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

[7]. M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. In KDD Workshop on Text Mining, 2000.

[8]. D. R. Cutting, J. O. Pedersen, D. R. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, 1992.

[9]. B. Larsen and C. 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, 1999.

[10]. D. Arthur and S. Vassilvitskii. k-means++ the advantages of careful seeding. In Symposium on Discrete Algorithms, 2007.
Reuters-21578 Distribution 1.0, available at

[11]. http://www.daviddlewis.com/resources/testcollections/reuters21578

[12]. F. Sebastiani, “Machine Learning in Automated Text Categorization,”ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[13]. Yanjun Li, Congnan Luo, and Soon M. Chung, “Text Clustering with Feature Selection by Using Statistical Data” , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. XX, NO. YY, 2008,pp.1-11

[14]. N. Sandhya,Y. Sri Lalitha, V.Sowmy, Dr. K. Anuradha, and Dr. A. Govardhan, “Analysis of Stemming Algorithm for Text Clustering”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011 ,pp. 352-359.