ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Survey of Text Classification Technique and Compare Classifier

[1] Mital Vala, [2] Prof. Jay Gandhi

[1] M.E Student, Dept of Computer Engineering , B.H.Gardi College of Engineering & Technology, Rajkot, India

[2]Assistant Professor, Dept of Information Technology, B.H.Gardi College of Engineering & Technology, Rajkot, India

ABSTRACT: Huge amount data on the internet are in unstructured texts can't simply be used for further processing by computer , therefore specific processing method and algorithm require to extract useful pattern. Text mining is process to extract information from the unstructured data. Text classification is task of automatically sorting set of document into categories from predefined set. A major difficulty of text classification is high dimensionality of feature space. Feature selection method used for dimension reduction. This paper describe about text classification process, compare various classifier and also discuss feature selection method for solving problem of high dimensional data and application of text classification.

KEYWORDS: Text mining, Text classification, Classifier, Feature selection.

I. Introduction

Text mining also known as text data mining. Text mining is similar to data mining, exact that data mining tool are design to handle structured data from data base but text mining can work with unstructured or semi-structured data sets such as word documents, email etc . Text mining is process of extracting interesting and non-trivial information from the unstructured text. Text mining task including text clustering, text classification, document summarization, entity modelling [1]. Here we describe text classification method in detail, Text classification is the act of dividing a set of input documents in to two or more class where each document can be said to belong to one or multiple class. The main aim of text classification is to identifying main themes of document by placing the document in to pre-defined set of topic.

Text Classification tasks can be broadly classified as Supervised Document Classification and Unsupervised Classification. In Supervised Document Classification some external mechanism provides information on the correct classification for documents or to define classes for the classifier, and in Unsupervised Document Classification, the classification must be done without any external reference and the system do not have predefined classes [4]. A major difficulty of text classification is high dimensionality of feature space. Feature selection method used to solve the problem of high dimensional data.

In this paper section 2 describe text classification process, section 3 compare various classifier, section 4 contain feature selection and section 5 describe application of text classification.

II. Text Classification Process

Text classification is the act of dividing a set of input documents into two or more class where each document can be said to belong to one or multiple class. Fig 1 show the different stages of text classification which include collection of documents, preprocessing , feature indexing, feature filtering, different classification algorithm and performance measure[3].

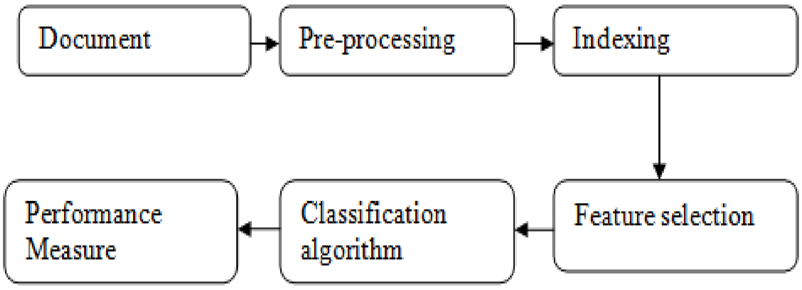

Fig1. Text classification process

(1) Documents:-First step is to collect the data from different type of format such as doc, html etc [3].

(2) Preprocessing:- In preprocessing phase converts the original textual data in a data-mining-ready structure, where the most significant text-features that serve to differentiate between text-categories are identified Commonly the steps taken are:
(a)Tokenization: A document is treated as a string, and then partitioned into a list of tokens.
(b)Removing stop words: Stop words such as “the”, “a”, “and”, etc are frequently occurring, so the insignificant words need to be removed.
(c) Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form, e.g. connection to
Connect , computing to compute[7].

(3) Indexing:- The document has to be changed from the full text to a document vector. The most commonly used Document representation is called vector space model here documents are represented by vectors of words. Usually, one has a collection of documents which is represented by word by word document Matrix.VSM representation scheme has its own disadvantages. Some of them are: high dimensionality of the representation, loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document to overcome these problems, term weighting methods are used to assign appropriate weights to the term.

(4) Feature selection methods:- Feature-selection methods used for reduction of the dimensionality of the dataset by removing features that are considered irrelevant for the classification .These feature selection methods possess a number of advantages such as smaller dataset size, smaller computational requirements for the text classification algorithms. To remove an irrelevant feature, a feature selection criterion is required which can measure the relevance of each feature with the output class [5].

