IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. III (Jan. 2014), PP 17-21
www.iosrjournals.org

Survey of Machine Learning Techniques in Textual Document Classification

S.W. Mohod[1], Dr. C.A.Dhote[2]
[1](Deptt. Computer Engineering, B.D. College of Engg. Sevagram, Wardha, India)
[2](Prof.,Ram Meghe Institute of Technology & Research, Badnera.Amravati, India)

Abstract: *Classification of Text Document points towards associating one or more predefined categories based on the likelihood expressed by the training set of labeled documents. Many machine learning algorithms plays an important role in training the system with predefined categories. The importance of Machine learning approach has felt because of which the study has been taken up for text document classification based on the statistical event models available. The aim of this paper is to present the important techniques and methodologies that are employed for text documents classification, at the same time making awareness of some of the interesting challenges that remain to be solved, focused mainly on text representation and machine learning techniques.*

Keywords: *Text mining, Web mining, Documents classification, Information retrieval, Event models.*

I. Introduction

With the rapid growth of the World Wide Web and increasing availability of electronic documents, the task of automatic categorization of documents became important for organizing the information and knowledge discovery. Proper categorization of electronic documents, online news, blogs, e-mails and digital libraries requires text mining, machine learning and natural language processing techniques to extract required knowledge information. The term "Text document" refers to written, printed, or online document that presents or communicates narrative or tabulated data in the form of an article, letter, memorandum, report, etc. The Text expresses a vast range of information, but encodes the information in the form that is difficult to decipher automatically. In the existing online word huge amount of textual information is available in textual form in databases and various sources. The information may be available in structured and unstructured form. Unstructured means data that does not reside in fixed locations. The term generally refers to free-form text, which is present everywhere. Data that resides in fixed fields within a record or file that data is termed as a structured data. Relational databases and spreadsheets are examples of structured data.

In reality a large portion of the available information does not appear in structured databases but rather in collections of text articles drawn from various sources. Unstructured information refers to computerized information that either does not have a data model or the one that is not easily usable by a computer program. The term distinguishes such information from data stored in field form in databases or annotated in documents. However, data mining deals with structured data, whereas text presents special characteristics and is unstructured. The important task is how these documented data can be properly retrieved, presented and classified. Extraction, Integration and classification of electronic documents from different sources and knowledge information discovery from these documents are important.

In data mining, Machine learning is often used for Prediction or Classification. Classification involves finding rule that partition the data into disjoint groups. The input for the classification is the training data set, whose class labels are already known. Classifications analyze the training data set and construct a model based on the class label. The goal of classification is to build a set of models that can correctly predict the class of the different objects. Machine learning is an area of artificial intelligence concerned with the development of techniques which allow computers to "learn". More specifically, machine learning is a method for creating computer programs by the analysis of data sets since machine learning study the analysis of data.

Some machine learning systems attempt to eliminate the need for human intuition in the analysis of the data, while others adopt a collaborative approach between human and machine. Human intuition cannot be entirely eliminated since the designer of the system must specify how the data are to be represented and what mechanisms will be used to search for a characterization of the data. Machine learning has a wide spectrum of applications including search engines, medical diagnosis, detecting credit card fraud, stock market analysis, classifying DNA sequences, speech and handwriting recognition, game playing and robot locomotion.

II. Document Representation

One of the pre-processing techniques is the document representation which is used to reduce the complexity of the documents. The documents need to be transformed from the full text version to a document

vector. Text classification is again an important component in most information management tasks for which algorithms that can maintain high accuracy are desired. Dimensionality reduction is a very important step in text classification, because irrelevant and redundant features often degrade the performance of classification both in speed and classification accuracy. Dimensionality reduction technique can be classified into feature extraction (FE) [1] and feature selection (FS) approaches given below.

1 Feature Extraction

FE is the first 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 tasks [2] [3]. The documents in text classification are represented by a great amount of features and most of them could be irrelevant or noisy [4]. DR is the exclusion of a large number of keywords, base preferably on a statistical process, to create a low dimension vector [5]. Commonly the steps taken for the feature extractions (Fig.1) are:

Tokenization: A document is treated as a string, and then partitioned into a list of tokens.

Removing stop words: Stop words such as "the", "a", "and"… etc are frequently occurring, so the insignificant words need to be removed.

