Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Fuzzy Based Approach for Multilabel Text Categorization and Similar Document Retrieval

Rubiya P U
M.Tech Student, Computer Science Department, Viswajyothi College, India

Cinita Mary Mathew
Asst. Professor, Computer Science Department, Viswajyothi College, India

Abstract— Text Categorization is a fundamental task in natural language processing aspects such as information extraction, information retrieval, and text mining. Text classification is the process of classifying an incoming stream of documents into predefined categories through the classifiers learned from the training samples. Machine Learning (ML) is the most popular approach used for text classification. Single label classification is concerned with learning from a set of samples that are associated with a single label. Since a text document often belongs to multiple categories in real tasks such as web pages and international patent categorization, text categorization is generally defined as assigning one or more predefined category labels to each data sample. Therefore, developing better classifiers with generalization ability for such multi-label categorization tasks is an important issue in the field of machine learning. Using a fuzzy based approach, multiple category labels can be obtained during this multilabel text categorization. By using these categories, another problem in machine learning, i.e., similar document retrieval can be solved. Similar document detection identifies groups of highly related documents.

Keywords— multilabel text categorization, fuzzy method, Dimensionality reduction, clustering, similar document detection.

I. INTRODUCTION

Nowadays the data stored in various forms is increasing beyond a limit. Hence the need for innovative and effective technologies to help, find and use the useful information and knowledge from a large variety of data sources is continually increasing. Text classification by Machine learning and Similar Document Retrieval are the important technologies for information organization and management. Text categorization is a fundamental task in such aspects of natural language processing such as information extraction, information retrieval, and text mining.

Today, most text categorization is done by people. People save hundreds of files, email messages, and URLs in folders every day. They are often asked to choose keywords from an approved set of indexing terms for describing the technical publications or areas of expertise on program committees. Human categorization is very time-consuming and costly, thus limiting its applicability especially for large or rapidly changing collections. Additional concerns such as the lack of consistency in category assignment and the need to adapt to changing category structures also limit the applicability of purely human systems. The goal of text categorization is the classification of documents into a fixed number of predefined categories. Each document can be in multiple, exactly one, or no category at all. Using machine learning, the objective is to learn classifiers from examples which perform the category assignments automatically. This is a supervised learning problem.

Traditional single label classification is concerned with learning from a set of examples that are associated with only a single label l from a set of disjoint labels L, $|L| > 1$. If $|L| = 2$, then the learning problem is called a binary classification problem (or filtering in the case of textual and web data), while if $|L| > 2$, then it is called a multi class classification problem. Most machine learning algorithms, such as Rocchio's method, k-nearest neighbor classifiers, probabilistic Bayesian models[4], decision trees, decision rules, and support vector machines[3], were designed for single label classification in which a document can only belong to one category.

Since a text document often belongs to multiple categories in real tasks such as web pages and international patent categorization, text categorization is generally defined as assigning one or more predefined category labels to each data sample. Therefore, developing better classifiers with generalization ability for such multi-label categorization tasks is an important issue in the field of machine learning.

For example, a newspaper article that concerns the reactions of the scientific circle to the release of the Da Vinci Code film can be classified to any of the three categories: Arts, Science, and Movies. Another example is in web search, each returned webpage with a given query may be labeled with more than one categories: consider the following webpage http://www.pbs.org/science/science health.html, which may be annotated as "Science", "Health", "Education", and "News and Media", four out of 14 top-level categories used by Yahoo! By using these categories, another problem in machine learning, i.e., similar document retrieval can be solved. Similar document detection identifies groups of highly related documents.

II. RELATED WORKS

The volume of text data, collected from different areas of research as well as over the Internet, has grown to be quite large over the years. Hence it is getting increasingly difficult to analyze these reports using human means. It has been pointed out in text research over a long period of time that classification is an important part of text data analysis. Now, with the increase of volume of text data, it has become necessary that the analysis is done through automated means. Text data sets can be binary, multi class or multi label in nature. For the first two categories, only a single class label can be associated with a document. However, in case of multi label data, more than one class labels can be associated with a document at the same time.

However, even if a data set is multi label, not all combinations of class labels appear in a data set. Also, the probability with which a particular class label combination occurs is also different. It indicates that there is a correlation among the different class labels and it varies across each pair of class labels. On studying the literature, it can be understood that, for multi label classification, most traditional approaches try to transform the multi label problem to multi class or binary class problem. There are also other techniques in the literature for MultiLabel Text Classification.

