DISCOVERY OF NEW CLUSTERS USING INTELLIGENT ADAPTIVE TECHNIQUES

Vennakula I s Saikumar*, V.Suresh**, Girish Reddy Ginni***
2. Assoc Professor, Dept of CSE, Pydah College of Engg & Tech, Vishakapatnam.

Abstract: In this new and current era of technology, advancements and techniques, efficient and effective text document classification is becoming a challenging and highly required area to capably categorize text documents into mutually exclusive categories. Fuzzy similarity provides a way to find the similarity of features among various documents. In this paper, a technical review on various fuzzy similarity based models is given. These models are discussed and compared to frame out their use and necessity. A tour of different methodologies is provided which is based upon fuzzy similarity related concerns. It shows that how text and web documents are categorized efficiently into different categories. Various experimental results of these models are also discussed. The technical comparisons among each model’s parameters are shown in the form of a 3-D chart. Such study and technical review provide a strong base of research work done on fuzzy similarity based text document categorization.

1. INTRODUCTION

Text classification is different from conventional classification approaches in construction of text documents. The dimensionality for text data is very large in comparison to other forms of data sets. Also each document may contain only a few of the features from the entire pool of feature set. Recently, text data processing techniques have attracted more and more attention. The curse of dimensionality creates a great problem to the classification. For example, two realworld data sets, 20 Newsgroups and Reuters 21591 top-10, both have more than 15 thousand features. Such high dimensionality is a severe problem for classification algorithms. To alleviate this problem, feature reduction approaches are applied before document classification tasks are performed. The main purpose of Feature Reduction is to reduce the classifiers computation load and to Increase data consistency.

There are mainly two techniques for feature reduction, those are feature selection and feature extraction. The feature selection methods use techniques like sampling which takes a subset of the features and the classifier only uses the subset instead of all the original features to perform the text classification task. The feature extraction methods convert the representation of the original documents to a new representation based on a smaller set of synthesized features. well-known feature selection approach is based on Information Gain [12] measure defined by the amount of decreased uncertainty given a piece of information. However, there are some problems associated with the feature selection based methods. In these methods, only a subset of the words is used for the classification of text data, therefore useful information may be ignored. The feature extraction techniques are used to translate the representations of the input documents to a new representation based on a smaller set of synthesized features. Feature clustering [1, 3, 2, and 10] is one of powerful techniques for feature extraction. Feature
clustering is nothing but grouping of the words with a high degree of pair wise semantic relatedness into clusters and each word cluster contains the grouped features treated as a single feature. In this way, the dimensionality of the features can be drastically reduced. There are some feature clustering techniques suggested by Baker and McCallum [1] derived from the “distributional clustering” idea of Pereira et al. [4]. An Information Bottleneck approach was proposed by Tishby et al. [2, 7] and showed that feature clustering approaches are more effective than feature selection ones. A Divisive Clustering (DC) method was proposed by Dhillon et al. [3], which is an information and theoretic feature clustering approach and more effective than other feature clustering methods. In these methods, each new feature is generated by combining a subset of the original words. A word is assigned to a word’s group or subset if the similarity of the word to the subset is greater than those to other subsets, despite the distinction is very small. All the feature selection and extraction methods mentioned above require the number of new features be specified in advance by the user. Later Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee propose fuzzy similarity-based self-constructing feature clustering algorithm [10], which is an incremental feature clustering approach to reduce the number of features for the text classification task. Words in the feature vector those are similar to each other are grouped into the same cluster. Each cluster is characterized by a fuzzy membership function with statistical mean and deviation. If a word is not match to any existing cluster, a new cluster is built for this word. The Fuzzy membership with mean and deviation represents the similarity between a word and cluster. When all the words have been fed in, a required number of clusters are formed automatically. Now we take each cluster as a reduced feature. The extracted feature represents to a cluster is a weighted combination of the words contained in the cluster. In this paper, we propose a novel fuzzy similarity-based clustering approach using an efficient split Gaussian fuzzy membership function. Sometimes the data distribution may be skewed so the usage of the exponential approach to data distribution may give more weak results. We are using efficient split Gaussian distribution function as fuzzy membership function. By this algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data. Besides, the user need not specify the number of extracted features in advance, and trial-and-error for determining the appropriate number of extracted features can then be avoided. Experiments on real world data sets show that our method can run faster and obtain better extracted features than other methods.

