International Journal of Research in Science advanced Technology and management studies (IJRSTMS)

ISSN: 2459-425X • Website: www.ijrstms.com

Class Specific Features using J48 Classifiers for Text Categorization

DR..CHANDRA MOHAN

Department of Computer Engineering

JSPM's Rajarshi Shahu College of Engineering,Pune

Pune, India

Dr.SAI

Department of Computer Engineering

JSPM's Rajarshi Shahu College of Engineering,Pune

Pune, India

***Abstract*:** Machine learning for text classification is the foundation of document classification, news filtering, document routing, and personalization. Text classification is continuing to be a standout amongst the most researched NLP issues due to the ever increasing amounts of electronic documents and digital libraries. In text domains, to make the learning task effective and more precise effective feature selection is essential. Automated feature selection is important for text classification to decrease the feature size and to speed the learning procedure of classifiers. In this paper, learned features are classified by using J48 classifier and prove that it outperforms Naive Bayes Classifier. This classification accurately categorized the documents. The performance of the system is estimated on 20-NEWSGROUPS dataset. The experimental results of extensive experiments demonstrate that the effectiveness of the proposed techniques.

Keywords: Machine Learning, Feature Extraction, Feature Selection, Classification, Document Categorization.

I. INTRODUCTION

In recent years, the number of electronic documents that are available on the Internet or on corporate intranets daily gains, forwarding or filtering, effective retrieval of textual information has become an important component in many information organization and administration tasks. An increasingly useful tool for managing this huge amount of data is the categorization of texts the task of assigning one or more predefined categories to a natural language text documents based on their content.

There are numerous statistical machine learning techniques and classification methods which have been applied to the categorization of text in recent years. Some, state of the art text, categorization algorithms include Support Vector Machine (SVM), Naive Bayes, Ridge Regression, K-Nearest Neighbor (KNN), Logistic Regression (LR) and Linear Least Squares Fit (LLSF).

Documents must be presented in this way is suitable for a general learning process. Mostly widespread representation of "The Bag of words" in the document is used a vector of features of which corresponds to a term or phrase in a vocabulary collected from a particular record. The value of each feature element represents the importance of term in the document for a specific characteristic measurement.

In text categorizations a big challenge is learning from high-dimensional data. On the one hand ten and hundreds of thousands of terms in a document can lead to a high computational effort for learning to process. On the other hand, some are irrelevant and redundant functions can affect the predictive power classifiers for categorizing text. To avoid the problem the "Curse of Dimensionality" and to speed up the learning process, it is necessary to perform feature reduction to reduce the size of features.

Typically, in text categorization the filter approach is predominantly used because of their efficiency and simplicity. However, the filter approach evaluates the goodness of a feature by exploiting only the self-serving characteristics of the training data without consideration the learning algorithm for discrimination, the can lead to unwanted classification performance. For a given learning algorithm, the choice is difficult for the best filter approach that generates the functions what the classifier is better for than any other discrimination in terms of theoretical analysis.

Machine learning techniques for automatic text categorization and for this they are using input as a pre-classified documents set and characteristics of categories. Effectiveness of system is very good for document classification. In this text filtering approach is classified steam of documents. Input to the filtering is asynchronous stream of documents. Classification of documents is done into the two categories such as irrelevant and relevant documents. In this they are improving productivities of human classifiers [1].

Mostly machine learning techniques are using for news filtering, personalization, document routing and document categorization. In this they also present the comparison of Twelve Feature Selection Methods and experiment conduct on the Benchmark of 229 text classification. They combines the BNS + F1-measure for best performance on greatest number of tasks considerable margin.

II. Review of Literature

Bo tang et al [1] they proposed a method to automatically classify text by using class specific features based on Bayesian classification. Selecting essential features for every class are allowed. Classification based on Class specific features on the basis of naive bayes rule designed from Baggenstoss PDF projection theorem. The major advantage is use of present feature selection conditions.

B. Tang et.al [2] in this they implemented a new supervised classification technique - the extended nearest neighbor- that guesses input patterns as per the most extreme pick up of intra-class coherence. Dissimilar to the conventional nearest neighbor (KNN) strategy, where nearest neighbors of a test are utilized to assess a group membership, the ENN technique guesses in a"2-way communication" style: it considers closest neighbors of test sample, along with who consider the test as their closest neighbors.

