A FUZZY NAIVE BAYES CLASSIFICATION USING CLASS SPECIFIC FEATURES FOR TEXT CATEGORIZATION

V. Bharathi¹, P.K. Jayanivetha², K. Kanniga³, D. Sharmilarani⁴
¹(Dept of CSE, UG scholar, Sri Krishna College of Technology, Coimbatore, India, bharathijayaBE2016@gmail.com)  
²(Dept of CSE, UG scholar, Sri Krishna College of Technology, Coimbatore, India)  
³(Dept of CSE, UG scholar, Sri Krishna College of Technology, Coimbatore, India)  
⁴(Dept of CSE, Assistant Professor, Sri Krishna College of Technology, Coimbatore, India)

Abstract— With the rapid growth of information text categorization has become one of the important technique for organizing and handling text data. It is most needed to label the documents automatically with pre-defined set of topics. There are various techniques has been proposed for automatic text categorization. A Bayesian classifier was used for automatic text categorization where the specific features subset for each class was selected. Then the selected features are given as input to the Bayesian classifier for text categorization. This approach has some drawbacks like computationally costlier, consuming more time and assumes independence of features. In order to overcome these issues and to improve the accuracy of text categorization Fuzzy naive bayes classifier is proposed. In this text categorization is performed by using class specific features in fuzzy naïve bayes classifier. The experimental results are conducted to prove the effectiveness of the proposed method in terms of accuracy, F-measure and G-Mean.

Keywords— Text Categorization; Bayesian Classifier; Fuzzy Naïve Bayes Classifier; Fuzzy Rule

1. INTRODUCTION

 Due to increasing the availability of documents in digital form, information retrieval has gained a prominent status in the information system field. Text categorization is an activity of labeling natural language texts with thematic categories from a predefined set. It is used to classify new stories, to guide a user’s search through hypertext and to find out the interesting information on WWW. Many machine learning techniques has been proposed to address text categorization task. The key challenge in text categorization is learning in huge volume of data space. The high dimensional features may even hurt the classification performance due to redundant and irrelevant features in the dataset. It can be handled by feature selection process and it also speed up the learning of classifiers.

 A Bayesian classifier is one of the machine learning techniques was used for automatic text categorization process. In this approach the redundant and irrelevant features are removed by selecting the most significant features for each class to reduce computation burden of Bayesian classifier. Then the class specific features are given as input to the Bayesian classifier which is followed by the Baggenstoss’s PDF Projection Theorem. This approach is computationally costlier, time consumption process and it has less accuracy. So an efficient automatic text categorization method is proposed based on Fuzzy naïve bayes classifier with class specific features. The proposed improved the accuracy and reduce the computation cost of text categorization.

2. LITERATURE SURVEY

 Tang, B., & He, H. [1] proposed Extended Nearest Neighbor (ENN) method for pattern recognition. It is based on the maximum gain of intra class coherence. This method improved the K nearest neighbor (KNN) method by analyzing the generalized class wise statistics. It has a capability of learning from the global distribution that improved the performance of pattern reorganization. During classification the proposed ENN considered the test samples as their nearest neighbors along with the nearest neighbor of test samples which improves the accuracy of classification. The major drawback of this method is high time consumption.

 Williams, K., & Calvo, R. A., [2] introduced a framework for text categorization. It managed the documents, categorization algorithms and collection of documents and so on. The proposed framework consists of both concrete classes and abstract classes like naïve bayes learner and Boolean learner respectively. The naïve bayes learner may be used without custom development and the Boolean learner required the user to implement certain behaviors before using them. It provided the reusability of design and implementation in application which uses text categorization.

 Yu, W., & Linying, X., [3] proposed an improved K Nearest Neighbor (KNN) classifier for text classification. Initially the k cluster centers of minority classes were calculated using K means. Then the sample data under these clusters were assumed as it satisfied the multivariate Gaussian distribution. A large number of sample points were generated for each cluster center which is based on the multivariate Gaussian distribution. From the large number of sample points choose only a limited number of sample points to add to the original dataset based on the weight measurement of the similarity between the cluster center and new generated sample points. Moreover this resolved the class imbalance problem. Thus
the improved KNN has better classification but with the decrease of the class imbalance degree.

