J. Basic. Appl. Sci. Res., 3(1s)31-35, 2013

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Text Classification with Machine Learning Algorithms

Nasim VasfiSisi[1] and Mohammad Reza Feizi Derakhshi[2]

[1]Department of Computer, Shabestar Branch, Islamic Azad University, Shabestar, Iran
[2]Department of Computer, University of Tabriz, Tabriz, Iran

Received: June 10 2013
Accepted: July 7 2013

ABSTRACT

By increasing the access to electronic documents and rapid growth of World Wide Web, documents classification task automatically has become a key method to organizing information and knowledge discovery. The appropriate classification of electronic documents, online news, weblogs, emails and digital libraries required for text mining, machine learning techniques and natural language processing is to obtain meaningful knowledge. The aim of this paper is to highlight the major techniques and methods applied in classification of documents. In this paper, we review some existing methods of text classification.
KEYWORDS: Text mining, text classification, machine learning algorithms, classifiers.

1. INTRODUCTION

In recent years, a dramatic growth has taken place in volume of text documents over internet, news sources and intranet throughout companies where the classification of these documents is required. The text automatic classification task is to use text documents for predefined classes of documents which could help in both well organization and finding information over these great resources. This work has several applications including automatic indexing of scientific articles based on predefined store of terminologies, archive inventions submission in inventions list book, spam filtering, identify different types of documents, automatic grading of articles and authorship documents and electronic government's repositories, articles news, biological databases, chat rooms, online associations, electronic mails and weblog pools [2].

Automatic classification of documents helps organizations to get rid of manual classification and also manual classification could be expensive and time consuming. The precision of modern text classification systems has become a competitor for professional trained people and as a result it is a combination of information retrieval technologies and machine learning technologies [2,3].

Today, text classification gives an individual challenge due to excess of existing features in datasets and excess of training samples and dependent features which lead to development of different types of text classification algorithms [10].

In text classification each document is placed either in none of the classes, in multiple classes or in one class. The mail goal of using machine learning methods is that the classifier learns the learning from the samples which previously have been classified in the previous classed automatically [1]. For example, we can label each of the automatically received news by a subject like "sport", "politic" or "art". Classification of a dataset like $d=(d_1,\ldots d_n)$ starts from labeled classes, c1, c2,…, cn (such as sport, political and etc) and then the same process is performed to determine a classification model which is able to determine the suitable class for a new document d from the text classification domain which has one label or multiple labels. Documents with one label belong to only one class and multiple labels belong to more than one class [4].

In this paper, we will have documents pre-processing steps in section two, different types of text classification methods are presented in section three and finally in section four we will have conclusion.

2. Pre-processing document

The first step in text classification is to transform documents into a string of characters with various formats which is represented for learning and classification algorithms. Always, it is better to find the word's root in information retrieval so that the word could be applied as a unit in documents and this unit word lead to representation of feature value in the text. Each separate word has one feature, where the value of this feature equals to the number of word occurrence in documents. To eliminate unnecessary feature vectors some words are considered as features which have occurred at least three times in training data and are not included in stop words [1]. Fig. 1 represents the text classification process:

Corresponding author: Nasim VasfiSisi, Department of Computer, Shabestar Branch, Islamic Azad University, Shabestar, Iran. Email address: Nasim_vasfi@yahoo.com

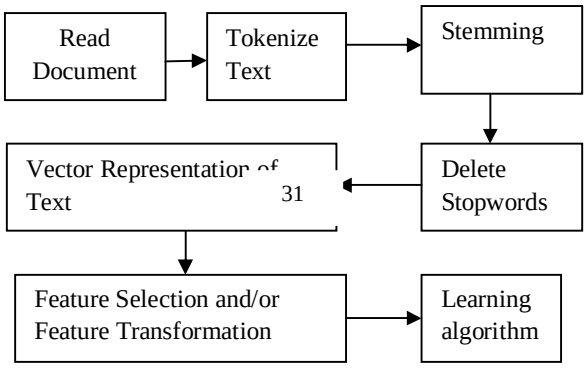

Fig 1. represents the text classification process [1]:

We briefly describe the fig. 1:

a) Read Document step: at first all of documents are read.
b) Tokenize text step: in this step the text is broken into tokens, meaningful words, terms, phrases, symbols or elements which is called Tokenization.
c) Stemming: step: the root of words is transformed into an original form.
d) Stop words step: words such as in, this, a, an, the, with and etc are removed.
e) Vector Representation of Text: In this step, a algebraic model is defined to represent text documents as a vector. Because the main goal of feature selection methods is to reduce dimensionality.
f) Feature Selection and/or Feature Transformation: In this step we reduce the dimensions of datasets using feature selection methods by removing the features not related to classification. After documents feature selection, according to the flexibility, we can use machine learning algorithms such as Genetic algorithm, Neural Network, Rule Induction, Fuzzy Decision Tree, SVM, K-NN (K-Nearest Neighbor) algorithm, Lsa, Rocchio algorithm and Naïve Bayesian [1].

