ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 5, Issue, 11, pp. 5952-5955, November, 2015

International Journal of
DEVELOPMENT RESEARCH

Full Length Review Article

A SURVEY ON TEXT ANALYTICS AND CLASSIFICATION TECHNIQUES FOR TEXT DOCUMENTS

Nihar Ranjan, Abhishek Gupta, Ishwari Dhumale, Payal Gogawale and *Rugved Gramopadhye

Department of Computer Engineering, Sinhgad Institute of Technology and Science, SavitriBai Phule Pune University, Pune, India

ARTICLE INFO

Article History:
Received 17th August, 2015
Received in revised form
15th September, 2015
Accepted 24th October, 2015
Published online 30th November, 2015

Key Words:
SVM,
Text Mining,
Text, Categorization,
NLP.

ABSTRACT

Text Mining is termed as extraction of relevant yet hidden information from the text document. One of the essential concepts in the field of text mining is Text classification (Also called Text Categorization). Through the sudden growth in digital world and available documents, the task of organizing text data becomes one of the principal problems. The classification problem has been widely studied in data mining, machine learning, database, and information retrieval. On the basis of text information processing, we have made a study of support vector machine in text categorization. By introducing the basic principle of SVMs, we described the process of text classification. Comparative Study of other classification algorithm is done and this paper states that how SVM is an effective machine learning algorithm for classification. A theoretical study of SVM and other machine learning techniques can be found in this paper along with their advantages and disadvantages.

INTRODUCTION

Text Mining (Kroeze, 2007) (Sebastiani, 2002) refers to the process of deriving high-quality information from text. 'High quality' in text mining means that information extracted should be relevant to the user, and according to the interest of the user. Text mining is similar to data mining, except that data mining tools (Navathe *et al*., 2000) are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc.

Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. The corporate data is becoming double in size. In order to utilize that data for business needs, an automated approach is Text mining. By mining that text required knowledge can be retrieved which will be very useful. Knowledge from text usually refers to some combination of relevance, novelty, and interestingness.

Corresponding author: Rugved Gramopadhye,
Department of Computer Engineering, Sinhgad Institute of Technology and Science, SavitriBai Phule Pune University, Pune, India.

Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities). Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/ annotation, information extraction, data mining techniques including link and association analysis, visualization and predictive analytics. A typical application of text mining is to scan given set of documents written in a natural language and either to model them for predictive classification or populate a database or search index with the information extracted.

Text (or Document) classification is an active research area of text mining, where the documents are classified into predefined classes. Text Classification tasks can be broadly classified as Supervised Document Classification and Unsupervised Classification. In Supervised Document Classification some external mechanism (such as human feedback) provides information on the correct classification for documents or to define classes for the classifier, and in Unsupervised Document Classification (also known as document clustering), the classification must be done without any external reference and the system do not have predefined classes.

There is also another task called Semi-Supervised Document Classification, where some documents are labeled by the external mechanism (means some documents are already classified for better learning of the classifier). There is a need to construct automatic text classifier using pre-classified sample documents whose accuracy and time efficiency is much better than manual text classification because to classify millions of text document manually is an expensive and time consuming task. In this paper we will be stating various algorithms and techniques used for text mining and categorizing them. Most of the techniques are explained by simplifying them as they can be understood by the reader. This paper provides reasons for choosing SVM in our project along with other algorithms like NLP as well. Improving classifier effectiveness has been an area of intensive machine-learning research over the last two decades, and this work has led to a new generation of state-of-the-art classifiers, such as support vector machines.

An SVM is a kind of large-margin classifier: it is a vector space based machine learning method where the goal is to find a decision boundary between two classes that is maximally far from any point in the training data. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples.

The following figure is been referred from (Upendra Singh and Saqib Hasan, 2015).

