Spam Filtering with Naive Bayes - Which Naive Bayes ?

Kaoruko Kawamoto

Keio University

May 1, 2023

Thesis Information

- Title: Spam Filtering with Naive Bayes – Which Naive Bayes?
- Author: Metsis, V., Androutsopoulos, I., Paliouras, G.
- Date: Accepted 28 July 2006
- Journal: 3rd Conference on Email and Anti-Spam - Proceedings, CEAS 2006

abstract

- A spam filter which distinguishes and classifies spam e-mails as legitimate e-mails is implemented using Naive Bayes classifier.
- When using Naive Bayes, we implemented the classifier using five kinds of probability distributions and recorded their recall rate and ROC curve as their performances.
- As a result of implementation, all classifiers were able to maintain a high recall rate regardless of the type of datasets, and it turned out that the multinomial Naive Bayes classifier using Boolean attributes had the highest performance.

Contents

1.1 Introduction About Spam Filtering

- Generally, the spam filter scans whether or not an e-mail contains specific words which tend to appear in a spam e-mail, and classifies it based on the result.
- ex)"special offer","dating service"

- Naive Bayes Classifer is one of the particulary popular in commercial and open-source spam filters because its structure is very simple.

2.0 Naive Bayes Classifier -Variable Explanation-

$$X = (x_1, x_2, x_3, , , x_i, , , x_m)$$

x_i : the values of attributes, tokens of the message
In the case of Boolean counts,

$$x_i = \begin{cases} 1, & \text{the token is in the message} \\ 0, & \text{the token is not in the message} \end{cases}$$

TF (Term Frequency) : how many times the token occurs in the message
normalized TF : divide TF by the total number of token occurrences in the message
m : number of attributes (tokens)

2.1 Naive Bayes Classifier -Main Structure-

From Bayes' theorem, the probability that a message with vector X belongs in category c is :

$$p(c|X) = \frac{p(c)p(X|c)}{p(X)}$$

The classification is done according to the inequality below :

$$\frac{p(c_s)p(X|c)}{p(c_s)p(X|c_s) + p(c_h)p(X|c_h)} > T$$

T : a threshold

c_s : The message is a spam e-mail.

c_h : The message is a ham (legitimate) e-mail.

2.2 Naive Bayes Classifier -Versions of NB-

1. Multi - variate Bernoulli NB
2. Multinomial NB, TF attributes
3. Multinomial NB, Boolean attributes
4. Multi - variate Gauss NB
5. Flexible Bayes

2.3 NB Classifier -Multi - variate Bernoulli NB-

$p(X|c)$ can be computed as :

$$p(X|c) = \prod_{i=1}^{m} p(t_i|c)^{x_i}(1 - p(t_i|c))^{(1-x_i)}$$

t_i : a member of the tokens list $F = (t_1, t_2, , , t_m)$
Then, the criterion for classifying a message becomes :

$$p(c|X) = \frac{p(c_s)\prod_{i=1}^{m} p(t_i|c_s)^{x_i}(1 - p(t_i|c_s))^{(1-x_i)}}{\sum_{c \in c_s, c_h} p(c)\prod_{i=1}^{m} p(t_i|c)^{x_i}(1 - p(t_i|c))^{(1-x_i)}} > T$$

$p(t|c)$ is estimated based on a Laplacean prior and computed below :

$$p(t|c) = \frac{1 + M_{t,c}}{2 + M_c}$$

$M_{t,c}$: the number of training messages of category c
M_c : the total number of training messages of category c

2.4 NB Classifier -Multinomial NB, TF attributes-

$p(X|c)$ can be computed as :

$$p(X|c) = p(|d|)|d|! \prod_{i=1}^{m} \frac{p(t_i|c)_i^x}{x_i!}$$

$|d|$: total values of t_i found in the message
Then, the criterion for classifying a message becomes :

$$\frac{p(c_s) \prod_{i=1}^{m} p(t_i|c_s)^{x_i}}{\sum_{c \in c_s, c_h} p(c) \prod_{i=1}^{m} p(t_i|c)^{x_i}} > T$$

$p(t|c)$ is estimated based on a Laplacean prior and computed below :

$$p(t|c) = \frac{1 + N_{t,c}}{m + N_c}$$

$N_{t,c}$: the number of occurrences of token t in the training message of category c
$N_c = \sum_{i=1}^{m} N_{t_i,c}$

2.5 NB Classifier -Multinomial NB, Boolean attributes-

- This type of NB is the same multinomial NB with TF attributes.
- The NB classifier is with Boolean attributes, but it does not take account into the case $x_i = 0$
- multinomial NB with TF attributes is equivalent to NB with attributes modelled as following Poisson distributions in each category.