(5) Classification Algorithms:-Databases are rich with hidden information that can be used for intelligent decision making. Classification algorithms can be used to extract models describing important data classes. Documents can be classified as supervised, unsupervised and semi supervised methods. There are several methods used to classify text such as Support Vector Machine, K Nearest Neighbor, Artificial Neural Network, Naive Bayes Classifier, and Decision Trees[7].

(6) Performance Evaluations:-This is Last stage of Text classification, in which the evaluations of text classifiers is typically conducted experimentally, rather than analytically. Precision and recall are the most common measures for evaluating an information retrieval system. Precision is the proportion of returned documents that are targets, while recall is the proportion of target documents returned[7].

III. COMPARISON OF CLASSIFIERS[3,7]

Classifier Name	Classifier principal	Advantages	Disadvantages
KNN	Distance is computed & k closest samples are selected the category of document on predicted based on nearest	Simple, effective and easy to implement.	Hard to find out the value of K, time cost is more.

	point.		
sNeural Network	Compute the input of unit j with respect to the previous layer , i.	Provide better result in complex domain.	Long training process.
SVM	SVMs find solutions of classification problems that have "generalization in mind", and also they are able to find non-linear solutions efficiently using the "kernel trick".	Compact description of the learned model, more capable to solve multi label classification.	Training speed is slow.
Naïve Bayes	Estimate the unknown probability distributions.	Easy for implementation & computation.	Very poor when features are co related to each other.
Decision Tree	Do the partition of data, which is set of training tuples & their associated class labels; then by making attribute list, select the attributes by attributes selection method, a procedure to determine the splitting criterion that gives the best partitions the data tuples in to individual class.	Simple and non expert user can understand.	Irrelevant attributes may affect badly the construction of a decision tree.

Table 1: comparison of classifier

IV. Feature Selection Method

Feature-selection methods used for reduction of the dimensionality of the dataset by removing features that are considered irrelevant for the classification. The aim of feature selection is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results.

One of the applications would be in gene microarray analysis .The standardized gene expression data can contain hundreds of variables of which many of them could be highly correlated with other variables .The dependant variables provide no extra information about the classes and thus serve as noise for the predictor. This means that the total information content can be obtained from fewer unique features which contain maximum discrimination information about the classes. Hence by eliminating the dependent variables, the amount of data can be reduced which can lead to improvement in the classification performance [5].

Feature selection techniques can be classified into two basic categories: filtering techniques and wrapper techniques. Filtering methods are independent of the learning algorithm .These methods regardless of learning algorithm and using statistical methods to feature selection and have low complexity. But wrapper methods uses from learning algorithm as the evaluation function. These methods have higher time complexity and accuracy than filter methods [10]

With the increasing size of the features in text classification, generally these methods could not be used because of the high complexity. Some filtering methods that can be used in many texts classification techniques such as Document Frequency (DF), Information Gain (IG), and Mutual Information (MI). Some wrapper methods are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS) and Neural Networks.

(1) ***Document frequency**:* Document frequency is number of documents in which a term occurs. DF thresholding is the simplest technique for the vocabulary reduction. We have to compute document frequency for each unique term in the training set and we have to discard all the terms whose frequency is less than the threshold k value from the feature space. The removal of the rare terms reduces the dimensionality of the feature space[5].

(2) ***Information gain:*** Information gain is frequently employed as a term goodness criterion in machine learning.

The prediction of category is done by knowing presence or absence of term in document and by measuring number of bits of information. In which give training corpus for each unique term we compute the information gain and remove from the feature space those terms whose information gain was less than some predetermined threshold. The computation includes the estimation of the conditional probabilities of a category given a term and the entropy computations in the definition [5].

(3) ***Mutual information****:* Mutual information is a criterion commonly used in statistical language modeling of, word associations and related applications .This is able to provide a precise statistical calculation that could be applied to a very large corpus to produce a table of association of words [5].

(4) ***Term strength:*** Term strength is originally proposed and evaluated for vocabulary reduction in text retrieval and later applied by yang and Wilbur to text categorization. This method estimates term importance based on How commonly a term is likely to appear in closely related documents. It uses a training set of documents to derive.

(5) ***X^2statistic***: The X^2statistic measures the lack of independence between t and c and can be compared to the X^2distribution with one degree of freedom to judge extremeness [5].

(6) ***Sequential Forward Selection***: Sequential Forward Selection starts with the empty set and sequentially adds one feature at a time. A problem with these Sequential Forward Selection techniques is that when a feature added in Sequential Forward Selection cannot be deleted once selected [9].

