Holy Grail of Hybrid Text Classification

Rupali Patil[1], R. P. Bhavsar[2] and B. V. Pawar[3]

[1] Department of Computer Science, S. S. V. P. S's L.K. Dr. P. R. Ghogrey Science College Dhule, Maharashtra, India

[2] School of Computer Sciences, North Maharashtra University Jalgaon, Maharashtra, India

[3] School of Computer Sciences, North Maharashtra University Jalgaon, Maharashtra, India

Abstract

Automatic management of ever increasing online digital data is a major challenge for computer science, which can be solved by using machine learning technique called automatic text classification. Automatic text classification is the process of assigning newly arrived text document to one or more predefined categories. Various feature selection and text classification techniques are available in machine learning literature. Various researchers have tried, to improve the accuracy of classification and reduce the time required, by combining different classification techniques and feature selection techniques under them. This generation of new technique by combination of existing techniques is known as Hybrid text classification. This paper aims to focus and discuss our study of the popular feature selection and text classification techniques available in machine learning literature at the same time discusses the various existing hybrid text classification techniques that are applied in the field of text document classification.

Keywords: *Hybrid text classification, Feature Selection, K Nearest Neighbor, Decision Tree, Naive Bayes, Support Vector Machine, Neural Network, Centroid Based Classifier, Vector Space Model.*

1. Introduction

From last few years, digital information is growing more rapidly. In this digital information most of the information is in the electronic textual format. Therefore there is a problem of organization and management of this information. In many organizations like Yahoo trained professionals are manually categorizing newly arrived information. But it is very costly and time consuming. Therefore there is a need to develop an automatic system which can handle this huge amount of digital information. Here text classification plays vital role, to handle this huge electronic data. Automatic text categorization categorizes text information into one or more predefined categories which can be used to extract information for various purpose. As there are so many categories, it is difficult to categorize this huge data, thus there is a need for automated categorization system. Automatic text classification helps us to organize and manage this information with the help of scientific techniques and computational tools. Various applications of text classifications include webpage classification [1], junk email filtering [2], online deception detection [3], Internet abuse detection [4]. Search engines can also use text classification techniques to return more accurate results to the user. To learn classification models automatically, various machine learning techniques based on training examples (training data) are used. Each of these techniques is termed as classifier. Category of new unseen documents (test data) can be predicted using these trained classifiers.

The rest of the paper is organized as follows: Section 2 summarizes Text Classification Process Overview, Feature Selection and some popular Existing Feature Selection Methods; while Section 3 gives Overview of Text Classifiers followed by Review of Hybrid Text Classification in section 4. Section 5 gives Conclusion.

2. Text Classification Process Overview

Automatic text classification falls under supervised machine learning. Phases of Text classification process can be generally divided into two phase's viz. Training phase and testing phase. The Training and testing phases consists of five different steps. These phases are shown in following block diagram (Fig.1.).

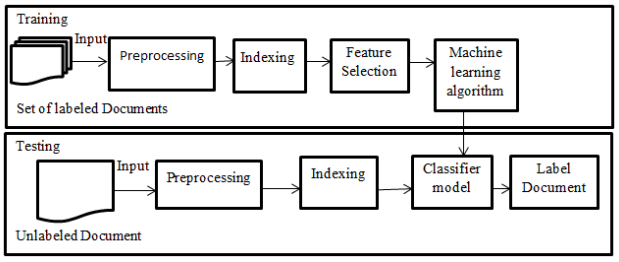

Fig.1 Text Classification Process Block Diagram

Each of the phases and their internal process steps are explained below.

2.1 Document Collection

First phase in classification is document collection. Total collection is divided into training and testing documents. Second phase is document pre-processing which is an important step in text mining. Preprocessing is used to transform text documents into a suitable format for automatic processing. In pre-processing documents are tokenized by removing punctuation marks, special symbols and numbers. The remaining character strings are considered as terms or tokens or features.

2.2 Document Representation

Second phase in classification is indexing. System cannot understand document in its raw form. In text classification system each document is represented as a bag of word. There are various approaches for document representation such as lexical, semantic, and syntactic, but most of the previous classification studies used bag-of-word approach for it, which is one of the straightforward methods of document representation. In this method number of times each term occurs in the document is represented by using vector. Each term in the document is a separate feature. Every different term in document is assigned a special weight which shows its importance in a document. There are different weighting schemes used for text classification or information retrieval, out of them Term Frequency (TF) which counts the number of occurrences of terms in a document. Term Frequency Inverse Document Frequency (TFIDF) in which TF weight is decreased by using the documents in the collection which contains the related term. Thus multidimensional feature vector is used to represent document terms and its weights. This is called Vector Space Model (VSM) where each weighted value for a different term within the document collection is corresponds to separate dimension. It ignores ordering of words or terms but considers number of occurrences of each term. The restriction of VSM is that it does not show the correlation and context of each term, which are essential to understand the document. There are so many text document representation models. From literature, it is clear that for text classification task, frequency-based Bag of Word model gives effective results. Indeed, till date the best multiclass, multi-labeled categorization results for well-known datasets are based on BoW representation [5].

2.3 Dimensionality Reduction

In text classification features increases in large size even for few numbers of documents. Because of this text classification takes more processing time. Also accuracy of text classification decreases due to curse of dimensionality. It faces the problem of overfitting for many classification models. Overfitting means classifier works well for training data set but not work for cases outside training data. Computational cost and overfitting can be reduced by decreasing dimensionality. We can achieve dimensionality reduction by using three basic methods which are stopword removal, stemming and feature selection.

2.3.1 Stopword Removal

Stopwords are meaningless words which are occurring frequently, such as articles, prepositions, conjunctions. These words are not very useful for describing different categories of documents.

2.3.2 Stemming

Stemming is the process of removal of suffixes and prefixes and finding root word. Words with common root have same meaning and can be grouped into a single term through stemming process. For example by stemming terms compute, computes, computed, computing, computation, computations can be group into a single term compute. Thus by eliminating features with less discriminative powers or by removing irrelevant features, feature dimensions can be reduced. Stopword removal and stemming both are language dependent ways of dimensionality reduction.

