© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Document Categorization by using Weighted J48 Classifier

Sonali Suskar, Dr. S. D. Babar
Department of Computer Engineering SIT College of Engineering, Lonavala, Maharashtra, India

ABSTRACT

In the field of information retrieval text categorization is the key research area in present. The text categorization selects entries from set of prebuilt categories and allots those to a document. Learning with high dimensional data space is challenging in a text categorization method. Learning with high-dimensional features may prompt a heavy calculation overhead and may affect the classification performance of classifiers because of unrelated and repetitive features. To improve the "scourge of dimensionality "issue and to accelerate the learning procedure of classifiers, it is important to perform feature reduction to reduce the size of features. This paper introduces a Bayesian arrangement approach and WeightedJ48 classifier for auto text categorization using class-specific features. For text classification, the proposed strategy selects a specific feature subset for every class. The presented system reconstructs PDF in raw data space from class specific PDF in low dimensional feature space and assembles Bayes classification rule utilizing Baggenstoss PDF Projection Theorem. The detectable importance of this methodology is that many feature selection criteria. The WeightedJ48 classifier saves the time and memory. The proposed system also uses Term weighting concept for pre-processing. These methods increase the accuracy of classification, feature selection process, and improve the system performance.

Keywords: Text categorization, class-specific features, Feature selection, PDF projection and estimation, dimension reduction, WeightedJ48, Term weighting.

I. INTRODUCTION

As data size on net as well as different companies will grow, there is huge requirement of a method for dealing with the huge size of information that can be filter and deals these information types.

The main categories is to separate the free text files in the categories that are defined previously, categorization of emails and files in folder tree, labelling of the topics, Particular processing operations, structures search as well as surfing or searching files which has long term interests or dynamic task depending interests. In different contexts professionals are selected to classes the new items, yet this procedure is especially time taking and in addition will as exorbitant so bounding its applicability apparently there is a more enthusiasm for the research and development work of the strategies for text categorization automatically. There are various classifications and machine-learning techniques are developed for categorization of text like the one rule learning algorithms nearest neighbour's classifiers, Support Vector Machines, decision trees etc.

Text categorization (TC) described as text classification, in this a documents are automatically classified by using predefined set. This process can be used in many systems; also in automated indexing of scientific articles based on predefined thesauri of terms, which are technical, filing patents inside the patent directories, chosen dissemination of data-to-data consumers, hierarchical catalogues for automated

population of resources present on web, filtering of spam and so on. The basic issue in text categorization is learning, it is a high dimensional in of ration space. Generally, files are depicted as the "sack of words", each word or phrase presenting the file once or more is considered as feature.

For TC one of the common feature reduction techniques feature selection, inside this only a subset of features is preserved as it is as well as the reaming of them are deleted. In common, feature selection method can be divided in three types: Embedded, filter and wrapper approach. The filter approach calculates the need of each specific feature with a score, which depends on characteristics of information, and only the features with higher score chosen. Totally oppose with these filter approach will not involve with the conditions with learning; the wrapper approach chooses best features as a learning condition. The greedily search of wrapper approach, requires training the classifiers at every step and has a high calculation overhead. The embedded approach is consolidation of filter and wrapper approaches that compute the importance of every feature but also employ a search technique supported by learning algorithm.

This study implements a Bayesian classification approach and WeightedJ48 classifier for auto text categorization, which makes use of class-specific features. Different from traditional approaches for text categorization, our designed technique chooses a particular feature subset for every class.
In addition, system used Term weighting concept for categorization of unstructured text documents. This paper presents study about the related work done in Literature review, the proposed approach modules description, mathematical modeling, algorithm and experimental setup in system overview and finally a conclusion is provided.

II. REVIEW OF LITERATURE

In the paper [1] authors have developed a system which is automatically categorize text by making use of class specific features which is Bayesian classification. The proposed method allows selecting the vital features for every class. Class specific features are used for classification by Naive Bayes rule designed by researchers from Baggenstoss PDF projection theorem. The major advantage of derived technique is it can make use of present feature selection conditions.

In the paper [2], authors developed a system offending optimal classification by making use of class-specific features known as the hypothetical establishment also provided utilization examples. To project PDFs in low dimensional feature space back towards raw information space is possible due to new PDF.

