PERFORMANCE ANALYSIS OF DIFFERENT TEXT CLASSIFICATION ALGORITHMS

Dr. N. Yuvaraj, Dharchana. K, Abhenayaa. B
Assistant Professor, UG student
Department of Computer Science and Engineering,
KPR Institute of Engineering and Technology.
Dharchana.kumar@gmail.com,abheluba29@gmail.com

Abstract

Today social media plays an important role in everyone's day-to-day life. Social media is a phrase which people use lot in these days to describe their post on sites and apps like facebook, twitter, instagram, snapchat and others. Social Network Analysis is primarily used by the companies with strong consumer focus-retail, financial, communication and marketing organization. With data mining a retailer can use point-of-sale records of customer purchase to develop products and promotions. Some of the Behaviour analysis applications are health care, education, fraud detection, customer relationship management (CRM) and others. This paper provides the survey on different text classification algorithms used for process of data analysis.

1. INTRODUCTION

There are different classification algorithms available in data mining. In particular to do text analysis, some of the text classification algorithm are decision tree algorithm, Rule based classifier, Support vector machine, Naïve Bayes classifier. Some of the domains in which text classification algorithm is used are Opinion mining, Email classification and spam filtering, News filtering and organization. The following section describes the various methods and algorithms for text classification process with their advantages and limitations.

2. FEATURE SELECTION FOR TEXT CLASSIFICATION

Feature Selection is one of the essential tasks to be performed before applying any classification algorithm. This process is especially important in text classification due to high dimensionality of text features and existence of noisy feature. Generally there are two ways of text representation. One method is represented as "bag of words" (set of words) that is independent of sequence of words. The another method is to represent text directly as strings, where each document is sequence of words. In both supervised and unsupervised learning the most common feature selection is stop-word removal and stemming. In stop word removal the common words in the documents are determined. In stemming process the same words in different forms are grouped into single word. The Classification problem makes use of class label for the feature selection process. The most common methods available for feature selection are Gini Index, Information Gain, Gain ratio.

3. DIFFERENT APPROACHES FOR TEXT CLASSIFICATION

There are two main approaches for text classification algorithms in data mining.

They are explained as follows. They are manual and rule based approach, statistical approach.

3.1 Manual approach

It is the simplest form of approach which is often referred as “bag of words”. This approach includes compiling a list of “key terms” which qualifies the type of content in question to specific topic.

3.2 Rule based approach

The rule based approach is flexible, powerful and easy to express. The text classification algorithm performs at its best, if the linguistic engine is based on true semantic technology. Rules can be written manually or generated with an automatic analysis and then validated manually.

3.2.1 Rule-based classifiers

In this classifier, set of rules is applied for classifying the data space. Generally, in text classification the left hand side of the rule is Boolean condition which is expressed as a simple conjunction of conditions. Whereas the right hand side is class label. The principle is to create a rule set such that all points in the dimension space are covered by atleast one rule. The most commonly used rule is IF…THEN rule set. The rule set is a model that is generated from training data. Inorder to generate the rules from training data the number of criteria’s are used. Among such criteria, Support and Confidence are the two most commonly used conditions for generation of rules. There are various issues with manual and rule-based approach which are listed in the following table 1.1

Rule-based approach	**Manual Approach**
Support and confidence values are not normalized	Labor intensive
Infinite chaining	Natural ambiguity of language problem
Inefficiency	Inaccurate
Opacity	Non-scalable
Possibility of contradiction	Results in false positive

TABLE 1.1

4. STATISTICAL APPROACH

The statistical approach is based on manual identification of “training set” from the data source, after feature selection process. It uses the text classification algorithms like Bayesian classifier, decision tree, support vector machine and many more algorithms. Some of them described as follows:

4.1 Decision tree

Decision trees are trees that classify the objects by sorting them based on the feature values. The core algorithm for building decision trees called **ID3** by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses *Entropy* and *Information Gain* to construct a decision tree. In particular text may use different terms for the partitioning process. Decision tree algorithm was used in Astronomy, for filtering noise from Hubble space telescope images, also it helped in star-galaxy classification, determining galaxy counts. Decision tree’s application in bio-medical engineering is to identify the features to be used in implantable devices. . Medicine, molecular biology, financial analysis, control system, remote sensing are the some other applications in which decision trees have been used. Recently, ID3 has been used for weather predictionThe following figure 1.1 represents how the decision tree classifies the weather based on training data set such as humidity, outlook, windy. An example to play tennis, the climatic condition is predicted by using decision tree which is described as follows.