2.6 NB Classifier -Multi - variate Gauss NB-

In the case of assuming each attribute follows a normal distribution in each category c, the normal distribution is expressed as follows :

$$g(x_i; \mu_{i,c}, \sigma_{i,c}) = \frac{1}{\sigma_{i,c}\sqrt{2\pi}} e^{-\frac{1}{(x_i-\mu_{i,c})^2}2\sigma_{i,c}^2}$$

$\mu_{i,c}, \sigma_{i,c}$: estimated from the training data

$$p(X|c) = \prod_{i=1}^{m} g(x_i; \mu_{i,c}, \sigma_{i,c})$$

and the criterion for classifying a message becomes :

$$\frac{p(c_s)\prod_{i=1}^{m} g(x_i; \mu_{i,c_s}, \sigma_{i,c_s})}{\sum_{c\in c_s,c_h} p(c)\prod_{i=1}^{m} g(x_i; \mu_{i,c}, \sigma_{i,c})} > T$$

2.7 NB Classifier -Flexible Bayes-

In the case of assuming the occurrences of tokens do not follow the normal distributions, FB models $p(x_i|c)$ and it is expressed as follows :

$$p(x_i|c) = \frac{1}{L_{i,c}} \Sigma_{i=1}^{L_{i,c}} g(x_i; \mu_{i,c}, \sigma_{i,c})$$

$L_{i,c}$: the number of different values which X_i has in the training data of category c
$p(x_i|c)$ means the average of $L_{i,c}$

3.1 Datasets and Methodology -Datasets-

In creating the datasets, 6 Enron's employees' e-mail datasets and 3 different public external data source (SH, BG, GP) are combined.

Table 1: Composition of the six benchmark datasets.

ham + spam	ham:spam	ham, spam periods
farmer-d + GP	3672:1500	[12/99, 1/02], [12/03, 9/05]
kaminski-v + SH	4361:1496	[12/99, 5/01], [5/01, 7/05]
kitchen-l + BG	4012:1500	[2/01, 2/02], [8/04, 7/05]
williams-w3 + GP	1500:4500	[4/01, 2/02], 12/03, 9/05]
beck-s + SH	1500:3675	[1/00, 5/01], [5/01, 7/05]
lokay-m + BG	1500:4500	[6/00, 3/02], [8/04, 7/05]

SH : mixture of a small scale dataset and covering a short period dataset

BG : recent dataset

GP : the only dataset which contains duplicated spam e-mails

ham-spam ratio is approximately 3:1, or 1:3

3.2 Datasets and Methodology -Methodology-

1. preprocessing the datasets
   - remove the message sent by the owner of the mailbox
   - remove html tags in the message
   - remove spam messages written in non - Latin characters
2. accumulate messages in the order they were sent regardless of ham or spam.
   - The purpose is to change the ham spam ratio over time and make it difficult to predict.
3. retrieve e-mails every new 100 messages at regular intervals. Carry out training every 100 new messages.
4. use spam recall, ham recall, ROC curve
   - spam recall : the proportion of spam messages correctly blocked
   - ham recall : the proportion of ham messages correctly passed

4.1 Experimental Results -Size of Attribute Set-

NB version	Enr1	Enr2	Enr3	Enr4	Enr5	Enr6
FB	7.87	3.46	1.43	1.31	0.11	0.34
MV Gauss	5.56	4.75	1.97	12.7	3.36	5.27
MN TF	0.88	0.95	0.20	0.50	0.75	0.18
MV Bernoulli	2.10	0.95	1.09	0.45	1.14	0.88
MN Boolean	2.31	1.97	2.04	0.43	0.39	0.20