2.3.3 Feature Selection

Another way for dimensionality reduction is feature selection also known as attribute selection which is the process of finding most relevant features for classification. The filter approach and wrapper approach are two general feature selection approaches. In wrapper approach a classifier build by using the learning algorithm is wrapped into the attribute selection procedure, so that based on different subset of attributes multiple classifiers can be generated and select the subset which gives the best performance. Because of the complete size of space of attribute subsets the wrapper approach become cost prohibitive so text classification are often forced to settle for the filter approach [6] . Although feature dependencies are considered in wrapper approaches and interaction between learning mode and feature subset search are provided, wrapper approaches are more costly than filter approaches. The wrapper approaches are dependent on learning algorithm whereas filter approaches are independent of learning algorithm. Although better subset of features could be find using wrapper approach, it is computationally expensive and is not suitable when the numbers of features are very large. Fig. 2 shows feature selection using wrapper approach.

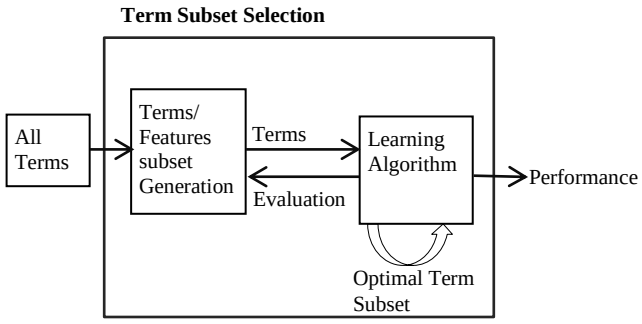

Fig. 2 The feature Wrapper approach

In filter approach by using some relevance measure features are evaluated and filtered without using any learning algorithm. Feature transformation or feature selection can be used in filtered approach for dimensionality reduction. The original feature space is estimated into a lesser dimensional subspace to carry more relevant or discriminative information, in feature transformation. Latent semantic analysis, linear discriminant analysis, principal component analysis are some of the most common feature transformation techniques used for text classification, whereas the original feature space is considered while applying feature selection. Aim of feature selection is to eliminate irrelevant dimensions and with their original values relevant dimensions are kept. Fig. 3 shows feature selection using filter approach.

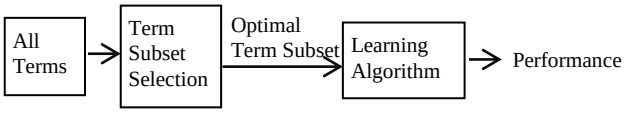

Fig. 3 The feature filter approach

Text clustering is another method for text data [7] in which words with high degree of pairwise semantic relatedness are grouped so that rather than using individual terms, groups may be used as dimensions of the vector. In literature generally individual filter or wrappers approach is applied. But in many studies filter and wrapper approaches were applied together. In feature selection scheme evaluation of features can be take place by univariate and multivariate approach. Features are independently examined in univariate approach. Univariate approach is fast and it offers individual discriminative powers of the features. But possible correlation among features is ignored in it. On the other hand, during feature evaluation feature dependencies are considered in multivariate approach but this approach is slow as compared to univariate approach. In text classification, there are large numbers of features and it requires large processing time and therefore univariate filter approaches are widely used in feature selection methods. After obtaining discriminative power of individual feature the best N features are selected and others are removed. Thus the compact subset of features is obtained although feature dependencies are ignored. Document frequency, mutual information, Chi-square, term strength, odd ratio and information gain are most popular examples of this approach. In literature numbers of comparative studies on feature selection metrics are found.

2.4 Existing Feature Selection Methods

As mention in previous section, in text classification, there are large numbers of filter based techniques for the choice of unique features. It is found from literature that document frequency, mutual information, chi square, information gain, Gini index have been confirmed to be much more popular techniques. The following subsection offers the mathematical background of these methods.

2.4.1 Chi-square (CHI2)

Chi-square is well known statistical feature selection method. Chi-square test in statistics is used to observe independence of two events. A and B events are assumed to be independent if

$$p(AB) = p(A)p(B)$$

These two events related to occurrence of particular term and class, respectively, in text feature selection. One can compute chi-square information using

$$\text{Chi} - \text{square}(t, C) = \sum_{t\in\{0,1\}} \sum_{C\in\{0,1\}} \frac{\left(N_{t,C} - E_{t,C}\right)^2}{E_{t,C}}$$

For each state of term t and class C, N is the observed frequency and E is the expected frequency [8]. Chi-square is used to measure how much expected counts E and observed counts N deviate from each other. High score of chi-square indicates that the occurrence of the term and class is dependent. In text classification, we can select the feature if they are dependent. Chi-square score of a term is calculated for individual classes. There are two ways to globalize this score over all classes. The first way is for all classes calculate weighted average score whereas the second way is among all classes select the maximum score. To globalize chi-square score for all classes, in this paper, the first approach is chosen

$$chi - square(t) = \sum_{i=1}^{M} P(c_i).chi - square(t, c_i)$$

where P(c_i) is the class probability and chi-square(t,c_i) is the class specific chi-square score of term *t*.
Monica Rogati and Yiming Yang in [9] performed controlled study on large number of filter feature selection methods for text classification. Their study show that chi-square statistic based feature selection methods consistently outperform other methods based on other criteria (including Information gain, document frequency) for NB, Rocchio, KNN and SVM classifiers.

2.4.2 Information Gain (IG) [10]

To make the correct classification decision, on any class, the presence or absence of a term donates how much information is measured by using information gain. For class relationship, if a term is an ideal indicator then information gain reaches its maximum value, that is, the document fit into the corresponding class if the term is present in it. We can get information gain for term t by using

$$IG(t) = -\sum_{i=1}^{M} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{M} P(c_i|t)\log P(c_i|t) + P(\bar{t})\sum_{i=1}^{M} P(c_i|\bar{t})\log P(c_i|\bar{t})$$

Where *M* is the number of classes, $P(c_i)$ is the probability of class *ci*, *P*(*t*) and $P(\bar{t})$are the probabilities of presence and absence of term *t*, $P(c_i|t)$ and $P(c_i|\bar{t})$ are the conditional probabilities of class c_i given presence and absence of term *t*, respectively.