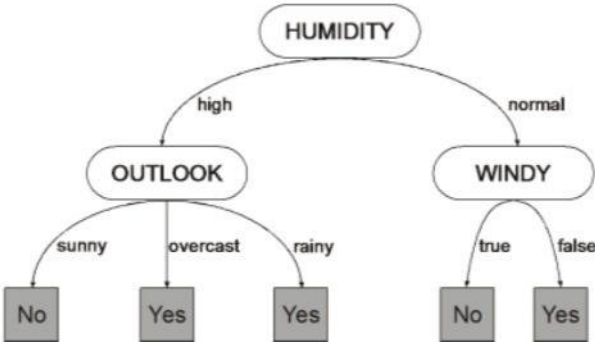


Figure 1.1

4.2 Support Vector Machine

Support Vector Machine is the supervised classification algorithm and also a form of linear classifier. It attempts to determine “good” linear separators between the different classes. The main principle of SVM is to determine separators in the search space that can best separate the different classes. The support vector machine works based on different classes and hyper-planes. Due to sparse high dimensional nature of text, text data is suited for SVM classification. In such texts certain features are irrelevant, even though they tend to correlate with one another and organized into linearly separable categories. Previously the work has been done which applies SVM method to classify the email data as spam or non-spam data. Thus support vector machine approach has been successfully used in the context of hierarchical organization of classes that often occurs in web data. An example for Support vector machine in sensor application with classification accuracy level based on number of features is given in the following graph figure 1.2

