# Feature Selection for text classification with Naïve Bayes

# 

***Submitted By***

- Suman Chatterjee.

Admn No.-15je001400.

**UNDER THE GUIDANCE OF**

***Dr. A.C.S RAO***

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**INDIAN INSTITUTE OF TECHNOLOGY(INDIAN SCHOOL OF MINES),**

**DHANBAD.**

**DHANBAD-826004.**

**Dec, 2018.**

## ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude and respect towards my project guide, Dr. ACS Rao, Assistant Professor, Department of Computer Science And Engineering, IIT(ISM) DHANBAD.

I am very much indebted to him for the generosity, expertise and guidance I have received from him while working on the project. Also, to Mayur Wankhede Sir who guided through the specific details of the project.

I wish to express my profound gratitude to Dr. Haider Banka, Head of Department Of Computer Science And Engineering, IIT(ISM) DHANBAD for his inspiration.

DATE-01/12/2018 NAME-SUMAN CHATTERJEE. ADMISSION NO.-15JE001400

## Abstract

The origin of text classification goes back to the early '60s [Sebastiani, 2002]. In the late '90s, Machine Learning techniques were successfully applied to text classification. Support Vector Machines were applied to text classification in [Joachims, 1998; Dumais et al., 1998]. Maximum Entropy Models were also applied in [Nigam et al., 1999].

Multi-label classification of multi-topic text has been investigated in the last years. AdaBoost was enhanced to handle multi-labels in [Schapire and Singer, 2000]. In this approach, the task of assigning multi-topics to a text is regarded as a ranking of labels for the text. This ranking-based evaluation was inspired by Information Retrieval. In a text classification problem, however, we need a definite set of topics for each document rather than the rankings of topic candidates. [McCallum, 1999] proposed to use the EM algorithm to train a mixture model of multi-labels. Parametric Mixture Models (PMM) were also proposed in [Ueda et al., 2002]. Maximum Entropy Models were extended to multi-labelled MEMs (MLME) in [Zhu et al., 2005].

As an important preprocessing technology in text classification, feature selection can improve the scalability, efficiency and accuracy of a text classifier. In general, a good feature selection method should consider domain and algorithm characteristics. As the Naïve Bayesian classifier is very simple and efficient and highly sensitive to feature selection, so the research of feature selection especially for it is significant.

***INTRODUCTION***

Due to the proliferated availability of texts in digital form and

the increasing need to access them in flexible ways, text classification

becomes an elementary and crucial task. In the past several

years, many methods based on machine learning and statistical

theory have been applied to text classification. Among this kinds of methods, **decision trees** (Lewis & Ringuette, 1994), **k-nearest neighbors (kNN)** (Cover & Hart, 1967; Tan, 2005; Yang, 1997; Yang & chute, 1994), **neural networks** (Wiener, Pedersen, & Weigend, 1995), **Naïve Bayes** (Lewis, 1998; McCallum & Nigam, 1998) and **support vector machines** (SVM) (Joachims, 1998) are all successful examples. As one of these successful methods, Naïve Bayes is popular in text classification due to its computational efficiency and relatively good predictive performance. For text classification a major problem is the high dimensionality of the feature space. It is very often that a text domain has several tens of thousands of features. Hence feature selection is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers. According to John, Kohavi, and Pfleger (1994) there are mainly two types of feature selection methods in machine learning: **wrappers and filters**.

* *Wrappers* use the classification accuracy of some learning algorithm as their evaluation function. Since wrappers have to train a classifier for each feature subset to be evaluated, they are usually much more time consuming especially when the number of features is high. So wrappers are generally not suitable for text classification.
* As opposed to wrappers, *filters* perform feature selection independently of the learning algorithm that will use the selected features. In order to evaluate a feature, filters use an evaluation metric that measures the ability of the feature to differentiate each class.

In general filters are much less time consuming than wrappers and have been widely used in text classification. Many feature evaluation metrics have been explored, notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index,etc.