2.4.3 Gini Index (GI)

Gini Index [11] [12] is a non-purity split method. It is another feature selection method which is an improved version of the method originally used to find the best split of attributes in decision trees [13]. It has simpler computation than the other methods in general [14]. Its formulation is given as

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2 + P(C_i|t)^2$$

where $P(t, C_i)$ is the probability of term *t* given presence of class C_i, $P(C_i|t)$ is the probability of class C_i given presence of term *t*, respectively.

2.4.4 Document Frequency (DF)

DF [15] is the number of documents containing term t. It is defined as

$$DF = \sum_{i=1}^{m} (A_i)$$

Where A_i: the number of the documents that contain term t and also belong to category c_i;
Low document frequency terms are often referred to as rare terms and high document frequency terms are often referred to as common terms. This feature selection method assumes that higher document frequency terms are more informative for classification. But this assumption does not true every time, for example, the stop words (e.g., a, an, the) hold very high DF scores, but they rarely contribute to classification. In general, this simple method performs very well. It defines the relevancy score of each term.

2.4.5 Mutual Information (MI)

Mutual information [15] is one of the common feature selection methods. Mutual dependencies of two random variables can be indicated by mutual information. Mutual information of term t and class c describes how much information is present or absent of a particular term t gives to make the correct classification decision on relevant class c. Therefore, MI can be calculated as:

$$MI(t, c) = \log \frac{P(t|c)}{P(t)}$$

Where $P(t|c)$ is the probability of term t given class c and $P(t)$ is the probability of term t.
K Raghuveer and Kavi Narayan Murthy [16] showed that Mutual Information is an effective language independent dimensionality reduction technique.

3. Overview of Text Classifiers

A text classifier is a model which allocates new unknown text document to predefined classes depending on contents of the document. Following subsection offers explanation of commonly used classifiers in the field of text classification.

3.1 K Nearest Neighbor (KNN)

KNN is an easy to understand and easy to implement classification technique. Although it is simple it can perform well in many situations. It is very efficient instance based algorithm for text categorization. It is a supervised learning algorithm. KNN works as follows
For a test document to be classified, first determine parameter K – number of nearest neighbor then calculate the distance between test document and all the training

samples, using some distance measures like Euclidean distance or using some similarity measures like cosine similarity. Sort the distances of all training samples to the test document and determine nearest neighbors based on K^{th} minimum distance if distance function is used. Once the nearest- neighbor list is obtained, the test document is classified based on the majority of its nearest neighbors.
Dr. Riyal Al-Shalabi, Dr Ghassan Kanaan and Manaf H. Gharaibeh [17] implemented KNN algorithm and applied to a dataset of 621 Arabic text documents. Their result shows that KNN is applicable to Arabic Text. They got 0.95 Micro Average precision and recall score.

3.2 Decision Tree

Decision tree is a simple yet widely used classification technique. It is inductive learning method. There are two basic steps in the technique. Building the tree and applying the tree to the database. In decision tree classifier terms are used to label internal nodes, weights of terms present in test document are used to label branches passing from them, and categories are used to label leaf nodes. For categorization of test documents d_j weights of the terms used for labeling the internal nodes are recursively tested up to a leaf node occur. This node label is then allocated to d_j. It has been used either as the main classification tool or as baseline classifier.
Fouzi Harrag, Eyas El-Qawasmeh, Pit Pichappan in [18] evaluated classification system based on decision tree algorithm. They performed their experiment over two self-collected data corpus. Comparison of their results with NB, Maximum Entropy and Vector space Model showed that decision tree based classification system is more accurate system in terms of global performance.

3.3 Naïve Bayes

NB is simple probabilistic classifier. It is widely used for document classification and shown to produce very good performance. It is based on Bayes' Theorem with strong independent assumptions which assumes all of the features are mutually independent, that the presence or absence of one feature does not affect the presence or absence of another feature.
Suppose there are total k categories denoted as C = $\{C_1, C_2, \dots, C_k\}$ in a training sample set. An arbitrary document d_i has p independent feature denoted as d_i = $\{X_{i1}, X_{i2}, \dots, X_{ip}\}$. Given a feature word X_i the probability that a related document, d_i, is in class C_j is described by $P(C_j|X_i)$. Training data can be used to determine $P(X_i)$ for each feature word, $P(X_i|C_j)$ and prior probability $P(C_j)$ for each class. From these values using Bayes Theorem we can calculate posterior probability $P(C_j|X_i)$ and then $P(C_j|d_i)$. To classify a target document the conditional and prior probabilities are generated from training set and used to make the prediction. This is done by combining the effect of different feature words from the document. Then we estimate $P(d_i|C_j)$ by

$$P(d_i|C_j) = \prod_{k=1}^{p}(X_{ik}|C_j)$$

Then posterior probability for each class is calculated as

$$P(C_j|d_i) = \frac{P(d_i|C_j).P(C_j)}{P(d_i)}$$

After that the class with highest posterior probability is chosen for the target document.
Ajay S. Patil and B. V. Pawar [19] applied NB approach to classify web sites based on the contents of their home pages and they have obtained 89.05% accuracy.

3.4 Support Vector Machine (SVM)

SVM is supervised learning algorithm. It is based on statistical learning theory. It is extensively and successfully used for text classification. This machine learning technique is introduced by Vapnik [20]. Based on statistical learning principle known as structural risk minimization, SVM classifier searches for a decision boundary which separates the training data points into two classes known as support vector. There are many such hyperplanes (Support Vector) possible. The classifier must choose decision boundaries with largest margin.
Consider a binary classification problem consisting of N linearly separable training examples. Example is denoted by a tuple (X_i, Y_i) (i= 1, 2,..., N) where X_i=$(X_{i1}, X_{i2}, .., X_{id})^T$ corresponds to the attribute set for i^{th} example. Let $Y_i \in \{-1, 1\}$ denote its class label. The decision boundary that bisects the training examples into their respective classes can be written in the following form

w.x+b=0

$$\text{And} \quad \begin{cases} w.x_i + b \geq 1 & if\ y_i = +1 \\ w.x_i + b \leq -1 & if\ y_i = -1 \end{cases}$$

Where i=1, 2,...,N where the dot product operation (.) is defined by

w.x= $\sum_i w_i x_i$

For vector w & x SVM imposes an additional requirement that the margin of its decision boundary must be maximal. Maximizing the margin is equivalent to minimizing the following objective function.

$$f(w) = \frac{\|w\|^2}{2}$$

Subject to

$$w.x_i + b \geq 1 \quad if\ y_i = +1$$
$$w.x_i + b \leq -1 \quad if\ y_i = -1$$

For i=1, 2… N

The method described is applicable to the case in which the positives and negatives are non-linearly separable.
Tarek Fouad Gharib, Mena Badieh Habib and Zaki Taha Fayed [21] applied a Support Vector Space model for Arabic text documents classification and compared with traditional Bayes Classifier, KNN classifier and Rocchio classifier. Their experimental result shows that when feature set size is large enough SVM outperformed the other classifiers.

