

A Comparative Survey on Different Text Categorization Techniques

Miji K Raju¹, Sneha T Subrahmanian², T.Sivakumar³

^{1,2}PG Scholar, ³Assistant Professor

^{1,2,3}Department of Computer Science and Engineering

^{1,2,3}Maharaja Institute of Technology, Coimbatore, India

mijikraju@gmail.com¹, snehats19@gmail.com², sivakumarrrts@gmail.com³

Abstract: To classify billions of documents manually is an expensive task and it is very time lagging. In internet huge amount of data are in the uncategorized form. Text Categorization is a task of automatically sorting a set of documents in to different classes from predefined set. It is mostly depends up on information retrieval and Machine Learning techniques. Text Classification (also called Text Categorization) performed on the basis of endogenous collection of data. Classifier algorithms should be used to classify various meaning of sentences. The major difficulty of this categorization approach is high dimensionality of feature space. Documents are classified on the basis of Supervised, Unsupervised or Semi-Supervised Learning. A key element is linking together of extracted data's together to form new hypothesis or facts to be analyze. This paper surveys of various methods such as Decision Tree, K –Nearest Neighbor, Bayesian Approaches, Support Vector Machine and Neural Network. This paper surveys of Text Categorization classifiers and Comparison. It also aims to various available classifiers on the basis of some criteria like complexity and performance.

Keywords: Decision Tree, K –Nearest Neighbor, Bayesian Approaches, Support Vector Machine, Neural Network, Text Categorization, Supervised Learning and Semi-Supervised Learning. Information retrieval and Machine Learning Techniques.

1. INTRODUCTION

Amount of text availability for analysis has increased hugely in recent years due to the micro blogging social networking, and various messaging systems. In this approach classify the text documents in to some classes. Classification is based on supervised form of machine learning aiming at identifying the class from a set of classes to which a selected text belongs. It is done on the basis of predefined training set. It consists of two types of approaches in Text Categorization: Rule based and Machine Learning based approaches. Rule based approach focused ones where classification rules are defined manually and documents are classified based on some rules. Machine Learning approach focused ones where classification rules or equations are defined automatically using various labeled documents. Machine learning approach has much higher recall but a lower precision than rule based approaches. It can provide conceptual views of document collections and has important applications in the real world. When categorizing a document, a computer program delimits the document as a “bag of words”. Now a day's dramatic increase in the amount of content available in digital forms give rise to a problem to manage the textual data. Various techniques used in Text Categorization applied on online news papers, online channels, e- papers, web search engines because these web technologies incorporate search and retrieval of data in the form of text. Classification problem is an activity of supervised learning, since the learning process is supervised by the knowledge of categories and of the training instances belongs to them.

1.1 DECISION TREE

Decision trees are designed for hierarchical decomposition of the data space. Trees are originally implemented in decision theory and statistics.[1] The benefits of decision tree in data mining 1) It able to handle variety of input data such as nominal, numeric and textual. 2) It processes the dataset that contain the errors and missing values.3) It is available in various packages of data mining and number of platform. In order to reduce the over fitting data, pruning is to be done. There are several different kinds of splits in the decision trees are available. The listed splits are

- Single attribute split
- Similarity-based multi-attribute split
- Dimensional-based multi-attribute split

These methods rebuild the manual categorization of training documents by constructing well defines true/false queries in the form of a tree structure where the nodes represent questions and the leaves represent the corresponding category of documents. After creating the tree a new document can easily be categorized by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf [2]. Advantage of decision trees is the fact that the output tree is easy to interpret even for the persons who are not familiar with the details of the model.

1.2 K - NEAREST NEIGHBOR

In this approach usually performed by comparing the category frequencies of the K- Nearest documents (neighbors). The evaluation of the documents is done by measuring the angle between the two feature vectors or calculating the Euclidean distance between the vectors. The advantage of this method is simplicity [3]. There is different number of training documents per category. It consists of too many documents from a comparatively large category appear under K- Nearest Neighbors and thus lead to an inadequate categorization and the risk factor increases. It is to test the degree of similarity between the documents and K training data.

The key element of this method is the availability of the similarity measure for identifying neighbors of a particular document. This method is non parametric, effective, easy for implementation. The examples of training are vectors in a multidimensional feature space, each of them have class label. The training phase of this algorithm consists only of storing the class labels and feature vectors of the training samples. It works on the principle of calculating centers again and again for each test term but Neural Networks works on the principle that it only calculates centre for one time and never update. It calculates the minimum distance with the help of Euclidean distance [4].

1.3 BAYESIAN APPROACHES

It contains two groups of Bayesian approaches in document categorization: Naïve and Non naïve Bayesian approach. The naïve part of the former is the assumption of word independence, the meaning that word is consequently that the presence of one word does not affect the presence or absence of another one and irrelevant. A disadvantage of this approach is they can only process binary feature vectors and have possibly relevant information [5].

The simple Bayesian classifier is mainly used for classification purpose. Using this approach learn profiles through the feedback collected from various websites.