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 etc.

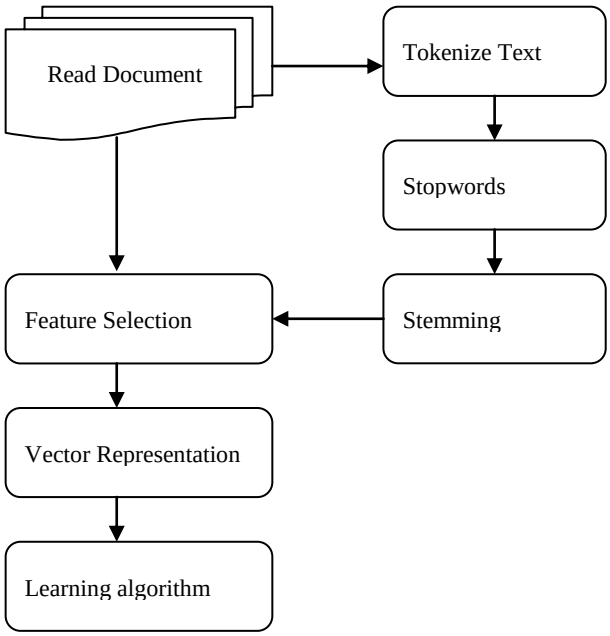

Figure. 1 Document Classification Process

2 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 [6]. 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 [4]. The selected feature retains the original physical meaning to provide a better understanding for the data and learning process [1]. 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 [7]. Hence FS is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

III. MACHINE LEARNING ALGORITHMS

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 existing literature. 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 labeled documents. Some of these techniques are described below.

1 Rocchio's Algorithm

Rocchio's Algorithm [8] is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents, i.e. the average vector over all training document vectors that belong to class ci, and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

2 K-nearest neighbor (k-NN)

The k-nearest neighbor algorithm (k-NN) [9] 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 [10]. 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 neighbors of a particular document [10]. 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.

3 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 better 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.

4 Decision Rules Classification

This method uses the rule-based inference to classify documents to their annotated categories [11]. The algorithms construct a rule set that describe the profile for each category. Rules are typically constructed in the format of "IF condition THEN conclusion", where the condition portion is filled by features of the category, and the conclusion portion is represented with the category's name or another rule to be tested. The rule set for a particular category is then constructed by combining every separate rule from the same category with logical operator, typically use "and" and "or". During the classification tasks, not necessarily every rule in the rule set needs to be satisfied. In the case of handling a data set with large number of features for each category, heuristics implementation is recommended to reduce the size of rules set without affecting the performance of the classification.

5 Naïve Bayes Algorithm

This 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 [12]. 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.

6 Artificial Neural Network

ANNs are constructed from a large number of elements with an input fan order of magnitudes larger than in computational elements of traditional architectures [13] [14]. These elements, namely artificial neuron are interconnected into group using a mathematical model for information processing based on a connectionist approach to computation. The neural networks make their neuron sensitive to store item. It can be used for distortion tolerant storing of a large number of cases represented by high dimensional vectors.

7 Fuzzy correlation

Fuzzy correlation can deal with fuzzy information or incomplete data, and also convert the property value into fuzzy sets for multiple document classification [15]. The researchers have shown great interest recently to use the fuzzy rules and sets to improve the classification accuracy, by incorporating the fuzzy correlation or fuzzy logic with the machine learning algorithm and the feature selection methods.

8 Genetic Algorithm

Genetic algorithm [16] aims to find optimum characteristic parameters using the mechanisms of genetic evolution and survival of the fittest in natural selection. Genetic algorithms make it possible to remove misleading judgments in the algorithms and improve the accuracy of document classification. This is an adaptive probability global optimization algorithm, which simulated in a natural environment of biological and genetic evolution, and is widely used for their simplicity and strength. Now several researchers used this method for the improvement of the text classification process. In [17] authors introduced the genetic algorithm for text categorization and used to build and optimize the user template, and also introduced simulated annealing to improve the shortcomings of genetic algorithm. In the experimental analysis, they show that the improved method is feasible and effective for text classification.

9 Support Vector Machine (SVM)

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 [18]. The idea of this principle is to find a hypothesis to guarantee the lowest true error. Besides, the SVM are well-founded that are very open to theoretical understanding and analysis [19]. 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 n-dimensional 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 [12].