3.5 Neural Network

Neural Network has been applied in document classification system to improve efficiency. A neural network classifier is known as a network of nodes (neurons), where input nodes represent terms; the output nodes represent the category or categories of interest. Each input node is connected via a weighted link to the output node. The weight on the edges represents dependence relations. For classifying a test document d_j, the input units are assigned by its term weights Wk_j, through the network, these units activation is propagated forward and the categorization decision(s) is determined by the value that the output unit(s) takes up.
Generally back propagation is used for training neural networks, whereby, into the input units the weights of term of a training document are loaded, and if misclassification occurs the error is back propagated so that the parameters of the network can be changed and remove or reduce the error.
The simplest model of neural network is the perceptron. It is linear classifier. It is a single layer, feed-forward neural network. Another complex structure of neural network is multilayer neural network. It is a nonlinear neural network. Fouzi Harrage, Abdul Malik Salman, Al-Salman and Mohammed BeMohammed [22] presented a Neural Network Based Model for Arabic text classification. Multilayer perceptrons (MLP) and Radial Basis Function (RBF) NN based learning techniques are used in their model. Their study concludes that MLP classifier is better than RBF classifier. Their study also showed that Single Value Decomposition (SVD) supported NN classifier outperforms the basic NN.

3.6 Centroid Based Classifier

It is a very simple classification algorithm. In this algorithm, centroid vector for each class in a set of classes is calculated using a set of documents $\{d_1, d_2,\ldots,d_n\}$ belonging to that class. Suppose, there are m classes then there are m centroid vectors $\{c_1, c_2,..,c_m\}$ and each centroid vector c_i represents centroid of corresponding class. For calculation of centroid there are two methods

1. Arithmetical Average Centroid (AAC): Most commonly used initialization method for centroid based classifier $\overrightarrow{ci} = \frac{1}{|ci|}\sum_{\vec{d}\in ci}\vec{d}$ where centroid is the arithmetical average of all document vectors of class c_i
2. Cumuli Geometric Centroid (CGC): $\overrightarrow{ci} = \sum_{\vec{d}\in ci}\vec{d}$ where each term will be given a summation weight.

This centroid vector can be used to classify new test document. To determine the class of new document j calculate similarity of document j with centroid of each class and based on this similarity, assign the new document j, to the class having maximum similarity centroid. We can use any similarity measure like cosine similarity, Jacarred similarity, dice similarity, inner similarity etc. The computational complexity of this algorithm is very low and is identical to fast document classifiers such as Naïve Bayesian.
In our previous study [23] we had implemented and compared NB, KNN and Centroid Based classifiers on R-52 of Reuters-21578 standard dataset. Our experimental results showed that Centroid Based classifier outperform among three classifiers.

3.7 Vector Space Model

Vector space model was proposed in the late 60s by Gerald Salton et. al. [24] to represent text by vector. VSM is most familiar method of the document weighting approaches. It implements Bag of words representation in which co-relation of adjacent words is not considered. In IR and text mining generally tf*idf weighting is used to find the importance of a word to a document.
In vector space model document d_k is represented by a set of terms $(t_1, t_2,.., t_n)$ where each t_j is a term that appears in text document and n is the total number of terms in the documents. After weight calculation each document is represented as a specific n dimensional vector d_k as

$d_k = (w_1, w_2,.., w_n)$

Once the document is represented as term weight matrix we can apply any one of the distance measures such as Euclidean distance to find distance of documents or similarity measure such as cosine similarity to find the similarity of documents. Based on distance measure or similarity measure category with minimum distance or maximum similarity is assigned to test document.
The Euclidean distance of two vectors can be computed as

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{n}(x_i + y_i)^2}$$

Cosine measure or normalized co-relation coefficient

$$cosine(\vec{x},\vec{y}) = \frac{\vec{x}.\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2}}$$

Smruthi Mukund and Rohini K. Shrihari [25] have developed Urdu language classifier that can distinguish subjective sentences from objective sentences. The result of their experiment based on SVM and VSM algorithms shows that their modified VSM based approach is more suitable as a sentence level subjectivity classifier.

4. Review of Hybrid Text Classification

BPSO (Binary Particle swarm optimization)-KNN

In Document classification process, feature selection plays an important role to improve accuracy and effectiveness of classifications. Therefore the selection of correct features indicates accurate classification with appropriate computational efficiency, so Hamouda K.Chantar and David W. Corne [26] proposed BPSO (Binary Particle swarm optimization)-KNN as a feature selection method for Arabic text classification. Three Arabic datasets (Alj-News, Alwatan, Akhbar-Alkhaleej) were used to test this method, and three well-known machine learning algorithms – SVM , Naïve Bayes and C4.5 decision tree learning (in its Weka implementation as J48) – were used to classify Arabic documents using features selected by this method. Their results suggest that the proposed method is effective for Arabic document classification. BPSO-KNN +SVM get highest accuracy with 96.1% on Alwatan dataset.