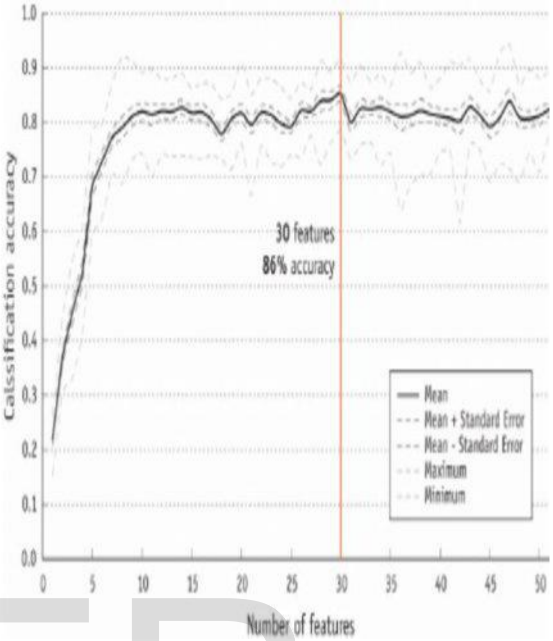


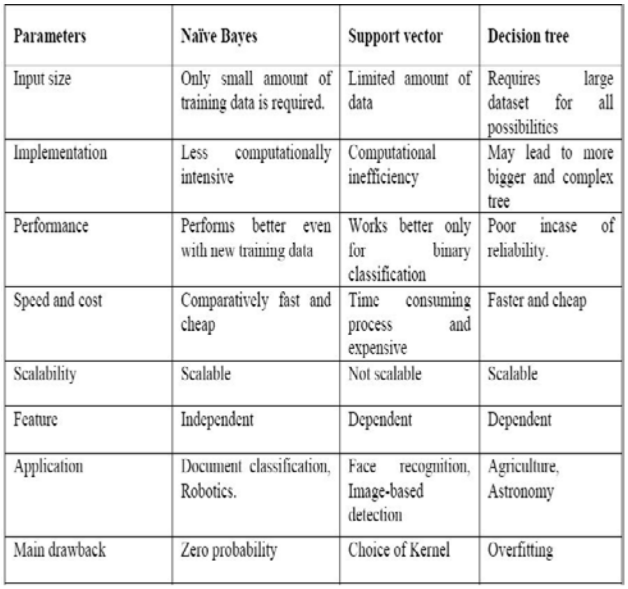
Figure: 1.2

4.3 Naive Bayes

Naive Bayes classifier is a simplest probabilistic classifier which is based on bayes theorem with strong and naïve independence assumptions. Naïve Bayes is one of the most basic text classification techniques with various applications. A classical use case for Naive Bayes is document classification: Determining whether a given (text) document corresponds to one or more categories. In the text case, the features used might be the presence or absence of key words. This approach can be implemented when system has only limited memory resources & CPU. Another important advantage of the naïve Bayes algorithm is it needs only small amount of data for training. Also this technique is more efficient where training time plays a crucial factor. Thus Naïve Bayes is used as a baseline in several researches. The Naïve Bayes classifier in one of the simplest probabilistic model works positively on text categorization and employed on Bayes rule with self feature collection works positively on text categorization and employed on Bayes rule with self-supporting feature collection. It is flexible in way of handling with any number of classes or attributes. There are many variations in this classifier some of them are as follows: Multinomial Naïve Bayes, Binarized Multinomial Naïve Bayes, Bernoulli Naïve Bayes. Among these, Binarized multinomial Bayes outperforms in case of Sentiment Analysis. Naive Bayes algorithm works based on Bayes theorem. Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

Comparison of algorithms

There are different text classification algorithms available. Some are the support vector machines (SVM), Decision tree, Maximum Entropy. The comparison of these different text classification algorithms based on certain parameters has been described in the table 1.2.



Parameters	Naïve Bayes	Support vector	Decision tree
Input size	Only small amount of training data is required.	Limited amount of data	Requires large dataset for all possibilities
Implementation	Less computationally intensive	Computational inefficiency	May lead to more bigger and complex tree
Performance	Performs better even with new training data	Works better only for binary classification	Poor incase of reliability.
Speed and cost	Comparatively fast and cheap	Time consuming process and expensive	Faster and cheap
Scalability	Scalable	Not scalable	Scalable
Feature	Independent	Dependent	Dependent
Application	Document classification, Robotics.	Face recognition, Image-based detection	Agriculture, Astronomy
Main drawback	Zero probability	Choice of Kernel	Overfitting

Table 1.2

New Naïve Bayes Approach

To avoid zero probability which is a drawback of Naïve Bayes, the small-sample correction was incorporated which is called pseudocount,in all possibility estimates such that no probability is said to be exactly zero. This way of regularizing Naïve Bayes is called Laplace Smoothing when a pseudocount is one. In general it is known as Lidstone Smoothing.

5.CONCLUSION

The different approaches for the text classification with different algorithms have been described. Along with these approaches the combination of algorithms as hybrid approaches is also proposed for automatic classification of documents. A new approach has been described which is the modified Naïve Bayes classification algorithm with Laplace Smoothing to provide more accurate and effective result for text classification process. The future work will be implementation of meta algorithms such as boosting, bagging, with the classifier algorithms to make the algorithm work efficiently with improved result.

6. REFERENCES

[1] Vapnik V (1995) “The nature of statistical learning theory”. Springer, New York

[2] Agrawal R, Srikant R (1994) “Fast algorithms for mining association rules” In: Proceeding of the 20th VLDB conference, pp 487–499

[3] Charu C. Aggarwal “ A Survey of text classification algorithms”.University of Illinois at Urbana-Champaign Urbana,IL

[4] Raj Kumar,Dr.Rajesh Varma “ Classification Algorithms for Data Mining:A Survey”International Journal of Innovations in Engineering and Technology (IJIET)Vol. 1 Issue 2 August 2012 14 ISSN: 2319 – 1058

[5] Pratiksha Y. Pawar and S. H. Gawande, Member, IACSIT “A Comparative Study on Different Types of Approaches to Text Categorization” International Journal of Machine Learning and Computing, Vol. 2 August 2012

[6] P. Bennett, N. Nguyen. “Refined experts: improving classification in large taxonomies”. ACM SIGIR Conference, 2009.

[7] S. Dumais, J. Platt, D. Heckerman, M. Sahami. “Inductive learning algorithms and representations for text categorization”. CIKM Conference, 1998.

[8]R. Gilad-Bachrach, A. Navot, N. Tishby. “Margin based feature selection – theory and algorithms”.ICML Conference, 2004.A Survey of Text Classification Algorithms 217

[9] E.-H. Han, G. Karypis, V. Kumar. “Text Categorization using Weighted-Adjusted k-nearest neighbor classification”, PAKDD Conference,2001.

[10]T. Joachims. “A Statistical Learning Model of Text Classification for Support Vector Machines”.
ACM SIGIR Conference, 2001.