Yang, W. et al [4] proposed an improved parallel algorithm based on MapReduce for text categorization. In this paper an improved MapReduce algorithm called as Rocchio algorithm is proposed which handles huge volume of data for text categorization. The MapReduce Rocchio algorithm consists of number of mappers and number of reducers to handle the huge volume of data. The mappers get input as key value pair and it produce intermediate key value pair and combine the results of each mapper in the reducers. Thus the parallel computation for text categorization is carried out which improves the speed and accuracy of text categorization process. However Rocchio algorithm has linear combination too simple for classification.

Tang, B., et al [5] presented a novel and efficient feature selection framework for text categorization. The proposed framework ranks the features with their discriminative capacity for text classification. A new information measure was introduced named as Jeffreys-Multi-Hypothesis (JM) divergence by analyzing asymptotic properties of two measures are Jeffreys divergence measure and Kullback divergence measure. The proposed divergence measure used to measure multi distribution divergence for multi class classification. Based on this measure two feature selection methods named as maximum discrimination and maximum discrimination-x2 was proposed for text categorization. However in this proposed text categorization methods has low discriminative capacity.

Fukumoto, F., & Suzuki, Y., [6] addressed problems involve in text categorization and proposed a technique for text categorization. The proposed technique was used to minimize the impact of temporal effects in both learning techniques and feature selection. It involves three step processes are collection of documents by Latent Dirichlet Allocation (LDA), feature selection using temporal based feature selection (TbFS) method and Document categorization using temporal based transfer learning (TbTL).The temporal based classification method used only a certain number of labeled training data for classification. Thus the proposed method was automated and can be easily applied to different languages and new domains.

Kumar, M. A. & Gopal, M. [7] proposed a new text categorization system that combined Least Square Twin Support Vector Machines (LST SVM) for document classification and clustering of words for document representation. The distributional clustering of words used all words features for feature selection from the input data. This feature selection creates a compact representation in low dimensional space. In addition to that LST SVM gain benefits from this compact representation as training Complexity depends on the input dimensions and it yields fast training with competitive results.

Fragoso, R. C. P. et al [8] presented Category-dependent Maximum f Features per Document (cMFDR) for text categorization. cMFDR is a filter method which is an extension of MFDR method both were used for feature selection. Where the MFDR select the features based on one global threshold and it degraded categorizes which few relevant features. So in the proposed text categorization method compute the one threshold per category to promise that every category contains different number of features. Thus the proposed text categorization method has high accuracy. But the major drawback of the proposed method was static value of f parameter may affect the performance of text categorization.

Pinheiro, R. H. W. et al [9] proposed a method based on prototype selection and dissimilarity representation for text categorization. Bag-of-Words has problems like high feature-to-instance ratio and it produce sparse high dimensional feature vectors. These problems were resolved by dissimilarity representation method. The prototype selection method was utilized to choose a smaller representation set that increased the benefits of utilizing dissimilarity representation. In this features were represented as cosine similarity between two documents. Then the similarity is computed using prototype selection algorithm which reduced the high dimensionality of the task and reduced the sparseness.

Yasotha, R., & Charles, E. Y. A. [10] proposed Latent Dirichlet Allocation (LDA) based approach for automatic text categorization. Most of all text categorization methods used string matching for automatic text categorization. But in LDA, the clusters are identified by labeling based on the underlying natural clusters on the domain in concern. However the proposed method has less accuracy.

3. PROPOSED METHODOLOGY

In this section, the proposed system is described in detail. In which the automatic text categorization is performed using fuzzy rule with the Bayesian approach which is called as fuzzy naive Bayes classifier. Initially the class-specific features are selected from the high dimensionality text features and irrelevant or noisy features. Then the class specific features are classified by combining fuzzy rule with Bayesian classification which improves the probabilities of each of the classes by considering which are incorrectly classified or left un-classified by applying Bayesian classifier. Thus the classification accuracy and the efficiency of the classification approach are improved effectively.