An enhanced hybrid classification method through the utilization of the naive Bayes approach and the Support Vector Machine (SVM) is implemented in [3]. In this approach, the Bayes formula was used to vectorize a document according to a probability distribution reflecting the probable categories that the document may belong to. Using this probability distribution as the vectors to represent the document, the SVM can then be used to classify the documents on a multidimensional level.

The underlying assumption in traditional machine learning algorithms is that instances are Independent and Identically Distributed (IID). In [4], the focus is on the development and analysis of a supervised machine learning algorithm that does not make the IID assumption, but instead moves beyond instance boundaries to exploit the latent information in higher order co-occurrence paths between features within data sets. The approach used is based on a NB algorithm that is applied to the problem of text classification. NB is the simplest of Bayesian classifiers in that it assumes that all attributes of the examples are independent of each other given the context of the class. Another method is implemented in [5], which uses a fuzzy similarity-based self-constructing algorithm for feature clustering. This is an incremental feature clustering approach to reduce the number of features for the text classification task. An approach based on fuzzy clustering is discussed in [6] to handle high dimensionality of data and using inter class correlation information in the form of class label pairs to enhance the prediction probabilities in multi label classification as a post processing step.

Learning Classifier Systems (LCSs) are rule-based systems with a discovery mechanism to find additional meaningful rules according to the results of its previous experiments. LCSs were designed to deal with both single and multistep problems. A system called Voting Based LCS (VLCS) is implemented in [7]. This method helps to guide the discovery mechanism by a prior knowledge. This prior knowledge is defined as a voting mechanism that realizes the quality of the existing rules and is used in discovering new rules.

Fuzzy rule-based classification systems (FRBCSs) have been widely employed in the field of pattern recognition and classification problems. IVTURS[8] is such a system, which is short for linguistic FRBCS based on an Interval Valued fuzzy reasoning method(IV-FRM) with Tuning and Rule Selection. The main contribution of IVTURS is a novel IV-FRM in which the ignorance represented by the IVFSs is taken into account throughout the reasoning process. Another approach called FSKNN[9], employs fuzzy similarity measure (FSM) and k nearest neighbors (KNN), for multi-label text classification. An efficient fuzzy based method is proposed in [1] which consists of a training phase and testing phase. The trained model obtained in the training phase is used to assign categories to unseen documents in the testing phase.

Among the different machine learning techniques, clustering can interpret the multi labelity of data in a more meaningful way. In fact, the notion of subspace clustering matches that of text data i.e., having high and sparse dimensionality and multi labelity.

III. PROBLEM DEFINITION

Multilabel text categorization task classifies an incoming stream of documents into predefined categories through the classifiers learned from the training samples. In text categorization, the number of the involved features is usually huge. This may cause the curse of the dimensionality problem. The high-dimensional documents appear to be sparse and dissimilar in many ways which prevent common data organization strategies from being efficient. Besides, a category can be a nonconvex region which is a union of several overlapping or disjoint subregions. Hence an automatic classification system may suffer from large memory requirements or poor performance.

A fuzzy based method should be implemented that should effectively categorize the documents into specific categories and find out the most similar documents. The proposed method should overcome the following issues:

1. Curse of dimensionality problem.
2. Nonconvex category boundaries.
3. Large memory requirements or poor performance.

The objective of similar document detection is to efficiently find a subset of documents within a large collection that are textually similar to a given query document.

IV. SYSTEM OVERVIEW

In this section, the major phases in the proposed method are described in detail:

- Preprocessing and Feature Extraction
- Training Phase
- Testing Phase

A. Preprocessing and Feature Extraction

A document set D will be given as the input. These documents will be stored in separate folders. The names of the folders are the different categories. This document set should be preprocessed before doing further operations. After preprocessing, a set of features, which are the main words in the document, are obtained. A function called stemmer is used for preprocessing. Stemmer performs the basic preprocessing tasks like stemming, lemmatization etc. For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set. The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:
am, are, is→ be
car, cars, car's, cars'→ car

The result of this mapping of text will be something like:
the boy's cars are different colors → the boy car be differ color

However, the two words differ in their flavor. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

The words obtained after preprocessing are separated by a space. These words are the features and are to be written into a feature file. Each document is converted into a vector format by comparing the words in the document with the words in this feature file. The vector will be the term frequency of the document and another vector, which shows the categories or folders the document is a part of, is also created. Thus after preprocessing and feature extraction a triplet consisting of a set of training patterns, a set of features and a set of categories is obtained.