Naïve Bayes Classifier results strong independence assumptions. The independence hypothesis of features make the features order is irrelevant .It has a result that the presence of one feature does not affect other features in classification tasks which make the computation of this approach is more efficient and more expressive term for the underlying probability model it would be a independent feature model. Theses classifiers can be trained powerfully by requiring a small amount of training data to estimate the parameters necessary for classification [6]. It is used for anti – spam filtering technique. It consists of two phases. The first phase has been applied for training set of data and the second phase has been applied employs the classification phase.

1.4 SUPPORT VECTOR MACHINE

In this approach have the advantages of simplicity and interpretability. Text data which are correlated with one another and organized in to linearly separable categories. It is mainly used in E-mail data classification. Compared to other techniques it should provide more robust and flexible performance. It is well suited for large amount of un-labeled data and small amount of labeled data. To solve quadratic programming and other similar types of problems, support machine is applied. The method should not need any human and machines help for tuning on a validation set of parameters, default choices. There is error estimating formulas are helpful for predicting the classification and eliminating the need of cross validation on the test and training set of data [7].

1.5 NEURAL NETWORK

It is a self –adaptive method. It means adjusting the weight themselves to the data without any specification. It should be having arbitrary accuracy. There several types of networks used for the classification task. Multilayer networks and Multilayer perceptron's are mostly used for neural network classifiers. Multilayer perceptron's have been applied successfully to solve many problems using the algorithm called Error back propagation neural network. The basic unit is neuron. Each unit receives set of inputs called X_i and associated set of weights W , corresponding to the term frequencies in the document. [8]. It consists of input and output layer; others build more sophisticated neural network with a hidden layer between two others. They can handle noisy or contradictory data very well. It consists of high computation costs. It is very to understand an average user; this may negatively influence the acceptance of these methods.

2. LITERATURE SURVEY

Senthil Kumar et al. (2016) [1] Feature selection methods are able to successfully reduce the problem of dimensionality in text categorization applications. The Process of text classification is well researched, but still there are many improvements can be made both to the classification engine itself to optimize the classification performance and feature preparation for a specific application. Research describes what adjustments should be made in specific situations is common, but a more framework is lacking. Effects of specific adjustments are also not well researched outside the original area of application.

Niharika S et al. (2012) [2] Text categorization plays a very important role in information retrieval, machine learning and text mining. It has been successful in implementing wide variety of real world applications. Key to this success have been the ever-increasing involvement of the machine learning community in text classification, which has lately

resulted in the use of the latest machine learning technology within text categorization applications.

Vishwanath Bijalwan et al.(2014) [3] Text Categorization (TC), also known as Text Classification, is the task of automatically classifying a set of text documents into different categories from a predefined set. Document belongs to exactly one of these categories, it is a single-label classification task; otherwise, it is a multi-label classification task. TC uses several tools from Information Retrieval (IR) and Machine Learning (ML) .It has received much attention in the last years from both researchers in the academic and industry developers. In this paper, we first categorize the documents using K-Nearest Neighbor based machine learning approach and then return the most relevant documents.

Shaifali Gupta et al. (2016) [4] the rapid growth of online information and data, text categorization has become one of the crucial methods for handling and standardizing text data. Different learning algorithms have been applied on text for categorization. Based on the efficiency and accuracy, KNN (K nearest Neighbor) algorithm proves itself to be very efficient algorithm as compared to any other learning algorithms. The framework of KNN with TF-IDF is studied and some changes need to be done for removing time complexity and improving accuracy so, proposed work is based on using imp-KNN (improved KNN) classifier which is helpful in splitting of training data and testing data .

Monica Bali et al. (2015) [5] It is the method of finding interesting regularities in large textual, where interesting means, hidden, non trivial previously unknown and potentially useful. The major goal of text mining is to enable users to extract information from textual resource and it deals with operation such as data mining, retrieval, classification, clustering, natural language preprocessing and machine learning techniques together to classify various patterns. A major difficulty of text categorization is high dimensionality of feature space. Feature selection is the reduction of dimensionality by selecting new attributes which is subset of old attributes

Menaka S et al. (2013) [6] Keywords are the subset of words that it contains the most important information about the content of the document. The process used to take out the important keywords from documents is known as Keyword extraction. In this proposed system keywords are extracted from documents using WordNet and TF-IDF. To select the candidate words we are using TF-IDF algorithm. Lexical database of English which is used to find similarity among the candidate words is known as WordNet. The words which have highest similarity are taken as keywords. Done experiments based on the analysis and performance of these Naive Bayes, Decision tree and K-Nearest Neighbor (KNN) algorithms.

Liwei wei et al. (2012) [7] Recent studies have revealed that emerging modern machine learning techniques are advantages to statistical models for text classification, such as SVM. In this study, we discuss the applications of the support vector machine with mix-ture of kernel (SVM-MK)to design a text classification system. Only a linear programming problem needs to be resolved and it greatly reduces the computational costs. More important, it is a transpa-rent model and the optimal feature subset can be obtained automatically. A real Chinese corpus from Fudan University is used to demonstrate the good performance of the SVM- MK.