IV. Conclusion

This paper provides a review of machine learning approaches and documents representation techniques. An analysis of feature selection methods and classification algorithms is also presented. It was observed from the study that information Gain and square statistics are the most commonly used and well performed methods for feature selection, however many other FS methods are recommended as single or hybrid technique. More work is required for the performance improvement and accuracy of the documents classification process. New methods and solutions are required for useful knowledge from the increasing volume of electronics documents.

REFERENCES

[1] Liu, H. and Motoda, ., "Feature Extraction, constraction and selection: A Data Mining Perpective.", Boston, Massachusetts(MA): Kluwer Academic Publishers.

[2] Wang, Y., and Wang X.J., " A New Approach to feature selection in Text Classification", Proceedings of 4th International Conference on Machine Learning and Cybernetics, IEEE- 2005, Vol.6, pp. 3814-3819, 2005.

[3] Lee, L.W., and Chen, S.M., "New Methods for Text Categorization Based on a New Feature Selection Method a and New Similarity Measure Between Documents", IEA/AEI,France 2006.

[4] Montanes,E., Ferandez, J., Diaz, I., Combarro, E.F and Ranilla, J., " Measures of Rule Quality for Feature Selection in Text Categorization", 5th international Symposium on Intelligent data analysis , Germeny-2003, Springer-Verlag 2003, Vol2810, pp.589-598, 2003.

[5] Manomaisupat, P., and Abmad k., " Feature Selection for text Categorization Using Self Orgnizing Map", 2nd International Conference on Neural Network and Brain, 2005,IEEE press Vol 3, pp.1875-1880, 2005.

[6] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li "An Optimal Svm-Based Text Classification Algorithm" Fifth International Conference on Machine Learning and Cybernetics, Dalian,pp. 13-16 , 2006.

[7] Jingnian Chen a,b,, Houkuan Huang a, Shengfeng Tian a, Youli Qua Feature selection for text classification with Naïve Bayes" Expert Systems with Applications 36, pp. 5432–5435, 2009.

[8] Rocchio, J; "Relevance Feedback in Information Retrieval", In G. Salton (ed.). The SMART System: pp.67-88.

[9] Tam, V., Santoso, A., & Setiono, R. , "A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization", Proceedings of the 16th International Conference on Pattern Recognition, pp.235–238, 2002.

[10] Eui-Hong (Sam) Han, George Karypis, Vipin Kumar; "Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification", Department of Computer Science and Engineering. Army HPC Research Centre, University of Minnesota, Minneapolis, USA. 1999.

[11] Chidanand Apte, Fred Damerau, Sholom M. Weiss.; "Towards Language independent Automated Learning of Text Categorization Models", In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 23-30,1994.

[12] Heide Brücher, Gerhard Knolmayer, Marc-André Mittermayer; "Document Classification Methods for Organizing Explicit Knowledge", Research Group Information Engineering, Institute of Information Systems, University of Bern, Engehaldenstrasse 8, CH - 3012 Bern, Switzerland, 2002.

[13] Miguel E. Ruiz, Padmini Srinivasan; "Automatic Text Categorization Using Neural Network",In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, pp. 59-72. 1998.

[14] Petri Myllymaki, Henry Tirri; "Bayesian Case-Based Reasoning with Neural Network", In Proceeding of the IEEE International Conference on Neural Network'93, Vol. 1, pp. 422-427. 1993.
[15] Que, H. -E. "Applications of fuzzy correlation on multiple document classification. Unpublished master thesis", Information Engineering Department, Tamkang University, Taipei, Taiwan-2000.
[16] Wang Xiaoping, Li-Ming Cao. Genetic Algorithm Theory, Application and Software[M]. XI'AN:Xi'an Jiaotong University Press, 2002.
[17] ZHU Zhen-fang, LIU Pei-yu, Lu Ran, "Research of text classification technology based on genetic annealing algorithm" IEEE,, 978-0-7695-3311-7/08, 2008.
[18] Vladimir N. Vapnik, "The Nature of Statistical Learning heory", Springer, NewYork. 1995.
[19] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features" ECML-98, 10th European Conference on Machine Learning, pp. 137-142. 1998.