MIIEP_FS (Mutual Information and Information Entropy Pair)

For the same task i.e. to reduce the redundant features in the process of selecting appropriate features and keep the mutual information value, Zhili Pei, Yuxin Zhou, Lisha Liu, Lihua Wang, YinanLu [27] discuss feature selection method (MIIEP_FS) based on a mutual information and information entropy pair which is based on information entropy pair concept and theory of information entropy. By using feature selected by mutual information method this method find out the classification effect and demonstrate the difference between features selected by this method and the features selected based on information entropy. From their experiment they show that the proposed method MIIEP FS (87.4% with 100 features) is more effective than MI (85.7% with 200 features) and CHI (85.7% with 800 features) methods. By using MIIEP_FS method Macro Fl degrees of Naive Bayes and KNN methods are higher; sometime it is higher than support vector machines. This approach not only considers the role of feature to classification, but also gives consideration to the redundancy between features.

Filter + Wrappers

As mention earlier, text classification requires large attribute space and for that it requires an efficient and effective attribute selection procedure. Best attribute subset selection is more effective using wrapper approach as compare to filter approach, but for most of the text classification applications it is very costly. So Chen-Huei Chou, Atish P. Sinha and Huimin Zhao [4] proposed an efficient and effective hybrid attribute selection approach for text classification problems and the proposed approach was applied by them for detection and prevention of Internet abuse in the workstation. The empirical evaluations they conducted using a variety of classification algorithms (NB, MNB, NN, SVM and J4.8), indexing schemes and attribute selection methods (filters (IG, GR, χ^2), wrappers, hybrid (filter + wrappers) demonstrate the utility of the proposed approach. They found that by combining the filter and wrapper approaches accuracies of text classifiers were improved also the computational costs were significantly reduced.

In text classification problem there are thousands of features in initial feature set. To select proper features from initial features set is an important task. In the literature, no major work is done which can find effective features combination, carefully chosen by different selection methods, under different conditions. So by considering both wrapper and filter feature selection, Serkan G¨UNAL [11] proposed a hybrid feature selection strategy. Filter methods DF-, CHI2-, MI-, and IG-based were used for selecting features in first stage, they combined features selected by filter methods and then in next stage, they passed it to wrapper method GS (genetic algorithm (GA)-based selection (GS)) to do the study of text features selected by different methods using different datasets, different classifiers and different measures. They tested their selected features obtained by hybrid approach with SVM and Decision Tree classifiers. Their experimental results show that as compare to features selected by single selection method, combinations of features selected by various feature selection methods are more effective. The features combinations are influenced by dataset used selection of classification algorithm and selection of measure used. The highest Micro-F1 score and Macro-F1 score obtained by the filters when applied to Reuters dataset are (84.89%) and (61.63%) respectively. After applying hybrid selection to the same dataset Micro-F1 is improved to 85.83% which is attained by using SVM and Macro-F1 is improved to 66.19% which is achieved using Decision Tree respectively. Similarly, for the Newsgroups dataset Micro-F1 score and Macro-F1 score obtained by the filters are (98.00% and 98.01%) and after applying hybrid selection it is increased to 98.48% and

98.44%, both are obtained by using Decision Tree classifier.

In text classification problem the number of features can easily raise to hundreds and thousands. This poses a big difficulty in applying many sophisticated learning algorithms to text classification. Feature reduction methods are included to reduce the number of features either by selecting the original features or transforming the features into new features as some functions of existing ones.

Enhanced Naïve Bayes Classification Technique with Genetic algorithm

Naïve Bayes Classification is the simplest and effective method for text classification. However, simple Naïve Bayes method degrades the performance of accuracy due to too many irrelevant features. So, to improve classification accuracy by selecting relevant feature, Ms. K. Umamaheswari, Dr. S. Sumathi, Ms. V. Aparna and Ms. A. Arthi [28] present an enhanced Naïve Bayes classification technique with Genetic algorithm to select a subset of more relevant features, which overcomes this problem. The experimental results are tested using the Mini-Newsgroups dataset with 20 categories, with 100 documents in each category taken from UCI machine learning repository. The classification accuracy achieved is more than 95% with all categories using proposed approach compared to simple Naïve Bayes with Chi – Square statistics. The computation time decreases by 50% in the proposed approach compared to the Naïve Bayes approach with Chi–Square statistics.

Naïve Bayes and Rough set

For information retrieval Vidhya K .A and G. Aghila [29] proposed a hybrid model using Naïve Bayes and Rough set theory for text document classification. They improved the classification accuracy of Naïve Bayes by overcoming the impression and vagueness of dataset. They used Rough set theory for feature reduction whereas for classification of documents into predefined categories they used Naïve Bayes theorem.

CHI-Square and CF method (CHCFW)

The drawback of CHI-Square statistic method is that it considered only the dependency between category and feature. But it doesn't consider whether it is a positive or negative dependency. Also Bayes model gives equal weight to each feature without considering the distribution of a feature. So to remove the drawback of CHI-Square and to improve the effects of Naïve Bayes Yaying Qiu, Guangming Yang and Zhenhua Tan [30] construct an extended model which assigns weights to some important features. To select features and assign weight to important features they proposed a method known as CHCFW which combines the CHI-Square statistic method and CF method. Based on new proposed method CHCFW they select best features by improving the importance of positive features and removing the distribution of negative features. After comparing with original Naïve Bayes method and other algorithm to assign weight to features the experimental results show that the performance of CHCFW method is better and it is more suitable to larger amounts of training documents.

DTE-(Discriminative Term Extraction) based NBKNN model:

In order to reduce learning time and increase the computational efficiency with high accuracy of multi-class, single label text categorization Abhishek Sanwaliya, Kripa Shanker and Subhas C. Misra [31] designed DTE-(Discriminative Term Extraction) based NBKNN model. Their experiment result on Reuters-21578 shows that in accuracy NBKNN(90.6%) out performs the NB(85.97), DT(83.00%), Rocchios(71.54%), KNN(77.30%) classifiers on precision, recall and F1. Result suggests that the performance of classification method significantly depends on characteristics of data also suggest that combining different classification method can increase the consistency, reliability and accuracy of the categorization.

Tree Augmented Naïve Bayes (TAN)

For text classification Naive Bayesian classifier is commonly used and most effective method, but it is not able to express the dependence among text terms due to its attribute independence assumption. Therefore for text classification Shi Hong-Bo, Wang Zhi-Hai, Huang Hou-Kuan, and Jing Li-Ping in [32] propose a method which combines the ease of Naive Bayesian with the ability to express the dependence among text term in Bayesian network known as TAN (Tree Augmented Naïve Bayes) model. In this paper authors present some existing text methods review. Also, TAN model is introduced by them and this TAN model is applied to text classification. They compare Naïve Bayes classifier with TAN classifier and showed that performance of TAN classifier is better.