J. J. Patil et al [3] authors proposed a technique for automatically categorize text of Marathi files based on user profile with browsing history of user. Vector Space Model is better over present Probabilistic Models. The precision of the outcomes related to the system is way 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.

F.S. Al Anzi et al [4] 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 for utilizing the stemming in classification of Arabic text.. They also use Arabic text condition to refuse the use of stemming in Arabic text classification

Paul M et al [5] in this the hypothetical establishment class specific feature based system of finding optimal classification is developed 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 every PDF back to raw data space for analysis, M-ary classifier 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 features fulfill adequacy prerequisites exclusively for every class.

H Liu et al [6] 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.

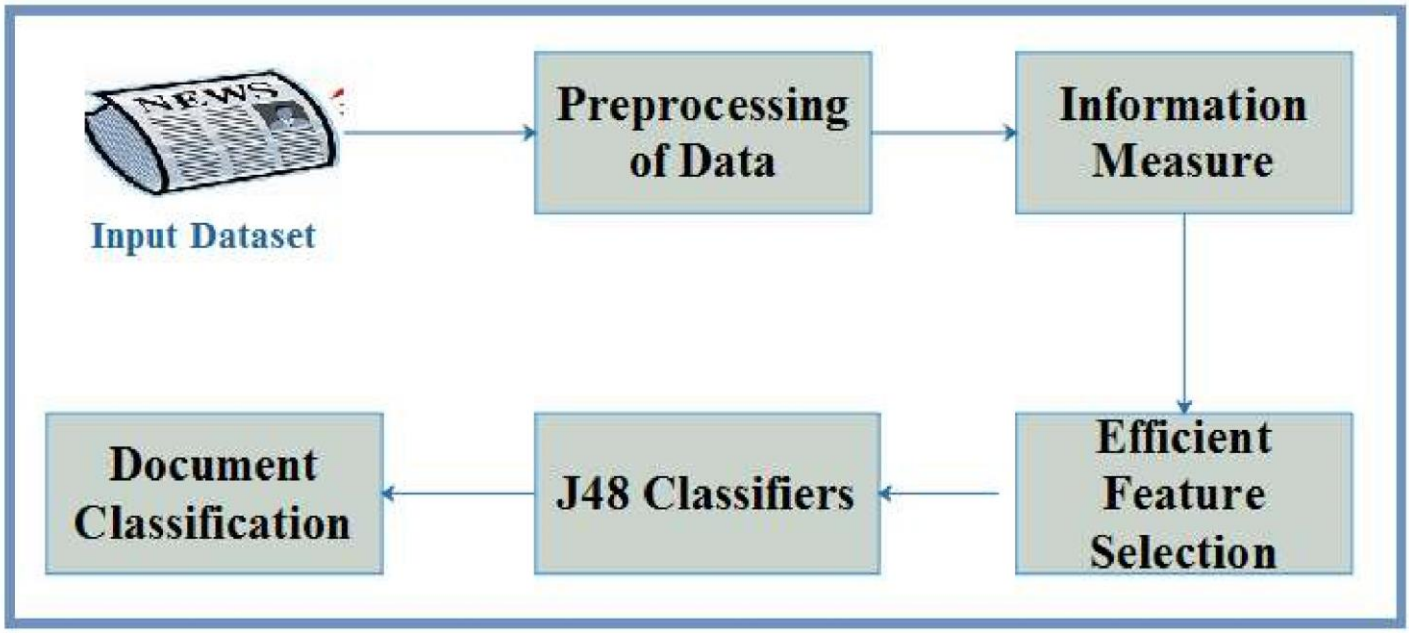
F Sebastiani et al [7] examines the principle method to deal with text categorization that are considered in the machine learning worldview The auto categorization of text in pre-specified classes as seen blasting enthusiasm for the most recent 10 years, because of the expanded accessibility of files 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.

W Lam et al [8] authors developed an automatic text categorization method and its applications for text retrieving. A categorization technique designed using a combination of learning 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. Also use of programmed automated categorization of text retrieval is explored.