2. RELATED WORK

Text categorization [1] [2] is an upcoming and vital field in today’s world which is most importantly required and demanded to efficiently categorize various text documents into different categories. Artificial Intelligence [3] – [5] provides many learning methods and paradigms to represent, interpret and acquire domain knowledge to help other documents in learning. Such categorization must produce the accurate and correct results with high performance. Due to the huge data size and complexity, data dimensionality reduction has also been a primary concern. Great levels of efforts have been put in this direction, so that the major problem of curse of dimensionality can be reduced.

Text documents clustering [1] [2] [6] has been paid good attention. Many models and techniques have been developed for clustering. The clustering techniques can be applied to the web documents also. In this way, they can be categorized into their major and respective categories of business, stock, sports, cricket, movie, news and many more. Therefore, the unsupervised learning paradigm [6] is used to make the document clusters. It does not include any prior information and knowledge, that’s why it requires complex text processing techniques. Nowadays, text classification [7] – [8] [9] is gaining more attention and focus for text categorization activities even at the overhead of increased cost. Research is also being done for the fuzzy association, signature, c-means, algorithms and methods for categorization tasks. Text classification with fuzzy logic base provides a better forum to sufficiently categorize the text and web documents. It also results in justified solutions with reduced efforts. When it is combined with the feature clustering technique, it highly improves the representation of features. It further improves the storage performance and decreases the risks of feature ambiguity. Therefore, text classification techniques provide prior information and classification knowledge, so that classifiers can be made learnable to further categorize text and web documents. Many researchers are doing well in this area. Some of the applications in this field are, text classification system SECTCS (Smart English and Chinese Text Classification System) [8],
segmenting handwritten text [9], nonlinear dimensionality reduction techniques [10] [11], complex linguistic features in context - sensitive text classification techniques [7] [12], cyber terrorism investigation [13], spam filtering [10] [11], topic spotting, email routing, language guessing, and many more. Text Classification and clustering are two opposite extremes with regard to the extent of human supervision they require. Real-life applications are considered somewhere in between, because unlabeled data is easy to collect but labelling data is more helpful.

3. TEXT CLASSIFICATION

Given a set D of training documents, text classification can be done as follows: We specify the similarity threshold $\gamma$ for (16), and apply our clustering algorithm. Assume that k clusters are obtained for the words in the feature vector W. Then we find the weighting matrix T and convert D to D0. Using D0 as training data, a classifier based on Support vector machines (SVM) is built. Note that any classifying technique other than SVM can be applied. Joachims proposed that SVM is works well and performs better than other methods for text categorization. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. A support vector machine (SVM) is a technique in computer science for a set of related machine learning methods that are used to process data and recognize the patterns, used for classification. It is known that support vector machines (SVM) are capable of effectively processing feature vectors of some 10 000 dimensions, given that these are sparse. Several authors have shown, that support vector machines provide a fast and effective means for learning text classifiers from examples. Documents of a given topic could be identified with high accuracy).Topic identification with SVM implies a kind of semantic space in the sense that the learned hyper plane separates those documents in the input space, which belong to different topics. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input, making the SVM a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm creates a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

SVM is a kernel method, which finds the maximum margin hyper plane in feature space separating the images of the training patterns into two groups. To make the method more flexible and robust, some patterns need not be correctly classified by the hyper plane, but the misclassified patterns should be penalized. Therefore, slack variables $i$ are introduced to account for misclassifications. The objective function and constraints of the classification

4. FUZZY CLUSTERING

Word patterns are considered one by one. No clusters exist at the beginning, and clusters can be created if necessary. For each word pattern, the similarity of this word pattern to each existing cluster is calculated to decide whether it is combined into an existing cluster or a new cluster is created. When the word pattern is combined into an existing cluster, the membership function of that cluster should be updated accordingly. Feature clustering methods are generally hard clustering where each word of the original features belongs to exactly one word cluster. Fuzzy clustering are defined as soft clustering where each feature has more than one similarity.