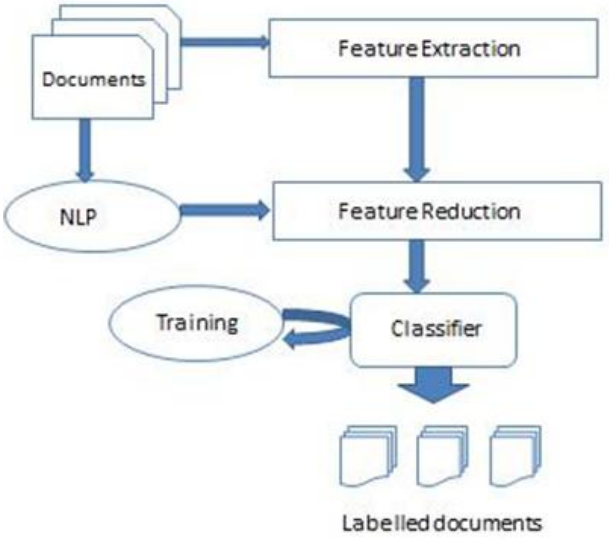

Fig. 1. Steps of Text Classification

CLASSIFICATION

With each passing day, automatic classification of documents in predefined categories is gaining active attention of many researchers. Supervised, unsupervised and semi supervised are the methods used to classify documents. The last decade has seen the unprecedented and rapid progress in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor (KNN), Support Vector Machines(SVMs), Neural Networks.

DIFFERENT CLASSIFIERS

Decision trees

A Decision Tree is a hierarchical (flow-chart like) tree structure used for classification of text documents. A Decision Tree text classifier in (Russell Greiner and Jonathan Schaffer, 2001) is a tree with internal nodes labeled by terms branches departing from them are labeled by the weight that the term has in the text document and leafs are labeled by categories. It can be also specified as each internal node represents a test on document, each branch represents an outcome of the test, and each leaf node holds a class label. It is a top-down method. Decision tree uses 'divide and conquer' approach for classification. The internal nodes which denote tests are represented using rectangles and the leaf nodes by circles.

Advantages

- Decision trees are simple to understand and easy to interpret.
- Modifications and addition of new possible scenario can be easily done.
- Their robustness to noisy data and their capability to learn disjunctive expressions seem suitable for document classification. Decision trees are simple to understand and interpret (Nalini and Jaba Sheela, 2014).

Disadvantages

- As levels of a tree increases the complexity of calculations also increases.
- Decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where decisions are made at each node locally and cannot guarantee to return the globally optimal decision tree.

Naïve Bayes classifier

Bayesian classifiers are statistical probabilistic classifiers used for text categorization. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Naïve Bayes classifier is based on Bayes' theorem. The Bayes' theorem provides posterior probability of the documents belonging to different classes. Naïve Bayes algorithm uses this posterior probability for classification of the documents. The document is assigned to the class if it has maximum posterior probability for that class.

This algorithm computes the posterior probability of the document belongs to different

Advantages:

- It is fast to train and classify the data or documents.
- It is not affected by irrelevant features.
- Streaming data is handled well.

Disadvantages

- It is independent feature model so that the present of one feature does not affect other features in classification tasks (5).

K-Nearest Neighbor

KNN is a classification algorithm is used for text classification. As given in (Tam *et al*., 2002) KNN classifies dataset or objects by voting several labeled training data with their smallest distance from each dataset or object. It uses the local neighborhood to predict the class of an object. The majority vote of its neighbors decides the class of an object. The object is assigned to the class most common among its k nearest neighbors.

Here k is a positive integer. If k= 1, the object is assigned to class of that single nearest neighbor. The classes of these neighbors are decided using the similarity of each neighbor to new document vector, where similarity may be measured by for example the Euclidean distance or the cosine between the two document vectors.

Advantages

- The cost of the learning process is zero
- No necessity of assumptions about the characteristics of the concepts to learn have to be done
- It is very simple.

Disadvantages

- The model cannot be interpreted.
- It is computationally expensive, requires more time to find the k nearest neighbors when there is large number of training datasets.
- It has to compute distance of each test objects with whole training dataset.