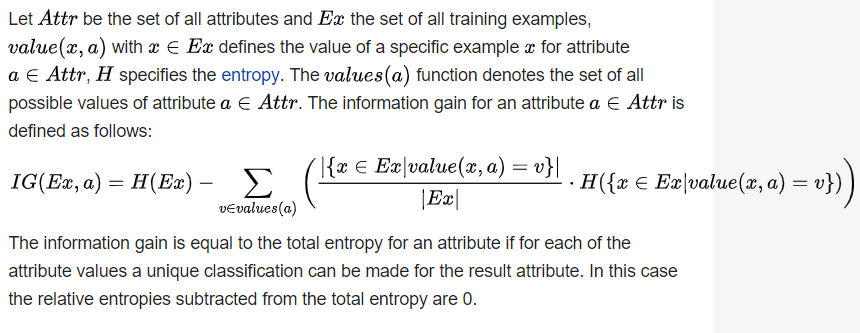
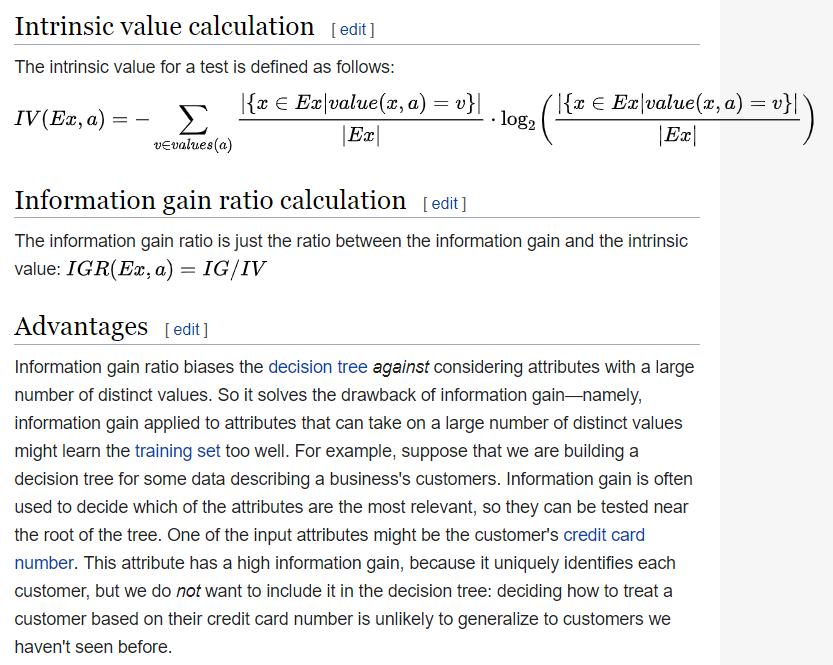
***Feature evaluation metrics for Naïve Bayes classifiers***

The Naïve Bayesian classifier is very simple and efficient. But it is highly sensitive to feature selection. So the study of feature evaluation metrics for it is very necessary. Mladenic and Grobelnik (1999) presented some good results in this area. They found that on all their binary-class domains Odds Ratio was among the best performing measures for the Naïve Bayes classifier. As numerous text datasets are multi-class, it is natural to adapt Odds Ratio for multi-class problems. As we have not expected that the directly extending of it performs badly. And so we present two effective metrics for the Naïve Bayes classifier applied on multiclass datasets: Multi-class Odds Ratio (MOR) and Class Discriminating Measure (CDM).

A good feature selection metric should consider problem domain and algorithm characteristics. A given evaluation metric may get much different selecting effect for different domain or algorithms. For example, Yang and Pedersen (1997) reported that IG was one of the best metric in their experiments. However, Mladenic and Grobelnik (1999) found that IG got worst selecting effect on the domain they studied. Besides the difference of problem domain, one important reason that leads to the dis-agreement in the performance of IG is the difference of algorithms used. In this paper, we will do some research on feature evaluation metrics specially for the Naïve Bayesian classifier applied on text data, which is very simple and efficient and highly sensitive to feature selection.