To assess the PDFs of class-specific features and the transformation of everyPDf back to raw data space for analysis, M-aryclassifier is produced. Albeit statistical adequacy is unimportant, the classifier in such a way created will get to be equal to the optimal Bayes classifier if featuresfulfill adequacy prerequisites exclusively for every class.

In paper [3], authors developed an automatic text categorization method as well as research its applications for text retrieving. A categorization technique designed using a combinationoflearning pattern known as instance based-learning as well as an advance document retrieving method i.e. feedback retrieval. The ability of suggested method is explained with 'MEDLINE' database by two actual document collections. Alsouse of programmed automated categorization of text retrieval is explored.

This paper [4] examines the principle method to deal with text categorization that are considered in the machine learning worldview The auto categorization

of text in respecified classes has seen blasting enthusiasm for the most recent 10 years, because of the expanded accessibility of files in digital structure and following required to sort out them. In investigation, group predominant style to deal with this issue depends on machine learning procedures: a general inductive process consequently constructs a classifier by learning from an arrangement of pre-classified documents as well as from classification characteristics. The upsides of this method compared to knowledge engineering methodology are a decent effectiveness, significant regarding work control, and versatility to various areas.

This paper [5] presents ideas feature selection methods, surveys related to existing feature selection methods for classification and clustering groups also contrasts distinctive algorithm and an arranging structure in view of pursuit methodologies, evaluation conditions, and information mining task, uncovers untried combinations, and gives rules for selecting highlight choice algorithms. With categorizing structure, they precede with attempt toward creating incorporated framework for intelligent feature selection.

The classical [six] Bayesian technique for classification needs information of the probability-density-function (PDF) of information or enough statistic for all class hypotheses. Hence, it is very hard to get a single low-dimensional sufficient statistic, sometime it is needed to use a sub-optimal yet still relatively high-dimensional feature set. The performance of methods is greatly limited by ability to estimate the PDF on a high-dimensional space separating training information.

This paper [7] implemented a new supervised classification technique the extended nearest neighbours that guesses input patterns as per the most extreme pick up of infraclass coherence. Dissimilar to the conventional k-nearest neighbour (KNN)strategy, where nearest neighbours of a test are utilized to assess a group membership, the ENN technique guesses in a ”2-way communication” style: it considers closest neighbours of test sample, along with who consider the test as their closest neighbours.

This paper [8] developed a method to consolidate multiple features in recognition of handwriting depending on two concepts: feature selection based combination as well as class dependent features. A nonparametric technique is utilized for feature evaluation, as well as first portion of paper focused on evaluation of features in their class separation as well as recognition capabilities. Further, multiple feature vectors are consolidated to generate a novel feature vector. The feature has diverse discriminating powers for various classes, another plan of selecting and consolidating class-dependent features is proposed. In addition, class is considered to have its own optimal feature vector differentiate itself from alternate classes.

In paper [9] authors has developed a system, which is automatically categorize text of Marathi files based on user profile with browsing history of user. Vector Space Model provides good outcomes compared with the present Probabilistic Models. The precision of the outcomes related to the system is very good compared with the Tamil language.

LINGO algorithm provides better cluster quality compared with different clustering methods. The categorization of Arabic texts has some issues, which demands to be solved particularly at the time utilizing stemming.

In paper [10] authors searched as well as observed that, a discussion in the various developers related the pros of utilizing the stemming in categorization of Arabic text. Due to this, authors have conducted the analysis of feature reduction technique for clearing the effect of this famous technique in the mining of text as well as classification of files. They also few

Arabic text condition to refuse the use of stemming in Arabic text categorization.

- Limitation:
  1. Feasibility of applying automatic text categorization to developed document filtering systems is not computed
  2. Energy efficiency is not considered with Time complexity is more

III. SYSTEM ARCHITECTURE / SYSTEM OVERVIEW

A. Problem Definition

The existing system uses the consolidation operation which that bias feature importance for discrimination. In addition, there are theoretical supports to choose the best consolidation operation. To overcome this problem, an alternative of using the combination operation to select a global feature subset for all classes, our system select a specific feature subset for every class, i.e. class-specific features. Also, build the Bayes decision rule for classification with these selected class specific features. The proposed system used WeightedJ48 classifier for classification and term weight concept to improve the accuracy of a system.