B. Training Phase

The different steps of the training phase are shown in Figure 1.

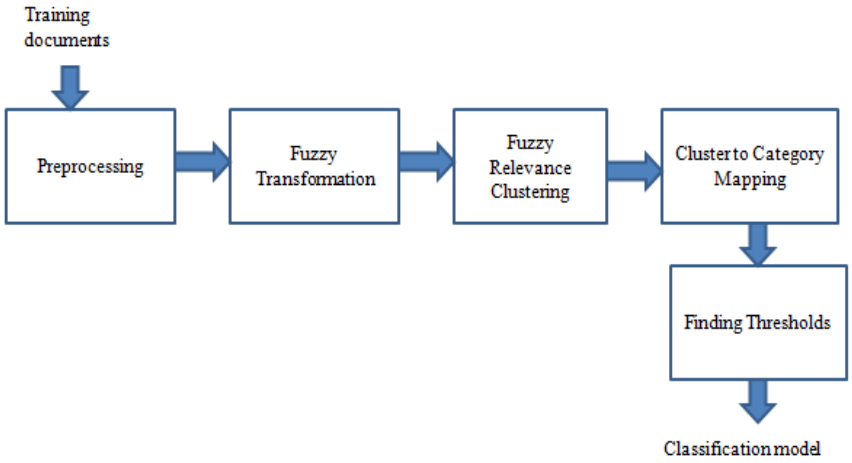

Fig. 1 Block diagram of Training Phase

After preprocessing of the documents, training phase can be started. First the high dimensional document is transformed into a low dimensional fuzzy relevance vector. This is called fuzzy transformation. Hence, a document with m dimensions is transformed into p dimensions, here m is the number of features and p is the number of categories. For this, the degree of relevance of feature to category and degree of relevance of feature to document is calculated. From this, the degree of relevance between document and category can be obtained. The combination of these values gives the fuzzy relevance vector.

The second step is grouping the obtained fuzzy relevance vectors into clusters. At the beginning, no clusters exist. At that time, the initial vector is considered as a cluster. After that, more vectors are put into the same cluster or new clusters are created based on the two factors: label similarity and cluster similarity. In the system, the number of clusters need not be specified in advance by the user. Rather, new clusters are created automatically and incrementally.

Then the clusters are mapped to the specific categories. The relationship for this mapping is linear. The cluster similarity vector calculated in the clustering step is converted into category similarity vector during this step.

After obtaining the category similarity vector, the categories that the document belongs will be obtained. Usually a document belongs to multiple categories. Thus a hard limiting function is provided based on the thresholds.

C. Testing Phase

1) Text categorization:

After training phase, a classification model will be obtained. On obtaining the classification model, testing can be started. When an unseen document d^t is given, the steps shown in Figure 2 are performed to determine the categories to which it belongs and also to find the similar documents. The path with arrows in blue color determines the categories that a document belongs to and the path with the arrows in green color determines the most similar documents from the training documents that a particular document belongs to. The category labels obtained can be used to determine the most similar documents.