***Information Gain***

In decision tree learning, **Information gain ratio** is a ratio of information gain to the intrinsic information. It was proposed by Ross Quinlan, to reduce a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute.



# The Multi-class Odds Ratio(MOR) and Class Discriminating Measure(CDM) Metrics

# The MOR Metric:-

The traditional Odds Ratio for binary-class domains is defined as:

# 

where **P(w|pos) is the probability of the occurrence of word w in the positive class**, and **P(w|neg) is the probability that word w occurs in the negative class**. The Odds Ratio extended directly for multi-class domains can be in the following two forms, named as Extended Odds Ratio (EOR) and Weighted Odds Ratio (WOR)

# 

where *P(cj) is the probability of the jth class value, P(w|cj) is the probability that word w occurs if the class value is j, and P(w|cj) is the probability that word w occurs when the class value is not j*. We can see that both EOR and WOR only prefer positive features. In domains where positive features dominate the classification results, just like the case described by Mladenic and Grobelnik (2003), the positive-feature-preferring metrics usually perform well. But for **multi-class text data**, negative features can usually contribute to the classification results. And positive-feature-preferring metrics may not perform well under such situation. Hence we ***present the MOR metric as***

# 

It can be seen from (4) that MOR metric prefers not only positive features but also negative features with high value of P(wi|cj) ***And so it can perform well.***

It should be pointed out that Zhou proposed MC-OR metric similar

# to MOR (Zhou, Zhao, & Hu, 2004). It is defined as

# 

The difference between MC-OR and MOR is that MC-OR weighted each term in (4) with class distribution, and so give more emphasis to features in large classes. This will worsen the classification effect for small classes that are often in the majority of a multi-class text data. And so, as experimental results indicate in *Section 4****, MOR performs obviously better than MC-OR.***

# The CDM Metric

# 

# Naïve Bayes Classifiers used on Text Data

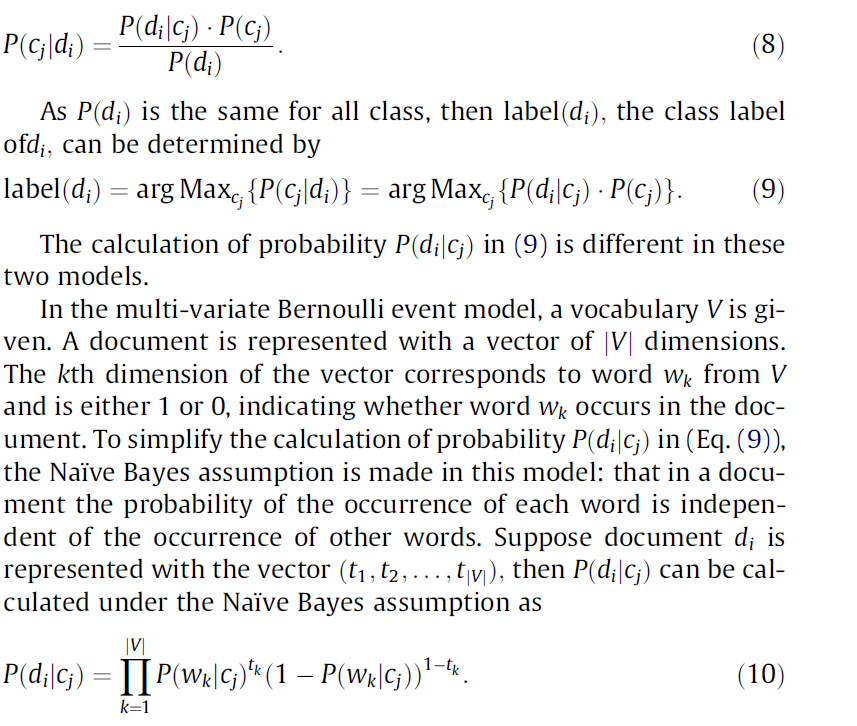
In the area of text classification there are two different models of Naïve Bayes classifiers in common use: the **Multi-Variate Bernoulli**

