Text Classification using Naïve Bayes

Kamaljeet Kaur
Department of Computer Science & Engineering
Guru Nanak Dev University Regional Campus Jalandhar, Punjab, India

Abstract— In text mining, one of the challenging and growing importance's is given to the task of document classification or text characterization. In this process, reliable text extraction, robust methodologies and efficient algorithms such as Naïve Bayes and other made the task of document classification to perform consistently well. Classifying text documents using Bayesian classifiers are among the most successful known algorithms for machine learning. Text mining refers to the process of taking high-quality information from text one of the classification method that can be used is Naïve Bayes Classifier.

Key words: Naïve Bayes, Text Classification, Hypothesis

I. INTRODUCTION

Text classification is the task of classifying documents by their content: that is, by the words of which they are comprised. Perhaps the best-known current text classification problem is email spam filtering: classifying email messages into spam and non-spam (ham). The NB approach, is one of the most effective and straightforward method for text document classification and has exhibited good results in previous studies conducted for data mining. The task is to assign a document to one or more categories and sub- or subjective categories, based on its text contents. There are two types of classification: supervised and unsupervised. Supervised classification is based on external source, for example, human feedback, which provides information on the correct classification. On the contrary, in un-supervised classification, the processing must be performed entirely without any external information.

II. HOW TO WORK TEXT CLASSIFICATION

Text classification (TC) is an important part of text mining.

– Input

a document d

a fixed set of classes C={c1,c2,.. cn}.

– e.g =(sport, Polities)
– Simple ("naive")classifica1on method Based on Bayes rule.
– Relies on Very simple represent a on of document(Bag of words)

A. *Bayes' Rule Applied to Documents & Classes*

For a document d and a class c

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

III. TEXT CLASSIFICATION PROCESS

The stages of TC are discussing as following points.

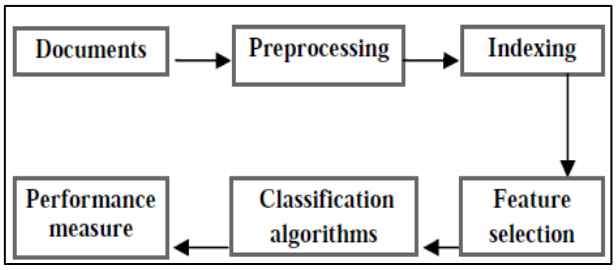

Fig. 1: Document Classification Process

A. *Documents Collection*

This is first step of classification process in which we are collecting the different types (format) of document like html, .pdf, .doc, web content etc.

B. *Pre-Processing*

The first step of pre-processing which is used to presents the text documents into clear word format. The documents prepared for next step in text classification are represented by a great amount of features. Commonly the steps taken are:

1) Tokenization

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

2) Removing Stop Words

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

3) Stemming Word

This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute.

C. *Indexing*

The documents representation is one of the pre-processing technique that is used to reduce complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector.

D. *Feature Selection*

After pre-processing and indexing the important step of text classification, is feature selection to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier.

E. *Classification*

The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi supervised methods

F. *Performance Evaluations*

Many measures have been used, like Precision and recall.

– Precision wrt ci (Pri) is defined as the as the probability that if a random document dx is classified under ci, this decision is correct.
– Recall wrt ci (Rei) is defined as the conditional that, if a random document dx ought to be classifie.d under ci, this decision is taken.

IV. TEXT CLASSIFICATION USING DOCUMENT MODELS

A. Document Models

Text classifiers often don't use any kind of deep representation about language: often a document is represented as a bag of words. (A bag is like a set that allows repeating elements.)
Two Type of Document model

B. Bernoulli Document Model

A document is represented by a feature vector with binary elements. Taking value 1 if the corresponding word is present in the document and 0 if the word is not present.

C. Multinomial Document Model

A document is represented by a feature vector with integer elements whose value is the frequency of that word in the document.

D. The Bernoulli Document Model

The Bernoulli model a document is represented by a binary vector, which represents a point in the space of words. If we have a vocabulary V containing a set of |V| words, then the t th dimension of a document vector corresponds to word wt in the vocabulary. Let bi be the feature vector for the i th document Di; then the t th element of bi, written bit, is either 0 or 1 representing the absence or presence of word wt in the i th document.

Let P(wt |C) be the probability of word wt occurring in a document of class C; the probability of wt not occurring in a document of this class is given by (1−P(wt |C)). If we make the naive Bayes assumption, that the probability of each word occurring in the document is independent of the occurrences of the other words, then we can write the document likelihood P(Di | C) in terms of the individual word likelihoods P(wt |C)

1) Define the vocabulary V; the number of words in the vocabulary defines the dimension of the feature vectors
2) Count the following in the training set:

– N the total number of documents
– Nk the number of documents labelled with class C=k, for k=1, . . . , K
– nk(wt) the number of documents of class C=k containing word wt for every class and for each word in the vocabulary

3) Estimate the likelihoods P(wt | C=k) using equation
4) Estimate the priors P(C=k) using equation

Example: Target Concept Interesting.
Document: (Positive or Negative).

DOC	Text	Class
1	I Loved the movies	+
2	I hated the movies	-
3	A great movies, good movies,	+
4	Poor acting	-
5	Great acting, a good movies	+

Table 1:

Find Unique Word.
I, Loved, the, movie, hated, a great, poor, acting, good=10

DOC	I	Loved	the	movie	hated	a	great	poor	acting	good
1	1	1	1	1						
2	1		1	1	1					
3			2			1	1			1
4								1	1	
5			1			1	1		1	1

Table 2:

Let's look at the probability per outcome (+ or -).
Document with Positive Outcome.