(7) ***Sequential Backward Elimination***: In Backward feature elimination, it starts with all the features and sequentially eliminates one feature at a time. A problem with this techniques is that when a feature is deleted, it cannot be re-selected. The issue which researchers faced when using statistical feature selection methods is it is not applicable when the variables are correlated. Thus, they discovered methods which are based on Shannon's Information Theory [9].

V. APPLICATION OF TEXT CLASSIFICATION

Text classification is an important task in document processing. The applications of classification are:

(1) **Automated survey coding: -** It is the task of assigning a symbolic code from a predefined set of such codes to the answer that a person has given in response to an open-ended question in a questionnaire). This task is usually carried out in order to group respondents according to a predefined scheme based on their answers. It has several applications, especially in the social sciences, where the classification of respondents is functional to the extraction of statistics on political opinions, health and lifestyle habits, customer satisfaction and patient satisfaction [2].

(2) **Automated authorship attribution:** - It is the science of determining the author of a text document, from a predefined set of candidate authors or inferring the characteristic of the author from the characteristics of documents written by that author [2].

(3) **Automatic Document Distribution: -** Text classification also allows the efficient automatic distribution of documents via email or fax by eliminating the time consuming, manual process of faxing or mailing [2].

VI. CONCLUSION

Text classification is task of automatically sorting set of document into categories from predefined set. This paper survey on text classification .The existing classification methods are compared and contrasted based on various parameters namely criteria used for classification, pros and cons. Finally we conclude that , Compare to other classifier SVM well, its accuracy, speed of learning, speed of classification, tolerance to irrelevant features and noisy data is much better than other classifiers. A major difficulty of text classification is high dimensionality of feature space. Feature selection method used for dimension reduction. Here also discuss various feature selection method and application of text classification.

REFERENCES

1. Vishal Gupta and Gurpreet S .Lehal,'A survey of text mining techniques and applications' ,Vol.1, Issue 1, Journal of emerging technology in web intelligence ,August 2009.

2. Nidhi and Vishal Gupta,'Recent trends in text classification techniques', Vol.35, Issue 6, International Journal of Computer Applications ,December 2011.
3. Anuradha and Patra, Divakar Singh,'A Surevy report on text classification with different term weighing methods and comparison between classification algorithem ', Vol.75, Issue 7, International Journal of Computer Applications , August 2013.
4. K.Nalini and Dr . L . Jaba Sheela ' Surevy on text classification ', Vol.1, Issue 6, International Journal of Computer Applications , July 2014.
5. S.Niharika , V.sneha Latha and D. R . Lavanya 'A survey on categorization', Vol.3, Issue 1, International Journal of Computer Applications 2012.
6. Meenakshi and Swati Singla'Review paper on text categorization technique' , SSRG International Journal of Computer science and engineering(SSRG-IJCSE)-EFES , April 2015.
7. Vandana Korde and C Namrata Mahender ,'Text classification and classifiers : survey', Vol.3, Issue 2, International Journal of Computer Applications 2012. International Journal of Artificial Intelligence & Applications, March 2012.
8. Upendra Singh and Saqib Hasan ,' Survey Paper on Document Classification and Classifiers', Vol.3 Issue 2, International Journal of Computer Science Trends and Technology, Mar-Apr 2015.
9. Inoshika Dilrukshi and Kasun de Zoysa. 'A feature selection method of machine learning and computing', Vol.4, Issue 4, International Journal of Machine Learning and Computing, August 2014.
10. Matthias Ring and Bjoern M.Eskfier,' Optimal feature selection for nonlineardata using branch-and-boundin kernel space' ,August 2015.
11. Alper kursat uysal,'An improved global feature selection schema for text classification' ,2015
12. Basant Agarwal and Namita Mittal, 'Text classification using machine learning method- A survey',2014, Springer India 2014
13. Kostas Fragos ,Petros Belsis and Christos skourlas,' Combining probabilistic classifiers for text classification'2014
14. Khalid Hussain zargar and Dr. Manzoor Ahmad Chachoo ,'Comparative study of text classification methods'Vol.2,Issue 8,International Journal for technological research on enginnering ,April 2015

BIOGRAPHY

Vala Mital is a PG student of Computer Engineering Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujrat, India.

Prof. Jay Gandhi is an Assistant Proffesor Computer Engineering Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujrat, India.