B. Proposed System Overview

In proposed system, a Bayesian arrangement approaches for auto text categorization utilizing class-particular features. Not like the ordinary methodologies for text categorization, has the proposed strategy chosen a particular feature subset for every class. To use these class-dependent features for classification, this system uses Baggenstoss PDF Projection Theorem (PPT) for reconstruction of PDFs in raw data space from class-specific PDFs in low-dimensional feature space and assembles a Bayes classification rule.

System assesses this technique's classification performance on some real-world benchmark data sets, contrasted with some feature selection approaches. To remove the unnecessary data, we use the term frequency concept with TFIDF.

According to this method, term frequency is calculated for each word to generate the training file. These training files provided to Feature selection process. In contribution, we use WeightedJ48 classifier for classification. These methods increase the accuracy of classification, feature selection process, and improve the system performance. Also system used Term weighting concept for categorization of unstructured text documents. During the categorization of text documents, term weighting allocates proper weights to various terms. It helps in to improve the categorization result.

1) Input and Read Dataset

In this module user, provide the dataset that is newspaper dataset with different topics and read the dataset.

2) Pre-processing approach

In the dataset, number of words present means it content numerous features, which can be hurt classification performance, are removed using stemming and stop ward operations.

Stop words: They are frequently occurring and insignificant words in a language that help construct sentences but do not represent any content of the documents.

Stemming: Stemming refers to the process of reducing words to their stems or roots.

3) Filter Approach

Here the number of words which has high score is filter out.

4) Term Weighted Concept:

To remove the unnecessary data in pre-processing step, we use the term frequency concept with TF-IDF. According to this method, term frequency is calculated for each word to generate the training file. During the categorization of text documents, term weighting allots proper weights to various terms.

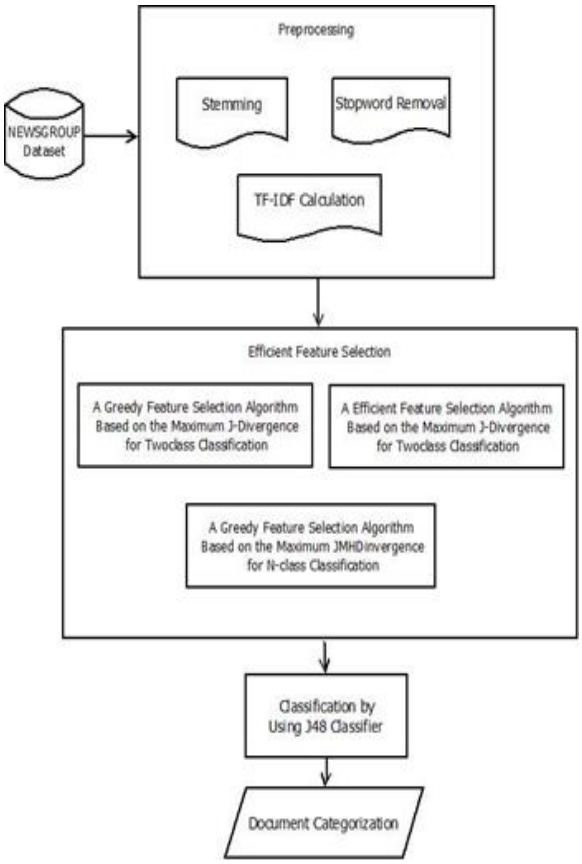

Figure 1. System Architecture

5) Feature Selection

This module takes the important words that are features as an input and select the features by applying feature selection method.

6) Classification

After selecting, the features apply classification. For Classification this system used WeightedJ48 classifier, which have high accuracy and get the text categorization output.

C. Mathematical Model

Let S is a system, where,

S = {Input, Process, Output}

- Input: I = Newsgroup Dataset
- Process = { p1, p2, p3, p4,p5 }

Where, P represent the total number of steps perform in system to get output.

1) P1 = Pre-processing Approach

In the dataset, number of words present means it content numerous features, which can be hurt classification performance, are removed.

p1 = {p11, p12}

Where, p11- stemming, p12-stopward

2) P2= Filter Approach

F = {F1, F2, Fan}

Where, F represent the set of filter words and F1, F2, Fan are number of high score words.