Mahalanobis distance based KNN (MDKNN)

Although KNN is oldest and simplest method of text classification, it is limitedly used in text classification and can cause low classification accuracy. KNN is sensitive to the distance (or similarity) function used in classifying a test instance (Jahromia et al., 2009). So for text classification area Suli Zhang & Xin Pan [33] present Mahalanobis distance, and based on this theory they suggested (MDKNN) algorithm. The experimental result shows that for text classification, performance of their MDKNN method is better than KNN Classifier and Naïve Bayes classifier.

Support Vector Machine and Maximum Entropy

Yu-Qiang Feng and Wei Jiang [34] proposed a technique for document classification in which maximum entropy and machine learning techniques were combined. In first step to extract the features effectively they offered the cross entropy and average mutual information methods. In second step to do the classification task they applied the support vector machine and maximum entropy model on

the selected features. Beside this rather than using binary features they used real-values as information of each word is helpful for document classification. After comparison of their method with traditional method they found that their method increased 2.78 % F-measures than basic ME model, and 0.95% than Naïve Bayes model.

Euclidean-SVM

For high accuracy of the SVM classifier it is necessary to decide the suitable kernel function and the suitable value of parameter C. To solve this problem Lam Hong Lee, Chin HengWan, Rajprasad Rajkumar and Dino Isa [35] implemented a new text document classification framework known as Euclidean-SVM in which SVM is used in training phase and for classification phase Euclidean distance is used. During training phase, using SVM the support vectors are identified for each category. In classification phase, using the Euclidean distance function the average distances between the support vectors from different categories and the new data point are calculated. The new data point is assigned to the category of support vector which has lowest average distance. Using many text datasets they tested their proposed framework and showed that Euclidean-SVM text classifier's accuracy has a low effect on the kernel functions implementation and soft margin parameter C.

SVM-NN

Although KNN is commonly used text classification algorithm, the accuracy of the KNN classifier is highly depends on the value of parameter k. KNN uses whole training data until classification, so it is known as lazy learning method. Therefore to reduce the parameters impacts on classification accuracy of KNN, Chin Heng Wan, Lam Hong Lee, Rajprasad Rajkumar and Dino Isa [36] proposed the SVM-NN hybrid classification approach by joining the training algorithm support vector machine (SVM) with the K-nearest neighbor (KNN) classification approach. Using SVM they reduce the training samples for each category to their support vectors (SV's) in training phase. These SVs of different categories are act as training samples. In testing phase, they find average distance between testing data point and SV's of different categories using Euclidean distance function. Then testing data point is assigned to the category with average minimum distance between its SVs and testing data point. They performed experiment using different datasets. The experimental result shows that as compare to conventional KNN classifier the SVM-NN's classification accuracy has small effect on the value of parameter. Among all dataset highest classification accuracy achieved by the SVM-NN approach is 94.17% with Vehicles dataset & parameter C is set to the values of 1000, 10,000 and 1,00,000.

Support Vector Machine and Decision Trees

In the past it has been shown that the result of the classification could improve by combining classifiers. So for multi class text classification Srinivasan Ramaswamy [37] made discussion on combination of Support Vector Machine and decision trees. At every level of the tree Support Vector Machines are trained for every class and the more successful SVM for predicting a class at that level is carefully chosen as the choice in that node. Thus in each node a tree is constructed with different SVM. And for classifying the multiclass text, the tree constructed is used. Author performed his experiment with Weka. He obtained 27.2641% accuracy with Naïve Bayes, 22.1205% accuracy with Ibk, 30.4512% accuracy with SVM and 42.64% accuracy with his Decision Tree based SVM classifier. Results had shown that as compare to the other classifiers like Naïve Bayes, Ibk and simple SVM his Decision Tree based SVM method perform better.

Homogenous Centroid-Based Classification

For classification of texts into a set of pre-defined classes commonly used supervised approach is Centroid-based text classification. In this approach a prototype class vector for each class is constructed using important terms in documents and its performance depends on how to weight and select these important terms. From literature it was found that one could use statistical term distributions to improve classification accuracy. However, the best weighting systems are different for different data sets. By considering this problem, Verayuth Lertnattee & Thanaruk Theeramunkong [38] suggested homogenous centroid-based classification method using multiple centroid-based classifiers with different term weighting systems which improves classification accuracy. Model selection and score combination are two factors considered by this method. Four different data sets (Drug information (DI), 20-Newsgroup, WebKB1, WebKB2) are used to test this model. The results showed that classification accuracy could improve by the system up to 7.5-8.5% as compare to k-NN classifier, as compare with naïve Bayes classifier it is 3.7-4.0% and 1.6-2.7% over the best single-classifier ($p < 0.05$).

5. Conclusion

Through this paper we have discussed the nuts and bolts of automatic text classification, including its definition, process overview with its phases and their internal steps. Thorough discussion on various popular Feature selection and Text Classification techniques is presented, these include feature selection techniques like Chi-square, Document Frequency (DF), Information Gain (IG), Mutual Information (MI), Gini Index (GI) and different classification methods like K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM) and Neural Network (NN). Since every feature selection method and Text Classification method has its own performance threshold due to their merits and demerits, research scholars have tried to fuse these

methods to develop a hybrid technique, which yields improved performance by utilizing their merits and complimenting each other. We have also presented the study of various hybrid feature selection techniques and hybrid text classification approaches.

Under feature selection hybrid techniques we have presented BPSO-KNN, MIIEP_FS, cross entropy and average mutual information methods, Naïve Bayes with Genetic algorithm, Naïve Bayes and Rough set theory, CH-CFW hybrids.

We have also presented study of hybrid text classification approaches such as NBKNN, TAN, Euclidean-SVM, MDKNN, SVM-NN, DT based SVM.

Our study revealed that in existing hybrid approaches studied by us Serkan G¨UNAL got highest 98.48% Micro-F1 score and 98.44% Macro-F1 score on Newsgroups dataset by using his proposed hybrid feature selection strategy in which he combined filter methods with wrapper method and both are obtained by using Decision Tree classifier.