Machine learning, natural language processing (NLP) and data mining techniques work for automatic classification and discovery of electronic documents' patterns. The main goal of text mining is to allow users to extract information from text resources and deal with actions like retrieval, classification (supervised, unsupervised and pseudo supervised) and summarization [3].

Development of computer hardware provides the adequate strength of computations in order to allow text classification to be used in applications. Text classification is usually used to deal with spam emails, classify large text collections in to subjective classes, knowledge management and also help to internet search engines [6].

3. Classifiers

3.1 SVM algorithm

The standard SVM (Support Vector Machines) has been purposed by Cortes and Vapnik in 1995 [8].

SVM is one of the supervised learning methods used for classification and regression. SVM classification method is from arithmetic learning theory based on Structural Risk Minimization principle. The idea of this principle is to find a hypothesis to guarantee the least error. SVM requires two positive and negative training sets which is unusual for other classification methods. This positive and negative training set is necessary for SVM to search a decision level in order to separate positive and negative data within n-dimensional space in a best way which is called hyper plane. Therefore, SVM creates a hyper plane or a set of large surfaces in a space with high dimensions [2,3].

In general, a useful separator for distance is obtained by a hyper plane which has the highest distance from neighbor training data points of both classed (which is called margin) and the highest margin produces the least error of classification [8]: In SVM method it is attempted to reduce the number of points classified wrongly and the logical way to goal consistence is as equation (1) [2]:

$$\begin{aligned} \min \quad & \frac{1}{2}\left\|w\right\|^2 + C\sum_{i=1}^{l}\xi_i \\ s.t. \quad & y_i\left[\left(w \cdot x_i\right) + b\right] \geq 1 - \xi_i \end{aligned} \tag{1}$$

3.2 Neural network algorithm

Neural network classification is a network of units, where input units usually represent words and output unit(s) represents a class or the label of class. for classifying a test document, the weight of words is determined for input units and activation of these units is performed through forward propagation in the network and the value of output unit is determined as a result in decision of classes. Some researches use single-layer perceptron, since its implementation is simple and multi-layer perceptron that is so complex requires an extensive implementation for classification. Using an effective feature selection method to reduce dimensionality improves efficiency in this method. The documents classification methods based on newly presented neural networks is so useful in companies to evident management of documents [4].

3.3 k-NN (K-Nearest Neighbor) algorithm

K-NN is a case-based learning method and is one of the simplest machine learning algorithms. In this algorithm, an example with majority vote from neighbor is classified and this example is determined in the most general class among k nearest neighbors. K is a positive integer and typically small. If k=1, then the example is simply assigned to the class of its nearest neighbor. The oddness of k is useful, since by this method, the equal vote is prevented [5]. K-NN has an application for most methods, since it is effective, non-parametric and has simple implementation, whereas its classification time is longer and it is difficult to find the optimal k value. The best selection from k depends on data, in general the high value of k reduces the noise effect on classification, but the margin among classes is differentiated less [4]. Fig. 2 is an example of K-NN classification algorithm [7]:

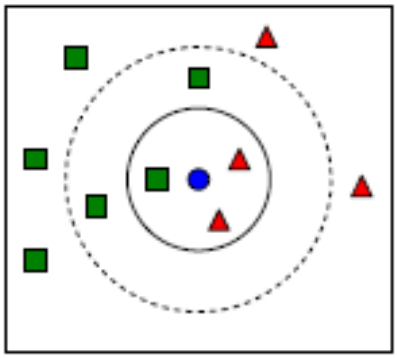

Fig 2. Example of K-NN classification algorithm [7].

Fig. 2 is an example of K-NN classification algorithm by using multi-dimensional feature vector where triangles represent the first class and squares show the second class. The small circle shows the test example. Now, if k=3 then the test example belongs to triangle class and if k=5, the example belongs to square class [5].

The training steps of this algorithm are as follows: this algorithm classifies a test document based on k nearest neighbor. The training examples are introduced as vectors in multi-dimensional feature space. The space is portioned into areas with training examples. A point in the space is assigned to a class in which the most training points belonging to that class within the K nearest training example, usually, Euclidean distance or Cosine similarity are used in this method. In classification phase, a test example is represented as a vector in feature space and Euclidean distance or Cosine similarity of test vector with whole training vectors is measured and the K nearest training example is selected. Of course, there are many ways to classify test vector and therefore, the classic K-NN algorithm determines a test example based on the maximum votes from the k nearest neighbors. Three main factors in K-NN algorithm are as follows [7]:

1. Distance or similarity criterion to find the K-Nearest Neighbor.
2. K is the number of nearest neighbors.
3. The decision rule is to determine a class for test document from k nearest neighbors.

3.4 Decision Tree

Decision Tree is a classification algorithm whose structure is based on "if-then" classification rules. In this method, at first we must determine the possible events and draw the tree from the root node. Each node describes a value taken from gain function [9].