L. S. Oh et al [9] they developed 2 concepts feature selection based combination and class dependent features for consolidating multiple features for handwriting recognition. A nonparametric technique is utilized for feature evaluation, as well as on evaluation of features in their class separation as well as recognition capabilities. Further multiple feature vectors are consolidated to generate novel feature vector. The feature has diverse discriminating powers for various classes, another plan of selecting and consolidating class-dependent features is proposed. Also class is considered to have its own optimal feature vector differentiate itself from alternate classes.

P.M Bagenstoss et al [10] they used Bayesian technique for classification needs information of the probability- densityfunction (PDF) of information or enough statistics for all class hypotheses. Hence it is very hard to get a single lowdimensional sufficient statistic, sometimes 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.



III. PROPOSED METHODOLOGY

A. Proposed System Overview

In proposed system we are working on a data set named as 20-NEWSGROUPS dataset which contains the data related to different news category this data set is taken as input to the system and as a output we will get documents classification in several categories (topic) as an output. This proposed system has several phases such as Preprocessing, Information measure, Efficient Feature Selection, J48 Classifier, Document Categories. First, in Preprocessing the Stemming, stop word are removed from the input file, next on the out of the phase the Information measure is performed using this information measure for efficient feature selection. These features are given to J48 classifier for the purpose of classification and at the end we well get the different document for each category.

Figure.1 : System Architecture

Modules of the system are as follows:

- Preprocessing :
  Some preprocessing tasks are usually performed on data before the dataset is used for retrieval. Certain features in documents can be hurt classification performances that are removed using Stemming and Stopwords operations.
  Stopwords: These 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.

- Efficient Feature Selection:
  A Greedy Feature Selection Algorithm Based on the Maximum J-Divergence for Two-class Classification, An Efficient Feature Selection Algorithm Based on the Maximum J-Divergence for Two-class Classification and An Efficient Feature Selection Algorithm Based on the Maximum JMHD in convergence for N-class Classification is used.

- J48 Classifiers:
  After selecting the features using feature selection method, these features are used for classification. Weighted J48 classifier classifies this system. Combining J48 with term weighting gives accurate results.

- Feature Selection:
  Here part of original features is sorted. It is essential in text classification to perform feature reduction. It increases learning speed of classifier. The module takes features as input and feature selection method is applied using this algorithm.

- Document Categorization:
  Finally we got text categorization from high dimensional data.

IV. MATHEMATICAL MODEL A. Mathematical Model

I. Input dataset
D= {d1, d2}
Where, d1 = 20-NEWSGROUPS dataset
d2 = REUTER dataset II.

Preprocessing
Stemming and Stop words are removed from the input file in this step PF = {pf1, pf2... pfn } Where,
PF is the set of preprocessed files belongs to input dataset.
Term Frequency Calculation for data filtering:
TF = $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d. Normalized TF,

$$tf_{t,d} = \frac{n^{i,j}}{\sum_k n_{i.j}} \quad \ldots\ldots\ldots (1)$$

Where, $n_{i,j}$ is the number of occurrences of the considered term (ti) in document dj , and the denominator is the sum of number of occurrences of all terms in document dj , that is, the extent of the document $|d_j|$ III. Feature Selection:
To make this process efficient and to reduce the size of extracted features, following 3 algorithms are used for feature extraction.

FEA = {fea1, fea2, fea3}
Where, FEA is the set of 3 Feature Extraction Algorithms:
FEA1 = A Greedy Feature Selection Algorithm
Based on the Maximum J-Divergence for Two-class
Classification
FEA2 = An Efficient Feature Selection Algorithm
Based on the Maximum JDivergence for Two-class
Classification
FEA3 = An Efficient Feature Selection Algorithm
Based on the Maximum JMHDivergence for N-class Classification

IV. Document Classification:

Documents are classified by using C4.5 classification. It performs n class classification. C = {c1, c2... cn}
Where, C is the set of n classes. Documents are categorized with this class labels.