Mr Rahul Patel et al.(2014) [8] In addition of that the efficient algorithms are also learned. According to the analyzed methods an improvement over this is suggested. In the future

proposed technique is implemented using JAVA technology and the comparative results are provided in the paper. Every supervised machine learning task, needs a initial dataset. A document can be assigned to more than one category (Ranking Classification), but based on this paper only researches on Hard Categorization are taken into consideration.

Zahid Hasan et al. (2013) [9] In this paper consists of database consisting of 200 paragraphs for each of the four different classes namely, business, sports, entertainment and politics. Frequency measure is a method as the feature to represent the characteristics of a specific paragraph. To develop a trained classifier consists of most frequent ten words of each of the paragraphs are used to develop a trained classifier. For predicting the class for an unknown paragraph automatically we are using trained classifier.

3. COMPARATIVE ANALYSIS

ALGORITHM USED	PROS	CONS
Decision Tree	It learns very fast compared to Neural networks. Easy to code. Reduce problem complexity	It has trouble dealing with noise. It is very expensive.
K- Nearest Neighbor	It achieves very good results and scales up well with the number of documents.	It requires more time for classification.
Bayesian Approach	It is simple Classifier which works very well on numerical and textual data.	Low classification performance. Performs very poorly when features are highly correlated.
Support Vector Machine	High dimensional input space. Many of the text categorization problems are linearly separable. Performance is very high.	It is very time consuming because of more parameters and requires more computation time.
Neural Network	It is used in recognizing complex patterns and performing nontrivial mapping functions. It is used in statistical modeling.	It is very hard to understand. Slow classification technique.

4. CONCLUSION

Text Categorization plays an important role in Information Retrieval, Machine Learning and Text Mining and it have been successful in tackling wide variety of real world applications. This paper gives an insight about the various methodologies that can be used for classifiers, After performing a review on different types of approaches and comparing existing methods based on various parameters it can be concluded that Support vector Machine (SVM) classifier has been recognized as one of the most effective text classification method in the comparisons of supervised machine learning algorithms. SVM have higher accuracy and can find and adjust automatically the parameter settings.


ACKNOWLEDGEMENT



The authors would like to thank Mrs. Leena P N and Mr. Sivaramakrishnan K N of Government Polytechnic College, Kunnankulam for their valuable suggestions and advices. Also we would like to thank the reviewers for the improvement of my paper.

REFERENCES

- [1] Bhavitha Varma E, Senthil Kumar B, "A Survey on Text Categorization", International Journal of Advanced Research in Computer and Communications Engineering, August 2016.
- [2] Niharika S, Sneha V Latha, Dr.Lavanya, "A Survey on Text Categorization", International Journal of Computer Trends and Technology, Vol 3, Issue 1, 2012.
- [3] Jordan pascual, Pinki Kumari, Vinay Kumar, Viswanath Bijalwan, "KNN Based Machine learning Approach for Text Document Mining", International Journal of Database Theory and Application, Vol 7, 2014.
- [4] Reena Rani, Shaijali Gupta, "Improvement in KNN Classifier (imp- KNN) for Text Categorization.", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 6, Issue 6, June 2016.
- [5] Deipali Gore, Monica Bali, "A Survey on Text Categorization with Different Types of Classification Methods", International Journal of Innovative Research in Computer", Vol 3, Issue 5, May 2015.
- [6] Menaka S, Radha N, "Text Classification using Keyword Extraction Technique", International Journal of Advanced Research in Computer Science and Software Engineering", Vol 3, Issue 12, December 2013.
- [7] Bing Wang, Bowei, Lewei Wei, "Text Classification using Support Vector Machine with Mixture of Kernel", A Journal of Software Engineering and Applications, 2012.
- [8] Mr Gaurav Sharma, Mr Rahul Patel, " A Survey on Text Mining Techniques", International Journal of Engineering and Computer Science , Vol 3, Issue 5, May 2014.
- [9] Krishnendhu Ghosh, Zahid Hasan, "A Decision Tree based Text Categorization for News Bulletin Data", Proc of Int. Conf on Emerging Trends in Engineering and Technology, 2013.
- [10] Goetz T ,Johnson D E, Oles F J, Zhang T . " A Decision Tre Based Symbolic Rule Induction System for Text Categorization", IBM Systems Journal, 2002.

ABOUT AUTHORS

	<p>MIJI K RAJU Pursuing M.E(Computer Science and Engineering) in Maharaja Institute of Technology. Have keen interest in the Artificial Neural Network, Data Mining, Big Data, Mobile Computing .Done projects in the Data Mining and GSM Technology. 2 years of Teaching experience in India.</p>
---	---

	<p>SNEHA T SUBRAHMANYAN Pursuing M.E(Computer Science and Engineering) in Maharaja Institute of Technology, Coimbatore, Tamilnadu, India. Interested in Mobile Computing, Networking and Cloud Computing</p>
	<p>T SIVAKUMAR Head of the Department and Assistant Professor in Maharaja Institute of Technology. Have keen interest in the Data Mining. Done projects in Data Mining, Data Source and Data Warehousing. 10 years of Teaching experience in India.</p>