Neural Networks

Neural network is related to electronic networks of 'neurons'. It is based on neural structure of the brain. It is an iterative learning process. The neural nework is given input as some input values with their associate weights. And the output of network is classified input. The errors made in previous classification can be corrected in neural network. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm the second time around, and so on for many iterations.

Advantages:

- Neural networks are self-adaptive and data driven.
- It provides high accuracy and noise tolerance.
- Neural networks are nonlinear models, which makes them flexible in modeling real world complex relationships.

Disadvantages

- Lack of transparency.
- Learning time is very long.

SVM

Support vector machines are supervised learning models with associated learning algorithms used for text classification. It is a non-probabilistic binary linear classifier (8). SVM constructs a hyper plane which is used for regression, categorization and similar tasks. A good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the argin the lower the generalization error of the classifier.

SVM can categorize multi-dimensional data or documents easily. It partitions the data-space using linear and non-linear definitions into two parts. SVM classifier uses hyper-plane for partitioning the data. SVM can handle larger feature spaces efficiently, as it uses over fitting protection, which does not necessarily depend on the number of features.

Advantages

- The accuracy of SVM is high.
- SVM is robust.
- The learned target function is evaluated fast.
- SVM can handle multi-dimensional data.

Disadvantages

- Learned functions are difficult to understand.

EXPERIMENTAL RESULTS AND ANALYSIS

The following experimental data, results and analysis has been referred from (Zhijie *et al*., 2010).

Experimental Data

Experimental data include four class documentations, and they are environment, sport, politic, and art. The training data are different from the testing data.

Experimental Results

Performance evaluation of text classification mainly includes accuracy rate, recall rate and F1 value. The following three sets of experimental data are the different classification results with the same training data sets and testing data sets under different classification methods.

Table 1.The result of KNN classification is shown

Category name	Training corpus/ Testing corpus	Accuracy rate	Recall rate	F1 value
environment	1800/200	97.44%	76.00%	85.39%
sport	1800/200	73.49%	91.50%	81.51%
politic	1800/200	78.48%	93.00%	85.13%
Art	1200/200	94.30%	74.50%	83.24%

Table 2. The result of Naive Bayesian classification is shown

Category name	Training corpus/ Testing Corpus	Accuracy rate	Recall rate	F1 value
environment	1800/200	95.92%	70.50%	81.27%
Sport	1800/200	76.86%	93.00%	84.16%
Politic	1800/200	91.76%	78.00%	84.32%
Art	1200/200	79.67%	96.00%	87.07%

Table 3. The result of SVM classification is shown

Category name	Training corpus/ Testing Corpus	Accuracy rate	Recall rate	F1 value
Environment	1800/200	86.03%	86.50%	86.26%
sport	1800/200	86.07%	86.50%	86.28%
politic	1800/200	97.31%	90.50%	93.78%
Art	1200/200	93.48%	86.00%	89.58%

Accuracy rate and recall rate reflect two different aspects of classification quality, while a comprehensive evaluation index of the two aspects is the F1 value. As shown in Fig.4, the figure reflects classification results of the various classifiers under the composite index F1 value.

Experimental Analysis

By comparing and analyzing Table 1, Table 2 and Table 3, we can draw the following conclusions: From the view of accuracy rate, SVM classification method all achieved 86.03%; although the accuracy rate of KNN classification method in environment and art is higher than SVM, classification results are lower than 80% in sport and politic; similarly, the accuracy rate of Naive Bayesian classification method is higher than SVM only in environment, the other three types are not better than SVM.

From the view of recall rate, SVM classification method has also reached 86%; for the KNN classification method and Naive Bayesian classification method, the recall rate has a fluctuation up and down, and the difference is obvious. That is to say, the overall effect is not better than SVM. From the view of F1 value, Fig.4 has made an intuitive comparison of classification results for different classifiers. We can clearly see that, SVM classification method in the four types of texts is higher than the other two classification methods. As a comprehensive evaluation index for text classification, F1 test value is better to reflect the effects of a good or bad classifier, so as a whole, SVM classification method is superior to other classification methods.