From the results of previous studies it is observed that by applying hybrid approaches one can improve the efficiency and accuracy of text classification.

Acknowledgments

We hereby acknowledge the financial and administrative support extended under SAP (DRS-I) scheme, UGC New Delhi at School of Computer Sciences, NMU, Jalgaon.

References

[1] Chen, R.-C., and Hsieh, C.-H., "Web page classification based on a support vector machine using a weighted vote schema", Expert Systems with Applications (31:2) 2006, pp 427-435.

[2] Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C., and Stamatopoulos, P. "A memory-based approach to anti-spam filtering for mailing lists", Information Retrieval (6:1) 2003, pp 49–73.

[3] Zhou, L., Burgoon, J.K., Twitchell, D., and Qin, T. "A comparison of classification methods for predicting deception in computer-mediated communication", Journal of Management Information Systems (20:4) 2004, pp 139-166.

[4] Chen-Huei Chou, Atish P. Sinha, Huimin Zhao, "A Hybrid Attribute Selection Approach for Text Classification", Journal of the Association for Information Systems, Vol. 492 11 Issue 9, September 2010, pp. 491-518.

[5] Bekkerman, R., Allan J.: Using Bigrams in Text Categorization. CIIR Technical Report, IR-408 (2004).

[6] Sebastiani, F. "Machine learning in automated text categorization, "ACM Computing Surveys (34:1) 2002, pp 1-47.

[7] Baker, L.D., and McCallum, A.K. "Distributional Clustering of Words for Text Classification,", Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998, pp. 96-103.

[8] Manning, C. D., Raghavan, P., & Sch¨utze, H. Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK, 2008.

[9] Monica Rogati and Yiming Yang, "High-Performing Feature Selection for Text Classification", CIKM '02 Proceedings of the eleventh international conference on Information and knowledge management, ACM NewYork,NY,USA, 2002 pp.659-661 ISBN:1-58113-492-4 doi: 10.1145/584792.584911

[10] Charu C. Aggarwal, ChengXiang Zhai "A Survey of Text Classification Algorithms" link.springer.com/chapter/10.1007%2F978-1-4614-3223-4_6

[11] Serkan G¨unal, "Hybrid feature selection for text classification", Turk J Elec Eng & Comp Sci, Vol.20, No.Sup.2, 2012, _c Tu¨Bi˙Tak, doi: 10.3906/elk-1101-1064.

[12] Wenqian Shang, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu, Zhihai Wang, "A novel feature selection algorithm for text categorization", Expert Systems with Applications 33 (2007) 1–5, 0957-4174/$, 2006 Elsevier, doi:10.1016/j.eswa.2006.04.001

[13] Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z, "A Novel Feature Selection Algorithm for Text Categorization", Expert Systems with Applications, 33(1), 2007, 1-5.

[14] Ogura, H., Amano, H., & Kondo, M., "Feature selection with a measure of deviations from Poisson in text categorization", Decision Support Systems, 36(3), 2009, 6826-6832.

[15] Yan Xu, Gareth Jones, JinTao Li, Bin Wang, and ChunMing Sun, "A Study on Mutual Information-Based Feature Selection for Text Categorization", Journal of Computational Information Systems, (2007) 3 (3). pp. 1007-1012. ISSN 1553-9105.

[16] K. Raghuveer and Kavi Narayana Murthy, "Text Categorization in Indian Languages using Machine Learning Approaches", IICAI 2007: 1864-1883.

[17] Al-Shalabi Riyad, Kanaan Ghassan, Gharaibeh Manaf H. (2006), "Arabic text categorization using kNN algorithm", Proceedings of the 4th international multi conference on computer science and information technology (CSIT 2006), Volume 4, Amman, Jordan.

[18] Fouzi Harrag, Eyas El-Qawasmeh, Pit Pichappan, "Improving Arabic Text Categorization using Decision Trees", First International Conference on, Networked Digital Technologies, NDT '09. 28-31 July 2009, pages 110 – 115, Print ISBN 978-1-4244-4614-8, Publisher:IEEE DOI:10.1109/NDT.2009.5272214

[19] Ajay S. Patil and B. V. Pawar, "Automated Classification of Web Sites using Naïve Bayesian Algorithm", proceeding of the International Multi Conference of Engineers and Computer Scientists 2012 Vol. IIMECS 2012, March 14-16, 2012 Hong Kong.

[20] Vladmimir Vapnik, (1995), "The Nature of Statistical Learning Theory", Springer,New York,1995.

[21] Tarek Fouad Gharib, Mena Badieh Habib, and Zaki Taha Fayed, "Arabic Text Classification Using Support Vector Machines", International Journal of Computers and Their Applications, VOLUME 16, NO. 4, December 2009.

[22]Fouzi Harrage, Abdul Malik Salman, Al-Salman, Mohammed BeMohammed, "A Comparative Study of Neural Networks Architectures on Arabic Text

Categorization using Feature Extraction", International Conference on Machine and Web Intelligence (ICMWI), 2010, 3-5 Oct. 2010, pg 102 – 107, ISBN:978-1-4244-8608-3, Publisher: IEEE, DOI10.1109/ICMW.2010.5648051

[23] Rupali P. Patil, R. P. Bhavsar, B. V. Pawar, "A Comparative Study of Text Classification Methods: An Experimental Approach", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), March 16 Volume 4 Issue 3, ISSN: 2321-8169, PP: 517 – 523.

[24] G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing", Communications of the ACM, Vol. 18, nr. 11, pages 613–620.

[25] Smruthi Mukund and Rohini K. Srihariin (2010), "A Vector Space Model for Subjectivity Classification in Urdu aided by Co-Training", COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, August 2010, pages 860–868.