A. Pre-processing

The input data are converted into understandable and easy format of further processing techniques called as pre-processing. One often applied transformation to the input text is the substitution of characters outside of the usual 26-letter English alphabet with a single space. Multiple spaces in the documents are reduced to single space and then the upper case is converted into lower case. Such mapping will make any punctuation indistinguishable from white space it is considered as uninfluential loss information. Here the input data are preprocessed by discarding those terms that occur in less than two documents and ignored those terms in a stop list.

B. Feature Selection

After the pre-processing step, the more relevant and more important features are selected for text
categorization. It is one of the most fundamental tasks which are done before any classification task which is needed to be accomplished is that of feature selection and document representation. The main intend of this feature selection process is to reduce the high dimensionality of text features and the discard the irrelevant and noisy features in the data. Considering a Text Classification problem (TC) with N predefined topics. Assume that c_i is the class label for values of i ∈ {1, 2, …, N}. A dictionary D is formed for a given dataset with M terms. The concept of Bag of words is explained as follows:

\[ \text{Output: Features} \]
\[ \text{Input: Documents for a given training data set with N topics.} \]

The features under the reference class \( c_0 \). Thus the features feature index vector and then estimate the mutual information (MI) and information gain (IG) the values of \( i \in \{1, 2, \ldots, N\} \). A dictionary D is formed for a given training dataset with M terms. The concept of Bag of words is explained as follows:

\[ \text{PROCEDURE:} \]
\[ \text{1. Form a reference class c0 which consists of all documents;} \]
\[ \text{2. Calculate the score of each feature based on a specific criteria,} \]
\[ \text{and rank the feature with the score in a descending order;} \]
\[ M1(x_k, c_i) = \log \frac{p(x_k, c_i)}{p(x_k)p(c_i)} \]
\[ IG(x_k, c_i) = p(x_k, c_i) \log \frac{p(x_k, c_i)}{p(x_k)p(c_i)} + p(\bar{x}_k, c_i) \log \frac{p(\bar{x}_k, c_i)}{p(\bar{x}_k)p(c_i)} \]

Where \( x_k \) represents binary feature \( c_i \) represents a category and \( \bar{x}_k \) represents the term does not occur in the document.

3. Choose the first K features \( z_i \), the index of which is denoted by \( l_i \).

A. Fuzzy Naïve Bayes Classifier

The performance of the naïve bayes classifier is improved by applying fuzzy logic is applied on the Bayesian classification. Fuzzy rule are in the form of conditional IF THEN rules. It is explained as follows:

\[ \text{IF} x \text{ is A THEN} y \text{ is B} \]

In the above fuzzy rule x and y represents linguistic variables and A and B represents linguistic values are determined to the universe of disclosure. These rules consider the missing values in between the precise values being defined by formulating the fuzzy rules accordingly. The rules are generated and given to fuzzy naïve bayes classifier.

\[ \text{FNBclassify}(x) = \text{argmax}_z \sum_{a_1 \in A_1} p(a_1 | z) \mu_{a_1} \ldots \sum_{a_n | z} p(a_n | z) \mu_{a_n} \]

In the above equation, \( \mu_{a_1} \in [0, 1] \) represents a membership function or degree of truth of the fuzzy rule of attribute \( a_1 \in A_1 \) in a new example x. To be conservative, it is needed that all degrees of truth are normalized in the current variable asignment, in this case \( \sum_{a_1 \in A_1} \mu_{a_1} = 1 \). The probabilities of the above equation can be calculated as follows:

\[ P(Z = z) = \frac{\sum_{\text{w} \in N} \mu_w^Z + 1}{|N| + |D(Z)|} \]

\[ P(A_i = a_i | Z = z) = \frac{\sum_{\text{w} \in A_i} \mu_w^Z + 1}{(\sum_{\text{w} \in A_i} \mu_w^Z) + |D(A_i)|} \]