3) P3= Term Frequency Approach

P3= {P31, P32...P3n}

Where, P3 represent the set of term frequency and P31, P32...P3n are number of term frequency of each word.

4) P4= Feature Selection P4 = {P41, P42...P4n}

Where, P4 represent the set of feature and P41, P42...P4n are number of feature.

5) P5 = Classification P5= {P51, P52, ... P5n}

Where, P5 represent the set of Classifier and P51, P52...P5n are number of classes.

Output:

TextCategorization : T = {T1, T2, ... TNT}

Where, T represent set of Text Categorization and T1, T2, TN, are number of text categories.

D. Algorithm Used

Algorithm 1: Text Categorization using Class-Specific Feature Selection Approach

Input: Newspaper Datasets for a given training data set with N topics.

Procedure:

1) Form a reference class l0, which consists of all documents;

For each class j = 1: N do

2) Calculate the score of every feature based on specific criteria and rank the feature with score in a descending order;

3) choose the first F features xi, the index of which is denoted by Juju;

4) Estimate the parameters θj|0 under the reference classl0 and the parameters under the class li; End

Output: Given a document to be classified,

Output the class label l using following equation

$$l * \propto_{j\in\{1,2,\ldots\ldots N\}}^{argmax} \sum_{f=1}^{F} x_f log \frac{\bar{p}_{n^j{}_{fj}}}{\bar{p}_{n^j{}_{f^0}}} + logp(l_i)$$

Algorithm 2. WeightedJ48 Classifier

WeightedJ48 classifier is classification algorithm used for detecting the novel and multi novel class. For the problem classification decision tree methodology is used.Tomodel classification process, first tree is build. When tree is generated, it is connected with each column of the database, which results in classification for that column. Perform the classification process by using WeightedJ48 Classifier instead of Naive Bayes classifier, which improves the classification accuracy. Based on different parameter such as time for building the model, correctly and incorrectly classified instances mean absolute error WeightedJ48 gives better result than Naive Bayes classifier.

Process:

WeightedJ48 builds decision tree classification from a set of training data.

Training data is a set S = s1, s2, s3 of already classified samples.

Each sample Si consists of a p-dimensional vector (X1i, X2i, and PA) where Ox represent attribute values or features.

After sample, as well as class in which Si falls.

DS: Dataset

DT: Decision Tree

1) Input: Training data DS

2) Output: Decision Tree DT

3) DSTBUILD (_DS)

4) {

5) DT=';

6) DT=Generating root node and labelling with splitting attribute;

7) DT=Add arch to root node for each splitting predicate and label;

8) DS=by applying split predicate to DS database is created;

9) If stopping point reached for this path, then;

10) DT 0 =generate leaf node and label with the appropriate class;

11) DT 0 =DSTBUILD (_DS);

12) Else

13) DT 0 =DSTBUILD (DS);

14) DT=add DT 0 to arc;

15)}

The WeightedJ48 classifier to establish the tree does not need any code. During construction of a tree, WeightedJ48 rejects missing qualities i.e. the quality for those things can be anticipated focused around which is the thought about characteristics qualities for the other record.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The setup is built on Java framework with Windows platform. As a development tool Net beans IDE is used. No Specific hardware needed to run the system; it can run on any standard machine.

B. Dataset: 20-NEWSGROUPS dataset, REUTER dataset.

Dataset Info: This contains data related to different news category (news dataset).

Input: Dataset

Output -: documents classification into different categories (topic) by using naïve byes.

C. Experimental Results

Table I describes the time required in Ms for classification using Naive Bayes and WeightedJ48 classifier. WeightedJ48 consume less time for classification than the Naive Bayes classifier.

Table 1. Time Efficiency Comparison

	Naïve Bayes	WeightedJ48
Time in Ms	2000	1000

Figure 2 represent the graphical comparison of time efficiency of Naive Bayes and WeightedJ48 classifier respectively. X-axisrepresents classifiers andy-axis represents time needed in ms.