B. Algorithm

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

Input: Newspaper Datasets for a given training data set with N topics.
Step 1: Start
Step 2: Form a reference class l0 which consists of all documents;
For each class j = 1: N do
Step 3: Calculate the score of each feature based on a specific criteria, and rank the feature with the score in a descending order;
Step 4: Choose the first F features x_i, the index of which is denoted by Jj;
Step 5: Estimate the parameters $\theta_j|0$ under the reference class $l0$ and the parameters θ_j under the class li;
Step 6: End
Output: Given a document to be classified, Output the class label l ˆa using following equation

$$l * \propto_{j \in \{1,2,\ldots\ldots N\}}^{argmax} \sum_{f=1}^{F} x_f log \frac{\vartheta_{nj}{}_{fj}}{\vartheta_{nj}{}_{f0}} + i \quad \ldots\ldots\ldots (2)$$

This algorithm of feature selection for classification attempts to select the minimally sized subset of features by select only those attributes and instance features which are as close to class as possible. It reduces the dimensionality of the dataset.

Algorithm 2: J48 Classifier

J48 classifier is classification algorithm used for detecting class of tuples. To classify the data decision tree methodology is used, it first creates decision tree based on available training data. As the tree is built it is applied to all records in the data and provides the class as an output for that record.

Step 1: Check for the below base cases:

- All the samples in the list belong to the same class.
  When this happens, it simply creates a leaf node for the decisions tree saying to choose that class.
- None of the features provides any information gain.
  In this case, J48 creates a decision node higher up the tree using the expected values of the class.
- Instance of previously unseen class encountered
  Again, J48 creates a decision node higher up the tree using expected values. Step

2: For each attribute a, find the normalized information gain ratio from splitting on a.
Step 3: Let a_best be the attributes with the highest normalized information gain.
Step 4: Create a decision node that splits on a_best.
Step 5: Recur on the sub lists obtained by splitting on a_best, and add those nodes as children of nodes.

V. Result and Disscussion

A. Experimental Setup

The system is constructed using Java framework (version jdk 1.8) on Windows platform. The Netbeans (version 8.0) is used as a development tool. The system does not require any special hardware to run; any standard machine is capable of running the application.

B. Dataset Used

The user provides the dataset that is newspaper dataset with different topics and read the dataset. The dataset 20-NEWSGROUPS collect 20, 000 documents that have been posted online with 20 different topics.

C. Experimental Results

Table.1 describes the time required in ms for classification using naive bayes and J48 classifier. J48 takes less time for text categorization as compared to naive bayes classifier.

Table.1: Time Efficiency Comparison

Algorithm	Time in ms
Existing System Using Naive Bayes	450
Proposed System Using J48	390

Figure.2 shows the time efficiency of naive bayes Vs J48 classifier. Time for J48 takes less time as compared naive bayes. X-axis represents classifiers and y-axis represents time needed in ms.

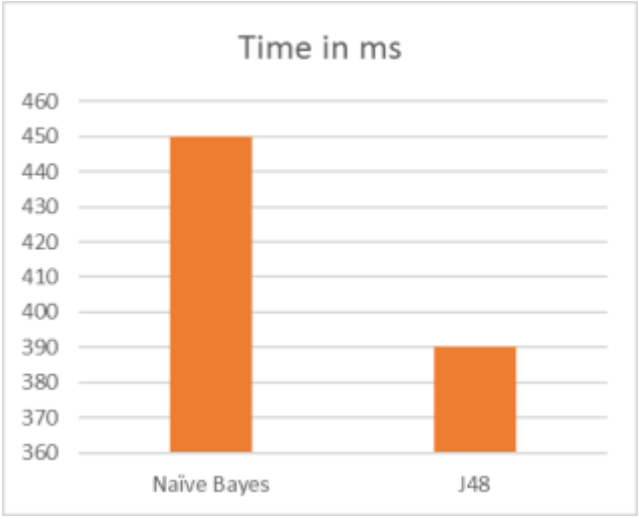


Figure.2 : Time Graph

Table.2 describes accuracy of existing system and proposed system in the form of percentage.