In the above equations, N represents the training set consisting of all rules \( e = \{a_1 = a_{1z}, \ldots, a_y = a_{yz}, Z = z\} \), \( \mu_w^Z \in [0, 1] \) represents the degree of truth of \( x \in Z \) in a example \( w \in N \) and \( \mu_w^Z \in [0, 1] \) is the membership attribute \( a_i \in A_i \) in such example. The Laplace correlation is applied to calculate the probabilities. For science writers is [7].

5. EXPERIMENT RESULTS

In this section the experimental results are conducted to prove the effectiveness of the proposed method. For the experimental purpose, 20 Newsgroup dataset is used. The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. Here we used 200 documents with 10 different topics for the experimental purpose. The experiments are conducted in terms of accuracy and F-measure and G-mean.

A. Accuracy

Accuracy is the measure of correctly categorized text in all documents. It can be calculated by

\[ \text{Accuracy} = \frac{(\text{Truepos} + \text{Trueneg})}{(\text{Truepos} + \text{Trueneg} + \text{Falsepos} + \text{Falseneg})} \]

A. Accuracy

Accuracy is the measure of correctly categorized text in all documents. It can be calculated by

\[ \text{Accuracy} = \frac{(\text{Truepos} + \text{Trueneg})}{(\text{Truepos} + \text{Trueneg} + \text{Falsepos} + \text{Falseneg})} \]

\[ \text{Fig. 1. Comparison of Accuracy} \]

Volume: 04 Issue: 03 2017 www.ijmtes.com 105
Fig. 1 shows the comparison of accuracy between existing Bayesian classifier based text categorization and proposed Fuzzy naïve Bayes classifier based text categorization. X axis takes the methods and Y axis takes the accuracy value in %. From the figure it is proved that the proposed method has high accuracy than the existing method.

B. F-measure

F-measure is a measure of test’s accuracy. It considers both the precision and recall score. Precision value is evaluated according to the relevant information at true positive prediction, false positive.

\[ \text{Precision} = \frac{\text{Truepositive}}{\text{Truepositive} + \text{Falsepositive}} \]

The Recall value is evaluated according to the classification of data at true positive prediction, false negative.

\[ \text{Recall} = \frac{\text{Truepositive}}{\text{Truepositive} + \text{Falsenegative}} \]

From the precision and recall score the F-measure score is compute as follows:

\[ F - \text{measure} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \]

Fig. 2 shows the comparison of F-measure between existing Bayesian classifier based text categorization and proposed Fuzzy naïve Bayes classifier based text categorization. X axis takes the number of features and Y axis takes the F-measure value. From the figure it is proved that the proposed method has high F-measure than the existing method.

C. G-Mean

G-Mean is a measure used to measure the performance of classification which is a combination of precision and recall.

\[ G - \text{Mean} = \sqrt{\text{Precision} \times \text{Recall}} \]

Fig. 3 shows the comparison of G-mean between existing Bayesian classifier based text categorization and proposed Fuzzy naïve Bayes classifier based text categorization. X axis takes the number of features and Y axis takes the G-mean value. From the figure it is proved that the proposed method has high G-mean than the existing method.

6. CONCLUSION

In this paper an efficient method called as fuzzy naïve bayes classifier is proposed for automatic text categorization. Here the most important features are selected for each classes those features are termed as class-specific features. The class specific features reduced the dimensionality of features and it speed up the process of fuzzy naïve bayes classifier. The texts are categorized by combining the fuzzy rule with Bayesian classification which facilitates improve in the probabilities of each of the classes by considering many of the features that are left un-classified or incorrectly classified by applying the Bayesian classification alone. The experimental results proved that the proposed fuzzy naïve bayes classifier has high accuracy, high F-measure and high G-mean than the existing method.

REFERENCES