In a decision tree, leaves show similar class of documents and branches represent the conjunction of features related to that class. A well-structured decision tree can place the class of a document simply in the root node of tree and allow performing the query structure until reaching a certain leaf which represents the aim of document. Fig. 3 represents a decision tree classification algorithm [3].

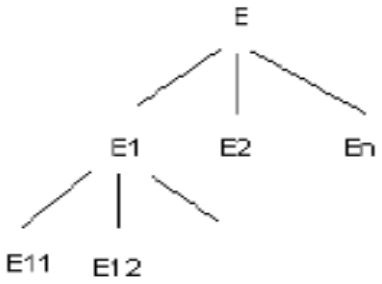

Fig 3. An example a Decision Tree [3]:

The decision tree classification method has dominant advantages over other decision support means. The main advantage of decision tree is its understanding and interpretation even for non-expert users. Furthermore, the interpretation of obtained results could be done conveniently by using a simple mathematical algorithms. Decision tree could experimentally show that the iteration of text classification includes so many appropriate and related features. An application of decision tree is to personalize advertisement in web pages. A major risk in implementation of a decision tree is to over fit of training data with the occurrence of an alternative tree that categorizes the training data worse but would categorizes the document to be categorized better [3].

3.5 Bayesian classification

The Bayesian classification is a simple possibility classification based on an application of Bayesian theorem with strong independent hypothesis. Description of probabilistic model is independent from description of features model. The features independency hypothesis makes the order to features unimportant and as a result, now one feature does not influence on other features in classification. These hypotheses have resulted in effectiveness of Bayesian classification method's computation, but this hypothesis limits its application significantly. According to the precise nature of probabilistic model, the Bayesian classifier could be trained more effectively with relatively low requirement of training data in order to estimate the necessary parameters for classification, since we have assumed parameters independent, it is only necessary to determine the variance of variants for each class, but not covariance of whole matrix [3].

4. Conclusion

Various algorithms or a combination of hybrid algorithms have been purposed for automatic classification of documents. The Bayesian classification is used well in filtering spam and emails and text classification and requires a few numbers of training data to estimate essential parameters for classification. Bayesian classification performs well over text and numerical data and has convenient implementation in comparison with other algorithms.

Although the hypothesis of conditional independency is contradicted by real world's data and when the feature are so dependent to each other it performs so weak and it does not have centralization in the words occurrence abundance. The advantage of Bayesian classification is that it requires a few training data to estimate the necessary parameters for classification and its main disadvantage is the relatively low efficiency of classification in comparison with other detection algorithms.

SVM classification has been known as one the most effective text classification methods in comparison with supervised machine learning algorithms and provides a perfect precision, but in this case recollection is reduced. SVM takes the main features from data and replaces it with Structural Risk Minimization (SRM) principle to minimize the upper bound in error generalization and also, capability to learn could be independent from feature vector dimensions. K-NN algorithm performs well when so local features of documents are introduced, while the classification time is longer in this method and it is difficult to find the optimal value to k. The major advantage of decision tree is its simplicity of understanding.

REFERENCES

1. Bhavani Dasari, D. and Gopala Rao. K, V., Text Categorization and Machine Learning Methods, Current State of the Art, Global Journal of Computer Science and Technology Software & Data Engineering, 2012. 12(11).
2. LIU, X. and FU, H., A Hybrid Algorithm for Text Classification Problem, 2011. Przegląd Eektrotechniczny (Electrical Review).
3. Khan, A. , Baharudin, B., Hong Lee, L. and Khan, Kh., A Review of Machine Learning Algorithms for Text-Documents Classification, Journal of Advances in Information Technology, 2010. 1(1).
4. Korde, V. and Mahender, C N. , Text Classification and classifiers, A survey, International Journal of Artificial Intelligence & Applications (IJAIA), 2012. 3(2).
5. Ananthi, S. and, Sathyabama, S. , Spam Filtering Using K-NN, Journal of Computer Applications, 2009. 2(3).
6. Mahinovs, A. and Tiwari, A., Text Classification Method Review. Decision Engineering Report Series, 2007.
7. Miah, M. , Improved k-NN Algorithm for Text Classification, In Proceedings of DMIN:2009. P. 434-440.
8. Xiao.li, CH., Pei.yu, L. , Zhen.fang, Z. and Ye, Q., A Method of Spam Filtering Based on Weighted Support Vector Machines, IEEE International Symposium on IT in Medicine & Education, 2009. 1.
9. Naksomboon, S. , Charnsripinyo, C. and Wattanapongsakorn, N., Considering Behavior of Sender in Spam Mail Detection. International Conference on Networked Computing (INC 2010), 2010. Gyeongju, South Korea.
10. Han, E. H. S. and Karypis, G., Centroid-based document classification, Analysis and experimental results, 2000, Springer Berlin Heidelberg. p. 424-431.