[26] Hamouda K.Chantar and David W. Corne, "Feature Subset Selection for Arabic Document Categorization using BPSO-KNN", IEEE, "Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress", 19-21 Oct. 2011, 546 – 551, ISBN: 978-1-4577-1122-0, DOI:10.1109/NaBIC.2011.6089647

[27] Zhili Pei, Yuxin Zhou, Lisha Liu, Lihua Wang, YinanLu, "A Mutual Information and Information Entropy Pair based Feature Selection Method in text classification", 201O International Conference on Computer Application and System Modeling (ICCASM 2010) 978-1-4244-7237-61101, 2010 IEEE

[28] Ms. K. Umamaheswari, Dr. S. Sumathi, Ms. V. Aparna, Ms. A. Arthi, " Text Classification Using Enhanced Naïve Bayes With Genetic Algorithm", International Journal Of Computer Applications In Xiu-Li Pang, Engineering, Technology And Sciences (IJ-CA-ETS) ISSN: 0974-3596, April '09 – September '09, Volume 1, Issue 2, Page: 263.

[29] Vidhya. K. A, G. Aghila, "Hybrid Text Mining Model for Document Classification", 978-1-4244-5586-7/10 C 2010 IEEE

[30] Yaying Qiu, Guangming Yang, Zhenhua Tan, "Chinese Text Classification Based on Extended Naïve Bayes Model with Weighed Positive Features", 2010 First International Conference on Pervasive Computing, Signal Processing and Applications , DOI 10.1109/PCSPA.2010.268, 978-0-7695-4180-8/10 © 2010 IEEE

[31] Abhishek Sanwaliya, Kripa Shanker and Subhas C. Misra, "Categorization of News Articles: A Model based on Discriminative Term Extraction method", 2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications 978-0-7695-3981-2/10 © 2010 IEEE Computer Society DOI 10.1109/DBKDA.2010.18

[32] Shi Hong-Bo, Wang Zhi-Hai, Huang Hou-Kuan, and Jing Li-Ping, "Text Classification B,ased on the TAN Model", Proceedings of IEEE 7-ENCNOZ, 0-7803-7490-8/02 2002 IEEE.

[33] Suli Zhang & Xin Pan, "A Novel Text Classification Based on Mahalanobis distance", 978-1-61284-840-2/11 ©2011 IEEE.

[34] Yu-Qiang Feng, Wei Jiang, "An Improved Document Classification Approach with Maximum Entropy and Entropy Feature Selection", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007 1-4244-0973-X/07 ©2007 IEEE.

[35] Lam Hong Lee · Chin HengWan · Rajprasad Rajkumar · Dino Isa, "An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization", Published online: 25 August 2011© Springer Science+Business Media, LLC 2011 Appl Intell (2012) 37:80–99 DOI 10.1007/s10489-011-0314-

[36] Chin Heng Wan, Lam Hong Lee, Rajprasad Rajkumar, Dino Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine", Expert Systems with Applications 39 (2012) 11880–11888, doi:10.1016/j.eswa.2012.02.068, 0957-4174/$ - 2012 Elsevier Ltd.

[37] Srinivasan Ramaswamy, "Multiclass Text Classification A Decision Tree based SVM Approach", http://www.cs.berkeley.edu/~asimma/294-fall06/projects/reports/ramaswamy.pdf

[38] Verayuth Lertnattee & Thanaruk Theeramunkong, "Combining Homogeneous Classifiers for Centroid-based Text Classification", Proceedings of the Seventh International Symposium on Computers and Communications (ISCC'02) 1530-1346/02 © 2002 IEEE.

Rupali P. Patil received her MSc (Computer Science) degree in 1998 from Department of Computer Sciences, North Maharashtra University Jalgaon, Jalgaon (India), M.Phil (Computer Science) degree in 2009 from YCMOU Nashik, Nashik (India) and currently she is pursuing her PhD. from North Maharashtra University Jalgaon, Jalgaon (India). She is having 8 years teaching experience. Presently she is working as Assistant Professor of Computer Science in the Department of Computer Science at S. S. V. P. S's Lk. Dr. P. R. Ghogrey Science College, Dhule, Dhule (India). Her research interests include Data Mining, Natural language processing and Information retrieval.

Dr. R. P. Bhavsar has finished his M.C.A. (1995) and Ph.D. degree (2016) in Computer Sciences from North Maharashtra University, Jalgaon (MS), India. He has experience of 21 years in the field of Computer Science which includes Academic (Lecturer, Reader, Associate Professor), Full time Research (Member, Technical Staff at Applied Artificial Intelligence Group, C-DAC, Pune) and Professional Experience (System Analyst at Computer Centre, NMU). Presently he is working as Associate Professor in Computer Sciences at North Maharashtra University, Jalgaon. His areas of interest primarily include natural Language Processing (NLP) with special focus on practical development of application like Machine Translation, Lexical resource creation and word/verb Morphology Generation tools, application of Machine Learning to NLP etc. He has delivered invited talks in various Workshops/Seminars and National Conferences on topics related to NLP. He was co-PI for TDIL, Deity's ambitious consortia ANUVADKSH (formerly E-ILMT phase II) project of MoCIT, Govt. of India. NMU received funds of the tune of 77.50 lakhs. He is also coordinating research activities pertaining to NLP thrust area under UGC's SAP (DRS-I) scheme (51 lakhs) at School of Computer Sciences, NMU, Jalgaon. He is life member of Linguistic Society of India.

Prof. Dr B. V. Pawar received his B. E. (Production) degree in 1986 from VJTI, Mumbai University, Mumbai, (India) and M.Sc (Computer Science) degree in 1988 from Department of Computer Science, Mumbai University, Mumbai, (India). He received his

Ph.D. degree in Computer Science in 2000 at North Maharashtra University, Jalgaon (India). He is having 27 years teaching experience. Presently he is working as Professor and Director, School of Computer Science at North Maharashtra University, Jalgaon (India). He is member of various professional bodies like CSI & LSI. He has been recognized as a Ph.D. guide for the subjects Computer Science, Information Technology & Computer Engineering by various Universities in the state of Maharashtra (India). Till date he has guided 07 students towards their Ph.D. degree. His research areas include pattern recognition, neural networks, Natural Language Processing, Web Technologies & Information Retrieval. His work has been published in various international and national journals and conferences. He has successfully executed major R&D and development projects of the tune of 3.5 crores which includes ANUVADKSH, SAP (DRS-I) scheme, UGC Infonet, NME-ICT, NMU CAN Network project etc.