CONCLUSION

This paper has stated that classification of documents is one of the most fundamental problems in the machine learning and data mining .With the drastic increase in the world digitization, there has been an explosion in the volume of documents. Text Classification is hence needed to classify the documents according to the predefined classes based on their content. A comparative study has been done among different techniques which are used for classification such as nearest neighbor classifiers, SVM classifiers, neural networks, decision trees, Bayes methods. When compared it was found that K-nearest neighbor algorithm (KNN) is the simplest method for deciding the class of the unlabeled documents and is a popular non-parametric method. But for the high dimensions, this method is not suitable for such documents. SVMs and Neural Network tend to perform much better when dealing with multi dimensions. For SVMs and Neural Network, large sample size is required to achieve maximum accuracy of the classifier, whereas Naïve Bayes may need a relatively less dataset and require little storage space. KNN, Neural Network is generally considered intolerant of noise; where association based classification and decision trees are considered resistant to noise because their pruning strategies avoid over-fitting the noisy data. Compared to other classifiers, SVM performs better as it has high accuracy, high speed of learning, high speed of classification, high tolerance to irrelevant features and noisy data than other classifiers. But still it seems difficult to recommend any one technique as superior to others as the choice of a modeling technique depends on organizational requirements and the data on hand.

There are still various open questions regarding implementation of SVMs. Will the algorithm support semi-structured documents? How does the algorithm works for unsupervised learning? What kind of relationship can be established between SVM and various machine learning algorithms for maximum effectiveness? More research is needed on these relationships along with the questions from learning theory. Also Text classification is a widespread domain of research encompassing Data mining, NLP and Machine Learning. One can use SVM with NLP for categorization of documents. The cost effectiveness, wide scope of further development, implementation in various sectors and profound study on algorithm are the key aspects to increase the use of support vector machine algorithms.

REFERENCES

https://en.wikipedia.org/wiki/Support_ve ctor_machine

Irina Rish, "An Empirical Study of the Naïve Bayes Classifier", Proc.of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence, Oct 2001. citeulike-article-id:352583.

Kroeze, J.H., Matthee, M.C. and Bothma, T.J.D. July 2007, "Differentiating between data-mining and text-mining Terminology.

Nalini, K. and Dr. Jaba Sheela, L. "Survey on Text Classification", July 2014. *International Journal of Innovative Research in Advanced Engineering*, (IJIRAE) ISSN: 2349-2163Volume 1 Issue 6 URl: http://ijirae.com

Navathe, Shamkant, B. and ElmasriRamez, 2000. "Data Warehousing and Data Mining", in "Fundamentals of Database System s", Pearson Education pvt Inc, Singapore, 841-872.

Russell Greiner and Jonathan Schaffer, "Exploratorium – Decision Trees", Canada. 2001. URL: http://www.cs.ualberta.ca/~aixplore/ learning/ Decision Trees

Sebastiani, F. 2002. "Machine learning in automated text categorization", ACM Computer Surveys 34(1), 1–47.

Tam, Santoso, A. and Setiono R. 2002. "A comparative study ofcentroid-based, neighborhood-based and statistical approaches for effective document categorization", ICPR '02 Proceedings of the 16th *International Conference on Pattern Recognition* (ICPR'02), vol.4, no. 4, pp.235–238

Upendra Singh, Saqib Hasan, 2015. "Survey paper of Document Classification and Classifiers." *International Journal of Computer Science Trends and Technology* (IJCST) – Volume 3 Issue 2, Mar-Apr 2015.

Zhijie Liu, Xueqiang Lv, Kun Liu, Shuicai Shi, 2010. "Study on SVM Compared with the other Text Classification Methods", *Second International Workshop on Education Technology and Computer Science*

****