**Event Model and the Multinomial Event Model** (McCallum &

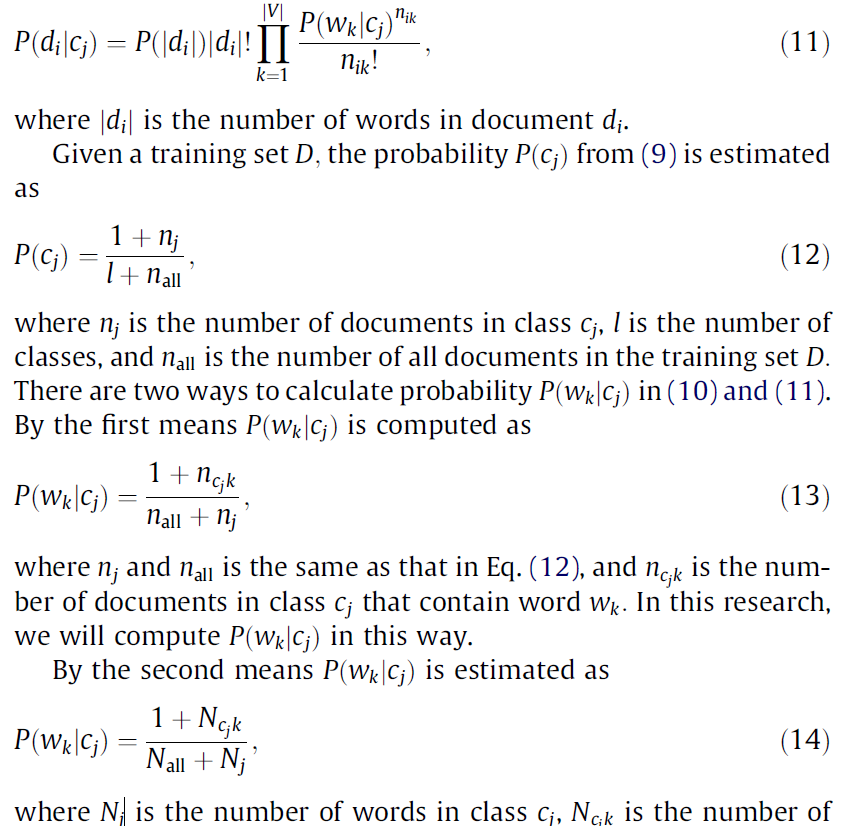
Nigam, 1998). Both of these two models use the Bayes rule to classify

a document. Given a document di; the probability of each class

cj is calculated as

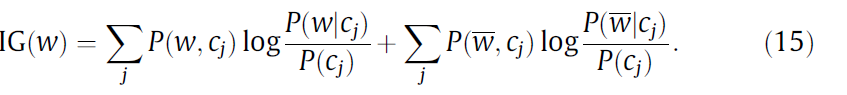


In the multinomial event model, a document is regarded as ‘‘a bag of words”. No order of the words is considered, but the frequency of each word in the document is captured. In this model, a similar Naïve Bayes assumption is made: that the probability of the occurrence of each word in a document is independent of the word’s position and the occurrence of other words in the document. Denote the number of times word wk occurs in document di as nik: Then the probability P(di|cj) from Eq. (9) can be computed by

