Development of Novel Machine Learning approach for Document Classification

A.N. Nihthiya Althaf[1], N.Priya[2], Mrs.Kalaivaazhi Vijayaraghavan[3]

[1,2]Student Members, Department of Information Technology
[3]Staff Member, Head of the Department of Information Technology
Anjalai Ammal Mahalingam Engineering College Kovilvenni, Tiruvarur.

Abstract— *Automatic classification of text document plays a vital area of research in the field of Text Mining (TM) ever since the explosion of online text information. The sources like digital libraries, emails, blogs, news groups, etc., make the rapid evolving growth of text documents in the digital era. In general the categorization of text document includes several field of interest like Information Retrieval (IR), Machine Learning (ML) and Natural Language Processing (NLP). Hence, this project focuses on the application of both supervised, unsupervised and semi-supervised ML techniques for classifying the text documents into pre-defined category labels.*

Keywords—*classification ; mining ; machine learning ; documents ;*

I. INTRODUCTION

Text mining is the part of data mining which is used to discover the previously unknown as well as the interesting information from a huge amount of textual data. It engages several fields like Information Retrieval (IR), Machine Learning (ML), Natural Language Processing (NLP) and Statistics. Document classification is one among the emerging research area in TM. It is a well proven approach to organize the huge volume of textual data. It also widely used in knowledge extraction and knowledge representation from the text data sets. The common classification applications are email categorization, spam filtering, directory maintenance, mail routing, news monitoring and narrow casting, etc. The solutions to the most of the applications are solved by using machine learning algorithms.

II. DOCUMENTS REPRESENTATION

The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector. Text representation is the important aspect in documents classification, denotes the mapping of a documents into a compact form of its contents. A text document is typically represented as a vector of term weights (word features) from a set of terms (dictionary), where each term occurs at least once in a certain minimum number of document. A major characteristic of the text classification problem is the extremely high dimensionality of text data. The number of potential features often exceeds the number of training documents. Text classification is an important component in many informational management tasks, however with the explosive growth of the web data, algorithms that can improve the classification efficiency while maintaining accuracy, are highly desired. Documents pre-processing or dimensionality reduction (DR) allows an efficient data manipulation and representation. Lot of discussions on the pre-processing and DR are there in the current literature and many models and techniques have been proposed. DR echniques can classified into Feature Extraction (FE) and Feature Selection (FS) approaches, as discussed below.

III. FEATURE EXTRACTION

The process of pre-processing is to make clear the border of each language structure and to eliminate as much as possible the language dependent factors, tokenization, stop words removal, and stemming . FE is the fist step of pre processing which is used to presents the text documents into clear word format. So removing stop words and stemming words is the pre-processing task. The documents in text classification are represented by a great amount of features and most of them could be irrelevant or noisy. Effective dimension reduction make the learning task more efficient and save more storage space. The steps involved in the feature extraction are given below:

Step 1 –Extract text (i.e. no preposition)
Step 2 –Remove stopwords
Step 3 –Convert all words to lowercase Step 4 –Stemming
Step 5 –Count the word frequencies Step 6 –Create an indexing file
Step 7 –Create the vector space model
Step 8 – Compute the inverse document frequency

Step 9 – Compute the weights of the words
Step 10 – Normalize all documents to unit length

IV. Feature Selection

After feature extraction the important step in preprocessing of text classification, is feature selection to construct vector space, which improve the scalability,efficiency and accuracy of a text classifier. In general, a good feature selection method should consider domain and algorithm characteristics . The main idea of FS is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word .The selected features retains original physical meaning and provide a better understanding for the data and learning process . For text classification a major problem is the high dimensionality of the feature space. Almost every text domain has much number of features, most of these features are not relevant and beneficial for text classification task, and even some noise features may sharply reduce the classification accuracy. Hence FS is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

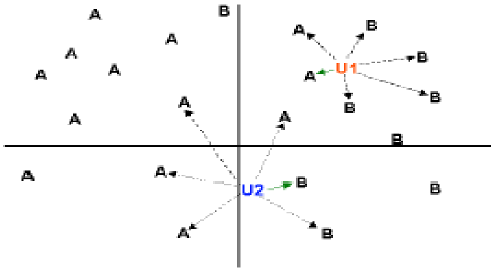
V. MACHINE LEARNING TECHNIQUES

The documents can be classified by three ways, unsupervised, supervised and semi supervised methods. Many techniques and algorithms are proposed recently for the clustering and classification of electronic documents.This section focused on the supervised classification techniques, new developments and highlighted some of the opportunities and challenges using the existingliterature. The automatic classification of documents into predefined categories has observed as an active attention, as the internet usage rate has quickly enlarged. From last few years , the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Latent Semantic Analysis, Rocchio's Algorithm, Fuzzy Correlation and Genetic Algorithms etc.Normally supervised learning techniques are used for automatic text classification, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents. Some of these techniques are described below.