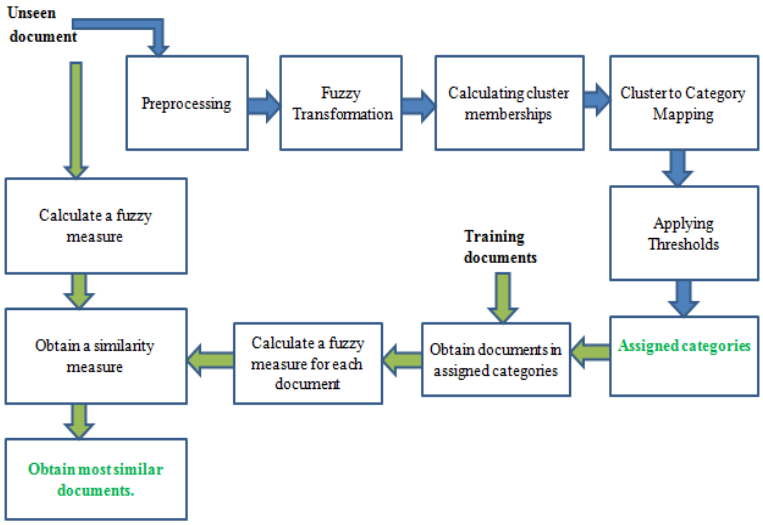

Fig. 2 Block diagram of Testing Phase

At first, the transformation to fuzzy vector step occurs. In that, the degree of relevance of feature to document is calculated. Then the relevance between the document and category can be calculated. The similarity of vector to each cluster, ie., the cluster similarity vector is being calculated as the next step. After that, this cluster similarity vector is transformed into category similarity vector. Then the thresholds are applied. Thus multilabel text classification or categorization can be achieved.

2) Similar Document Retrieval:

By using the categories obtained after performing multilabel classification, similar document retrieval can be performed. Similar document detection plays important roles in many applications, such as file management, copyright protection, and plagiarism prevention. The vector form of the query document as well as each document in the categories obtained should be calculated. Usually, the dimensionality of a document is large and the resulting vector is sparse, i.e., most of the feature values in the vector are zero. Such high-dimensionality and sparsity can be a severe challenge for similarity measure which is an important operation in text processing algorithms. This challenge can be overcomed by transforming into a fuzzy vector.

Using the vector a similarity measure is calculated between the query document and each training document of specific categories. The steps for obtaining the similarity measure are described in the following steps below:

- Obtain the difference between each component in the vector.
- Take the absolute of the differences.
- Obtain the sum of the values obtained.
- Sort the results.
- Return the documents that have minimal value.

The documents with similar value will be the most similar documents.

V. PERFORMANCE EVALUATION

In this section, RCV1(Reuters dataset) is used as dataset. The performance of the system is calculated mainly based on two measures:

A. Accuracy based on number of training documents.

B. Accuracy based on rho (ρ) and epsilon(ε) values.

A. Accuracy based on number of training documents

The dataset consists of a number of folders where each folder consists of a number of documents. The name of the folder shows various categories. Accuracy has been checked by using different number of training documents in each folder. It has been found that there is not much variation in accuracy even if the number of training document varies. Accuracy remains constant even if the number of training documents is reduced. This can be understood by analyzing the graph shown in Figure 3.

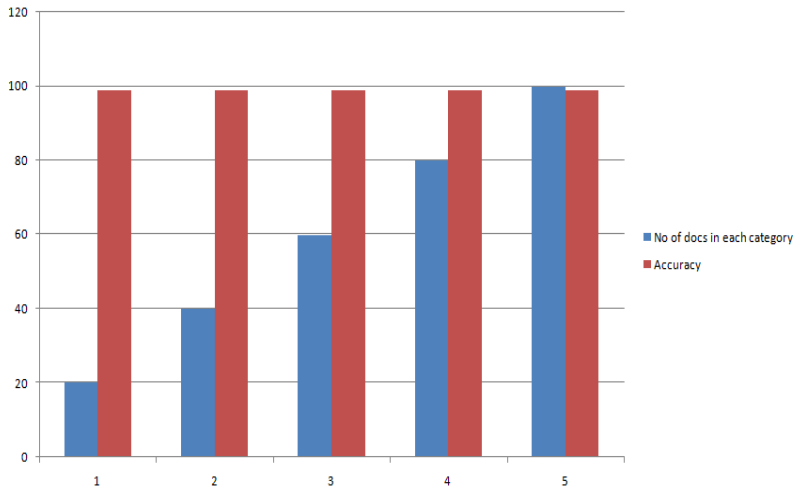

Fig. 3 Graph showing accuracy based on number of training documents

B. Accuracy based on rho and epsilon values

The rho (ρ) and epsilon (ε) values has a major effect on accuracy of the system. The rho (ρ) value is a predefined threshold for checking the cluster similarity. The epsilon (ε) value is a threshold for checking the label similarity. These are used mainly in the clustering step. By analyzing, it is found that very high and very low values of these two thresholds highly affect the accuracy. This is because if the values are low, very small clusters are created and if the values are high, very large clusters will be created. Both the cases affect the accuracy. Hence, a medium value of the threshold is best suited for obtaining high accuracy. Results can be analyzed in detailed by verifying the graph shown in Figure 4.