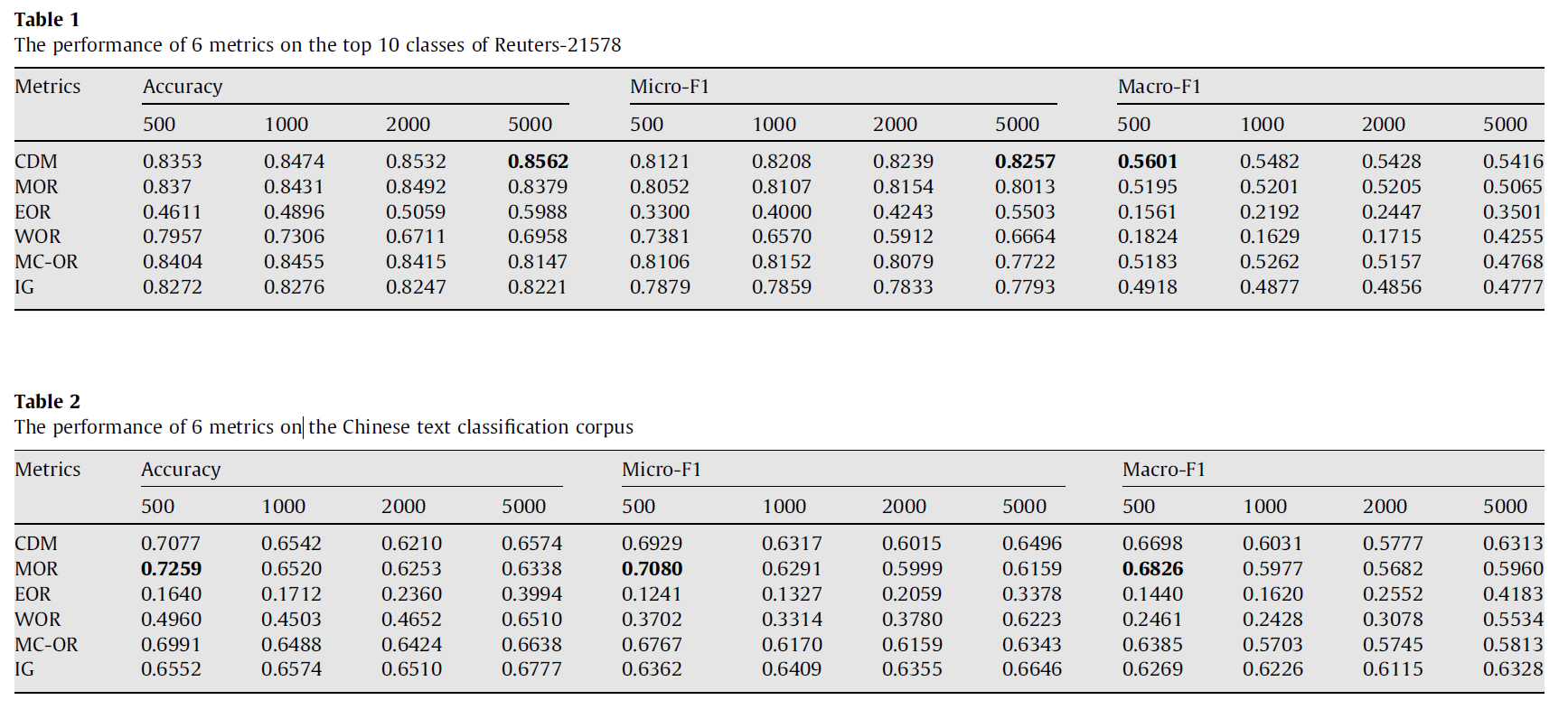


word wk in class cj; and Nall is the number of all words in the training

set D.



***Tables For Results***



***Conclusion***

This paper presents two feature evaluation metrics (CDM and

MOR) for the Naïve Bayes classifier applied on multi-class text collections.

We compared CDM and MOR with EOR, WOR and MC-OR, three variations of Odds Ratio for multi-class datasets. We also compare them with IG, which is usually among the best performing metrics for many text datasets. Experimental results on two data sets show that CDM and MOR are among the best performing metrics for the Naïve Bayes classifier applied on multi-class text datasets. Moreover, the computation of CDM metric is simpler than other feature evaluation metrics.