A. K-nearest neighbor (k-NN)

The k-nearest neighbour algorithm (k-NN) is used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents. This method is an instant-based learning algorithm that categorized objects based on closest feature space in the training set . The training sets are mapped into multi-dimensional feature space. The feature space is partitioned into regions based on the category of the training set. A point in the feature space is assigned to a particular category if it is the most frequent category among the k nearest training data. Usually Euclidean Distance is typically used in computing the distance between the vectors.

The key element of this method is the availability of a similarity measure for identifying neighbours of a particular document . The training phase consists only of storing the feature vectors and categories of the training set. In the classification phase, distances from the new vector, representing an input document, to all stored vectors are computed and k closest samples are selected. The annotated category of a document is predicted based on the nearest point which has been assigned to a particular category.



$$\mathrm{Dist}(X, Y) = \surd\Sigma\,(\quad i- \quad i)2$$

B. Decision Tree

The decision tree rebuilds the manual categorization of training documents by constructing well-defined true/false-queries in the form of a tree structure. In a decision tree structure, leaves represent the corresponding category of documents and branches represent conjunctions of features that lead to those categories. The well organized decision tree can easily classify a document by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf, which represents the goal for the classification of the document.

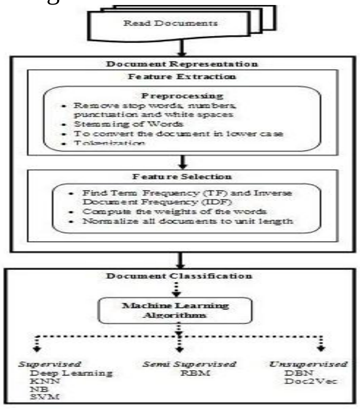
C. Naïve Bayes Algorithm

Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. These independence assumptions of features make the features order is irrelevant and consequently that the present of one feature does not affect other features in classification tasks [99]. These assumptions make the computation of Bayesian classification approach more efficient, but this assumption severely limits its applicability.

Depending on the precise nature of the probability model, the naïve Bayes classifiers can be trained very efficiently by requiring a relatively small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. Due to its apparently over-simplified assumptions, the naïve Bayes classifiers often work much better in many complex real-world situations than one might expect. Thenaïve Bayes classifiers has been reported to perform surprisingly well for many real world classification applications under some specific conditions.

D. Support Vector Machine (SVM)

Support vector machines (SVMs) are one of the discriminative classification methods which are commonly recognized to be more accurate. The SVM classification method is based on the Structural Risk Minimization principle from computational learning theory . The idea of this principle is to find a hypothesis to guarantee the lowest true error. Besides, the SVM are well-founded that very open to theoretical understanding and analysis.The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the ndimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. The performance of the SVM classification remains unchanged if documents that do not belong to the support vectors are removed from the set of training data


Read Documents
Document Representation
Feature Extraction
Preprocessing
Remove stop words, numbers, punctuation and white spaces
Stemming of Words
To convert the document in lower case
Tokenization
Feature Selection
Find Term Frequency (TF) and Inverse Document Frequency (IDF)
Compute the weights of the words
Normalize all documents to unit length
Document Classification
Machine Learning Algorithms
Supervised
Deep Learning
KNN
NB
SVM
Semi Supervised
RBM
Unsupervised
DBN
Doc2Vec

VI. HYBRID TECHNIQUES

Many new hybrid methods and techniques are proposed recently in the area of Machine Learning and text mining. The concept of combining classifiers is proposed as a new direction for the improvement of the performance of individual classifiers. Recently many methods have been suggested for the creation of ensemble of classifiers. A hybrid algorithm is proposed , based on variable precision rough set to combine the strength of both Support Vector Machine (SVM) and Single Value Decomposition (SVD) to improve the text classification accuracy and overcome the weaknesses of another algorithms.

VII. CONCLUSION

Combining classifiers has become a promising research area now a day. This paper gives an insight about the various methodologies that can be used for combining classifiers. Some of the relevant works for each method of combination are also discussed. And it is clear that the results obtained from the combination of classifiers are much better than the results obtained by the same classifiers individually. This further gives a boost to the research headed in providing faster and more efficient text classification process through classifier combinations.

REFERENCES

[1]. DASGUPTA, “FEAture selection methods for text classification.”, In Proceedings of the 13th ACMSIGKDD international conference on Knowledge discovery and data mining, pp. 230 - 239, 2007.

[2]. Raghavan, P., S. Amer-Yahia and L. Gravano eds., “Structure in Text: Extraction and Exploitation.” In. Proceeding of the 7th international Workshop on the Web and Databases(WebDB), ACM SIGMOD/PODS 2004, ACM Press, Vol 67, 2004.

[3]. Oracle corporation, WWW,oracle.com, 2008.

[4]. Merrill lynch, Nov.,2000. e-Business Analytics: Depth Report. 2000.

[5]. Pegah Falinouss “Stock Trend Prediction using News Article’s: a text mining approach” Master thesis -2007