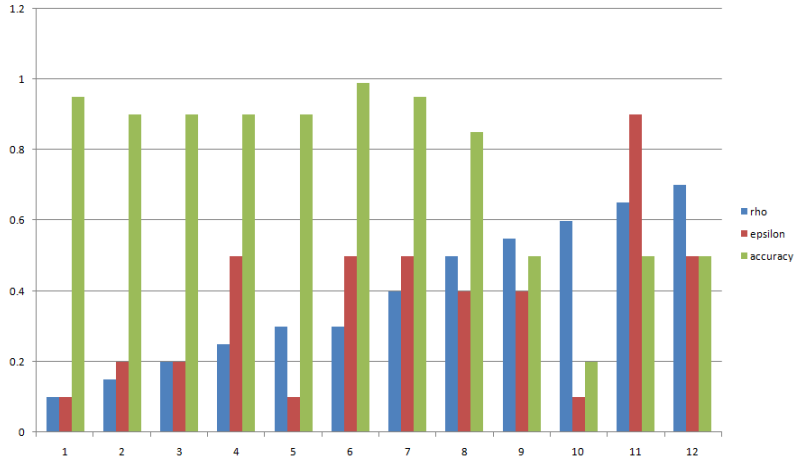

Fig. 4 Graph showing accuracy based on rho and epsilon values

VI. CONCLUSION

Text categorization and similar document retrieval are the fundamental task in such aspects of natural language processing such as information retrieval, information extraction, and text mining. Text classification is the process of classifying an incoming stream of documents into predefined categories through the classifiers learned from the training samples. Usually a text document often belongs to multiple categories in real tasks. Similar document detection efficiently finds a subset of documents within a large collection that are textually similar to a given query document. A fuzzy based approach has been developed to overcome these two major problems. Similar documents can be easily retrieved with the help of categories found out during multilabel text categorization. In future, the system can be modified by improving the accuracy of the results and speed of computations.

ACKNOWLEDGMENT

First of all the authors would like to thank god almighty for showering them with his grace and blessings without which the paper would not have been completed. We are also grateful to the students of M.Tech, Department of Computer Science and Engineering, Viswajyothi College of engineering and Technology, Vazhakulam for their patience and enthusiasm in all the stages of training that was required for improving the quality and presentation of the paper. Also the authors gratefully acknowledge all the staff members of the Department of Computer Science and Engineering, Viswajyothi College of Engineering and Technology, Vazhakulam for their whole hearted support and encouragement throughout the work.

REFERENCES

[1] Shie-Jue Lee, and Jung-Yi Jiang, "Multilabel Text Categorization Based on Fuzzy Relevance Clustering,"IEEE Transactions on Fuzzy Systems, Vol. 22, No. 6, pp. 1457-1471 December 2014.

[2] Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, vol. 26, No. 7, July 2014.

[3] Dino Isa, Lam Hong Lee, V.P.Kallimani, and R.RajKumar, "Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine,"IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 9, pp. 1264-1272, September 2008.

[4] Murat Can Ganiz, Cibin George, andWilliam M. Pottenger, "Higher Order Naïve Bayes: A Novel Non-IID Approach to Text Classification,"IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 3, pp. 1022-1034, March 2011.

[5] Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification,"IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 7, pp. 335-349, July 2011.

[6] Mohammad Salim Ahmed, Sourabh Jain, Fahad Bin Muhaya and Latifur Khan, "Predicted Probability Enhancement for Multi-Label Text Classification using Class Label Pair Association,"IEEE Conference on Evolving and Adaptive Intelligent Systems, pp. 70-77, April 2013.

[7] Kaveh Ahmadi-Abhari, Ali Hamzeh and Sattar Hashemi, "Voting Based Learning Classifier System for Multi-Label Classification,"Proceedings of the 13th annual conference companion on Genetic and evolutionary computation, pp. 355-359, July 2011.

[8] Jose Antonio Sanz, Alberto Fernandez, Humberto Bustince and Francisco Herrera, "IVTURS: A Linguistic Fuzzy Rule-Based Classification System Based on a New Interval-Valued Fuzzy Reasoning Method with Tuning and Rule Selection,"IEEE transactions on Fuzzy Systems, Vol. 21, No. 3, pp. 399-411,June 2013.

[9] Jung-Yi Jiang, Shian-Chi Tsai and Shie-Jue Lee, "FSKNN: Multi-Label Text Categorization based on fuzzy similarity and k nearest neighbors", Vol. 39, Issue 3, February 2012.