4.1. ADVANTAGES OF PROPOSED SYSTEM

1. A fuzzy incremental feature clustering (FIFC) algorithm which is an incremental clustering approach to reduce the dimensionality of the features in text classification.

2. Determine the number of features automatically.

3. Match membership functions closely with the real distribution of the training data.

4. Runs faster than other methods.

5. Better extracted features than other methods.
4.2. THE TEXT CLASSIFICATION FRAMEWORK

![Text Classification Framework Diagram]

Fig.1.Block diagram for Text Classification

4.3. EXPERIMENT AND RESULTS

The proposed algorithm is implemented in Matlab 7.8.0(R2009a) and applied on various data sets. Results of these experiments are summarized in Table 1. Datasets used for training are Reuter-21578 R8, 20 Newsgroup and WebKB data set. All of these data sets are obtainable from http://web.ist.utl.pt/~acardoso/datasets/. The input data set files are pre-processed, which provide require input set for the learning. The input data set files is converted into numeric data on the basis of number of occurrence of words in the documents. Proposed algorithm determines number of iterations in obtaining cluster center. Furthermore we use F-FFC and FFC to represent Fast Fuzzy Feature Clustering and Fuzzy Self Constructing Feature Clustering respectively. Execution time for both F-FFC and FFC is calculated in second.

Table 1. Sample Execution Time (in sec) on Different Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of Instances</th>
<th>No. of Features</th>
<th>No. of Classes</th>
<th>F-FFC</th>
<th>FFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuters-21578 R8</td>
<td>391</td>
<td>3125</td>
<td>8</td>
<td>2.734</td>
<td>88.536</td>
</tr>
<tr>
<td>20 Newsgroups</td>
<td>400</td>
<td>9005</td>
<td>20</td>
<td>30.498</td>
<td>1792.285</td>
</tr>
<tr>
<td>WebKB</td>
<td>400</td>
<td>3491</td>
<td>4</td>
<td>3.110</td>
<td>20.756</td>
</tr>
</tbody>
</table>

5. PERFORMANCE EVALUATION

UCI-Adult by removing tuples with missing values. Thus it contains 30,162 tuples. SFU-Adult is configured by adding 15,060 extra tuples to UCI-Adult, so it has 45,222 tuples. For k-anonymization, our quasi identifier attributes are selected from the following attributes:age, final-weight, education-number, capital-gain, capital-loss, hours-per-week, education, martial-status, occupation, nation. The first six of them are continuous, and the left four are categorical. We adopt from [7] the hierarchies for categorical attributes, and the intervals for continuous attributes.

6. CONCLUSION

In this paper we have presented a novel Fuzzy clustering algorithm for text classification which uses a different fuzzy membership function that solves the problem of skewness. The skewness of data in the distribution gives more weak results which effects the resulting clusters which may affects the performance of the text classification. In this paper we are proposing the split Gaussian distribution function as fuzzy membership function that works well for all types of data. We have presented a fuzzy based incremental feature clustering (FIFC) algorithm, which is an incremental clustering approach to reduce the dimensionality of the features classification of text. Feature that are similar to each other are placed in the same cluster. New clusters formed automatically, if a word is not similar to any existing cluster. Each cluster so formed is characterized by a membership function with statistical mean and deviation. By our work the derived membership functions match closely with the real distribution of the training data. We reduce the burden on the user in specifying the number of extracted features in advance. Experiments results shows that our method can run faster and obtain better extracted features methods.

7. REFERENCES