For future work, we will experiment on more multi-class datasets.

Furthermore, we will explore feature selection metrics for high skewed datasets and text collections containing unlabeled documents.

***References And Bibliography***

1. Feature selection for text classification with Naïve Bayes

Jingnian Chen a,b,\*, Houkuan Huang a, Shengfeng Tian a, Youli Qua

* School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
* Department of Information and Computing Science, Shandong University of Finance, Jinan, Shandong, 250014, China

1. Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE

Transaction on Information Theory IT, 13(1), 21–27.

1. Forman, G. (2003). An extensive empirical study of feature selection metrics for text

classification. Journal of Machine Learning Research, 3, 1289–1305.

1. Frank, E., & Bouchaert, R. R. (2006). Naive Bayes for text classification with

unbalanced classes. In Proceedings of the 10th European conference on principles

and practice of knowledge discovery in databases (pp. 503–510). Berlin: Springer.

Joachims, T. (1998). Text categorization with support vector machines: Learning

with many relevant features. In Proceedings of the 10th European conference on

machine learning (pp. 137–142). New York: Springer.

1. John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset

Selection Problem. In Proceedings of the 11th International Conference on machine

learning (pp. 121–129). San Francisco: Morgan Kaufmann.

1. Kim, S., Han, K., Rim, H., & Myaeng, S. (2006). Some effective techniques for Naive

Bayes text classification. IEEE Transactions on Knowledge and Data Engineering,

18(11), 1457–1466.

1. Lewis, D. D. (1998). Naive Bayes at forty: The independence assumption in

information retrieval. In Proceedings of the 10th European conference on machine

learning (pp. 4–15). New York: Springer.

1. Lewis, D.D., & Ringuette, M. (1994). Comparison of two learning algorithms for text

categorization. In Proceedings of the third annual symposium on document

analysis and information retrieval (pp. 81-93). Las Vegas, NV.

1. McCallum, A. & Nigam, K. (1998). A comparison of event models for naive Bayes text

classification. In AAAI-98 workshop on learning for text categorization.

1. Mladenic, D., & Grobelnik, M., 1999. Feature selection for unbalanced class

distribution and Naive Bayes. In Proceedings of 16th international conference on

machine learning (pp. 258–267). San Francisco.

1. Mladenic, D., & Grobelnik, M. (2003). Feature selection on hierarchy of web

documents. Decision Support Systems, 35(1), 45–87.

1. Shang, W., Huang, H., Zhu, H., Lin, Y., et al. (2007). A novel feature selection

algorithm for text categorization. Expert System with Applications, 33(1), 1–5.

1. Tan, S. (2005). Neighbor-weighted K-nearest neighbor for unbalanced text corpus.

Expert System with Applications, 28(4), 667–671.

1. Wiener, E. D., Pedersen, J. O., & Weigend, A. S. (1995). A neural network approach to

topic spotting. In Proceedings of SDAIR-95, 4th annual symposium on document

analysis and information retrieval (pp. 317–332).

1. Yang, Y. (1997). An evaluation of statistical approaches to text categorization.

Information Retrieval, 1(1), 76–88.

1. Yang, Y., & Chute, C. G. (1994). An example-based mapping method for text

categorization and retrieval. ACM Transactions on Information System, 12(3),

252–277.

1. Yang, Y., & Pedersen, J.O. (1997). A Comaprative Study on Feature Selection in Text

Categorization. In Proceedings of the 14th international conference on machine

learning (pp. 412-420). Nashville, USA.

1. Zhou, Q., Zhao, M., & Hu, M. (2004). Study on feature selection in Chinese text

categorization. Journal of Chinese Information Processing, 18(3), 17–23.

1. https://en.wikipedia.org/wiki/Information\_gain\_ratio#Information\_gain\_calculation