DOC	I	Loved	the	movie	hated	a	great	poor	acting	good
1	1	1	1	1						
3				2		1	1		1	
5				1		1	1		1	1

Table 3:

P(+)=3/5=0.6
Compute:P(I|+);P(Loved|+);P(the|+);P(movie|+);P(hated|+); P(a|+);P(great|+);P(poor|+);P(acting|+);P(good|+).
Let 'n' be the no.of word in the (+).
nk the no.of time word 'k' occurs in these cases(+).

$$\text{Let } P(Wk|+)\frac{nk+1}{n+\text{vocabulary}}$$

P(I|+):$\frac{1+1}{14+10}=\frac{2}{24}$=0.0833 P(Loved|+):$\frac{1+1}{14+10}$=0.0833.
P(the|+):$\frac{1+1}{14+10}=\frac{2}{24}$=0.0833; P(Movies|+):$\frac{4+1}{14+10}$=0.2083;
P(a|+):$\frac{2+1}{14+10}$=0.125; P(great|+):$\frac{2+1}{14+10}$=0.120;
P(acting|+):$\frac{1+1}{14+10}$=0.0833; P(good|+):$\frac{2+1}{14+10}$=0.125;
P(hated|+):$\frac{0+1}{14+10}$=0.0417; P(Poor|+):$\frac{0+1}{14+10}$=0.0412;

Now . Let's look at the Negative Document.

DOC	I	Loved	the	movie	hated	a	great	poor	acting	g o o	
2	1		1	1	1						
4								1	1		

Table 4:

P(-)=$\frac{2}{5}$=0.4

P(I|-):$\frac{1+1}{6+10}$=0.125; P(the|-):$\frac{1+1}{6+10}$=0.125;
P(movie |-)$\frac{1+1}{6+10}$=0.125; P(hated |-):$\frac{1+1}{6+10}$=0.125;
P(Poor|-):$\frac{1+1}{6+10}$=0.125; P(acting|-):$\frac{1+1}{6+10}$=0.125;
P(Loved|-):$\frac{0+1}{6+10}$=0.0625; P(a|-):$\frac{0+1}{6+10}$=0.625;
P(great|-):$\frac{0+1}{6+10}$=0.0625; p(good |-):$\frac{0+1}{6+10}$=0.0625;

Now that we have trained our classifier,
Let's classify a new sentence according to:

$$v_{NB}=\frac{\text{argmax}}{vjv}\prod_w P(w|vj)$$

If v_j= - p(-) p(I/-)P(HATED/-)P(the/-)p(poor/-)p(acting/-)=$1.22*10^{-5}$.
If v_j =+ p(+)p(I|-)P(HATED|+)P(the /+)p(poor/+)p(acting/+)=$6.03*10^{-7}$

V. CONCLUSION

We have shown how the Naive Bayes approximation can be used for document classification, by constructing distributions over words. The classifiers require a document model to estimate P(document | class). We looked at two document models that we can use with the Naïve Bayes approximation: The growing use of the textual data witch needs text mining, machine learning and natural language processing techniques and methodologies to organize and extract pattern and knowledge from the documents. This review focused on the existing literature and explored the documents representation and an analysis of feature selection methods and classification algorithms were presented. It was verified from the study that information Gain and Chi square statistics are the most commonly used and well performed methods for feature selection, however many other FS methods are proposed. This paper also gives a brief introduction to the various text representation schemes.

REFERENCES

[1] F. Sebastiani, "Text categorization", Alessandro Zanasi (ed.) Text Mining and its Applications, WIT Press, Southampton, UK, pp. 109-129, 2005.
[2] Kjersti Aas & Line Eikvil "Text Categorization: A Survey" Report No. 941. ISBN 82-539-0425-8. , June, 1999.
[3] Dasgupta, P. Drineas, B. Harb, "Feature Selection Methods for Text Classification", KDD'07, ACM, 2007.
[4] Muhammed Miah, "Improved k-NN Algorithm for Text Classification", Department of Computer Science and Engineering University of Texas at Arlington, TX, USA.
[5] Vidhya. K.A G.Aghila, "A Survey of Naive Bayes Machine Learning approach in Text Document Classification", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, 2010.
[6] Bayes Jingnian Chen a, b, Houkuan Huang a, Shengfeng Tian a, Youli Qua a "Feature selection for text classification with Naive", China Expert Systems with Applications 36 5432–54352009.
[7] D. E. Johnson F. J. Oles T. Zhang T. Goetz, "A decision-tree-based symbolic rule induction system for text Categorization", by IBM SYSTEMS JOURNAL, VOL 41, NO 3, 2002.
[8] Anirban Dasgupta "Feature Selection Methods for Text Classification "KDD'07, August 12–15, 2007.
[9] Wei Zhao "A New Feature Selection Algorithm in Text Categorization "International Symposium on Computer, Communication, Control and Automation 2010.
[10] Fang Lu Qingyuan Bai, "A Refined Weighted K-Nearest Neighbours Algorithm for Text Categorization", IEEE 2010
[11] Yiming Yang "An Evolution of statistical Approaches to Text Categorization" Information Retrieval 1, 69-90 1999.
[12] Dino Isa "Text Document Pre-Processing Using the Bayes Formula for Classification Based on the Vector Space Mode", Computer and Information Science November, 2008