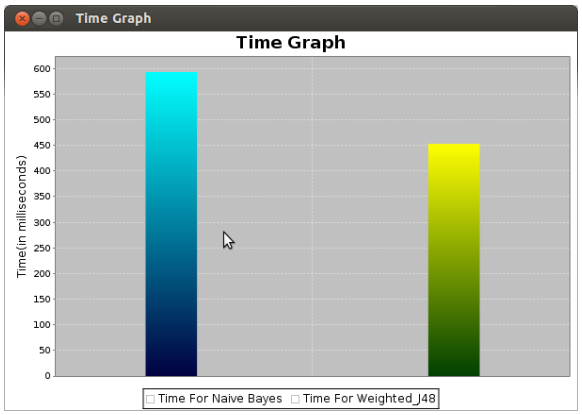

Figure 2. Time Graph

Table II describes the memory required in bytes for storing classifier result of Naive Bayes and WeightedJ48 classifier. WeightedJ48 consume less memory than the Naive Bayes and improve the classification result.

Table 2. Memory Comparison

	Naive Bayes	WeightedJ48
memory in bytes	23000	18000

Figure 3 represent the graphical comparison of memory consumption in Naive Bayes and WeightedJ48 classifier respectively. X-axis represents classifiers and y-axis represents memory consumed in bytes. Naive Bayes requires 23000 bytes memory and WeightedJ48 requires 18000 bytes memory for classification.

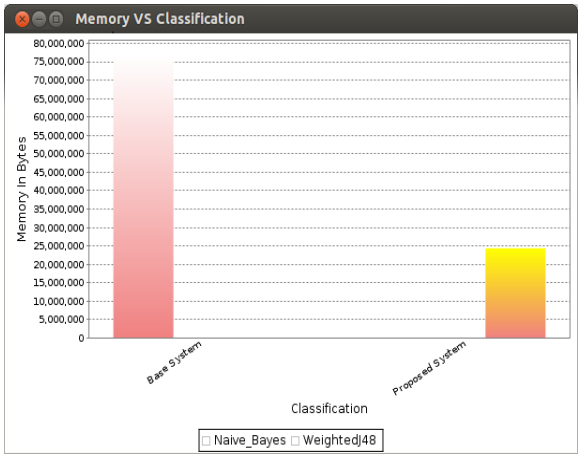

Figure 3. Memory Graph

V. CONCLUSION

For applications of information retrieval, machine learning and text mining, text categorization plays vital role. This paper presents the recent text categorization and feature selection methods used for text document classification. The feature selection method chooses the important features from classification. This system also used term weight concept to categorized unstructured data. During the classification of text documents, term weighting allots proper weights to various terms. It helps to improve the categorization results. The experimental result shows that system save both time and memory and improves the system performance.

VI. REFERENCES

[1]. Bo Tang, Haibo He, Paul M. Baggenstoss, and Steven Kay, "A BayesianClassification Approach Using Class-Specific Features for Text Categorization",1041-4347 (c) 2015 IEEE, Transactions on Knowledge and DataEngineering.

[2]. Paul M. Baggenstoss, "The pdf projection theorem and the class-specific method," IEEETransactions on Signal Processing, vol. 51, no. 3, pp.672-685, 2003.

[3]. W. Lam, M. Ruiz, and P. Srinivasan, "Automatic text categorization andits application to text retrieval," IEEE Transactions on Knowledge and Data Engineering, vol. 11, no. 6, pp. 865-879, 1999.

[4]. F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1-47, 2002.

[5]. H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491-502, 2005.s

[6]. P. M. Baggenstoss, "Class specific feature sets in classification," IEEE Transactions on Signal Processing, vol. 47, no. 12,pp. 3428-3432, 1999.

[7]. B. Tang and H. He, "ENN: Extended nearest neighbor method for pattern recognition research frontier]," IEEE Computational Intelligence Magazine, vol. 10, no. 3, pp. 52-60, 2015.

[8]. I.-S. Oh, J.-S. Lee, and C. Y. Suen, "Analysis of class separation and combination of class-dependent features for handwriting recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 1089-1094, 1999.

[9]. J. J. Patil and N. Bogiri, "Automatic text categorization: Marathi documents," 2015 International Conference on Energy Systems and Applications,Pune, 2015, pp. 689-694.

[10]. F. S. Al-Anzi and D. AbuZeina, "Stemming impact on Arabic textcategorization performance: A survey," 2015 5th International Conferenceon "Information Communication Technology and Accessibility"IEEE (ICTA),Marrakech, 2015, pp. 1-7.