Table.2: Accuracy Comparison

Algorithm	Accuracy in %
Existing System Using Naive Bayes	50
Proposed System Using J48	60

Figure.3 Accuracy of Naive Bayes vs J48 classifier Naive Bayes accuracy is 50 % and weighted J48 is 60%

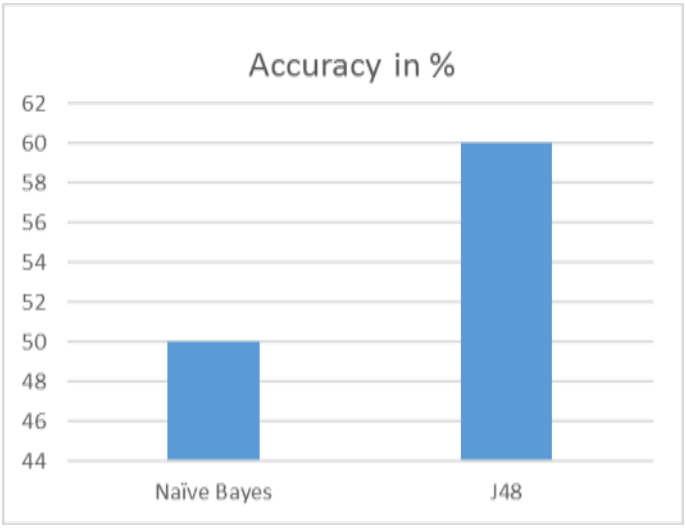


Figure.3 : Accuracy Graph

Table.3: Memory Comparison

Algorithm	Memory in Bytes
Existing System Using Naive Bayes	79
Proposed System Using J48	53

Figure.4 shows the memory comparison between naive bayes Vs J48 classifier. Memory required for J48 is 53 bystes and naive bayes is 79 bytes.

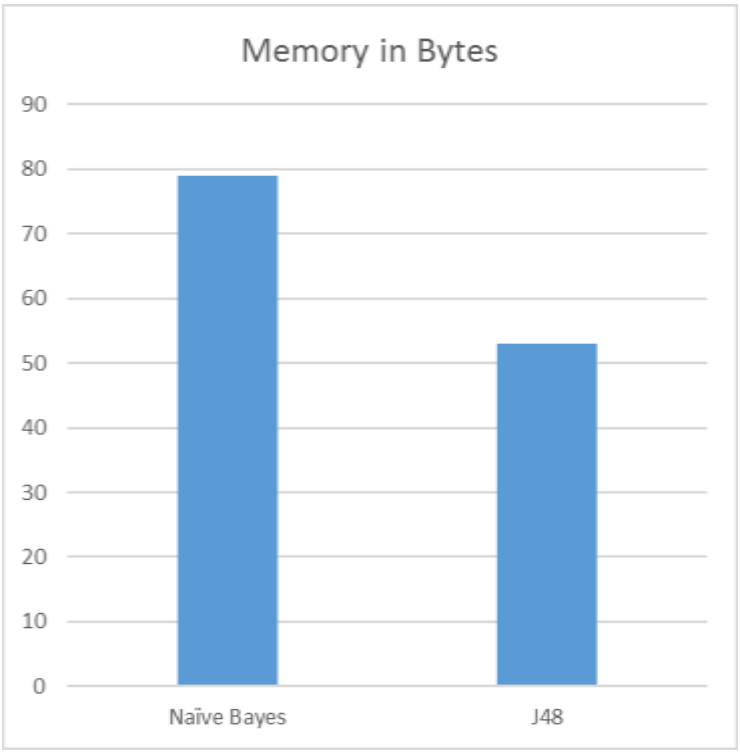


Figure.4 : Memory Graph

VI. CONCLUSION

For applications of information retrieval, machine learning and text mining, text categorization plays vital role. This paper presents the recent text categorization as naive bayes and J48 classifier and feature selection methods used for text document classification. This system also used term weighting 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.

REFERENCES

[1] Bo Tang, Haibo He, Paul M. Baggenstoss, and Steven Kay, "A Bayesian Classification Approach Using ClassSpecific Features for Text Categorization", 1041-4347 (c) 2015 IEEE, Transactions on Knowledge and Data Engineering.

[2] 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.

[3] J. J. Patil and N. Bogiri, "Automatic text categorization: Marathi documents," 2015.

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

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