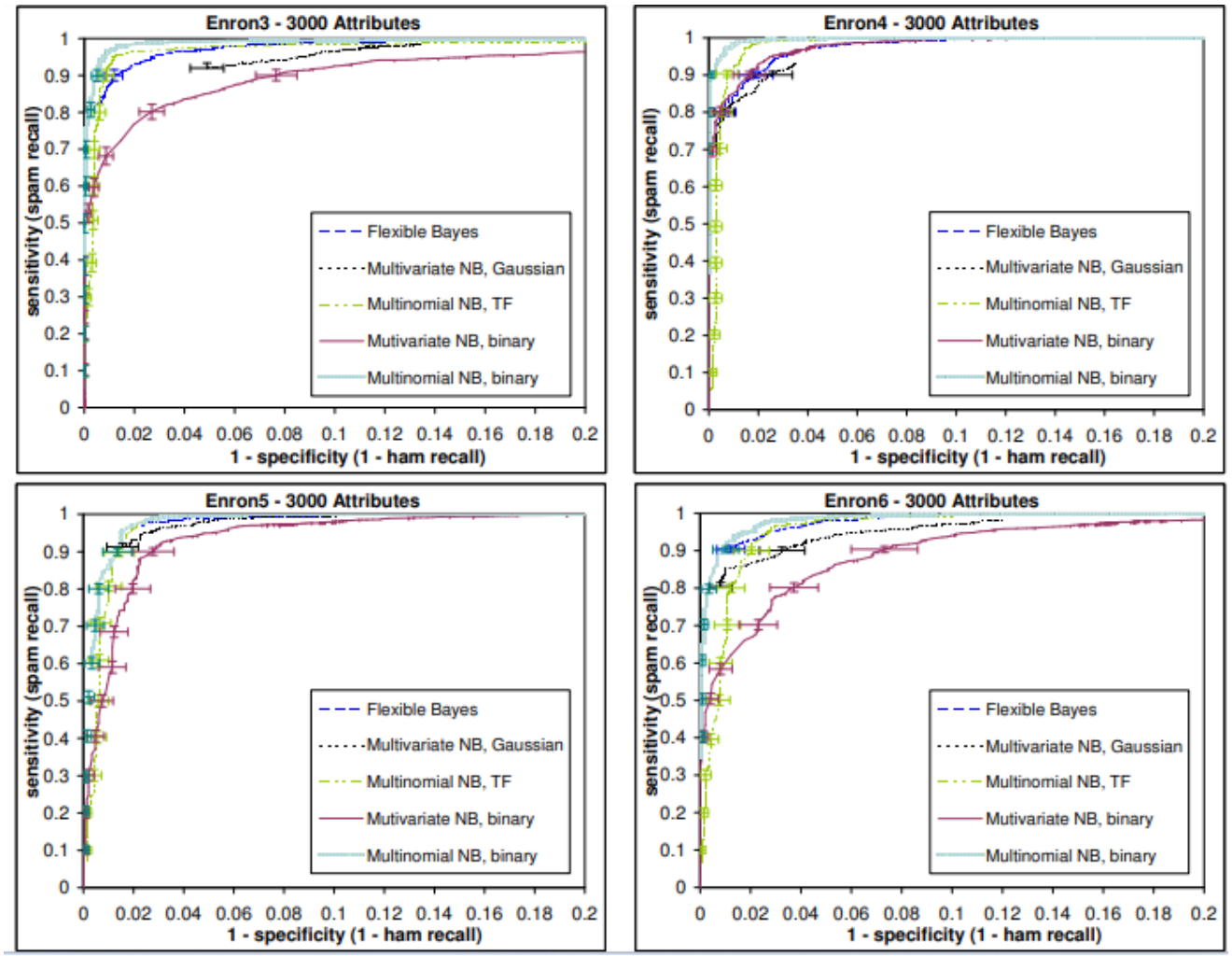
Table 2: Maximum difference ($\times 100$) in spam recall across 500, 1000, 3000 attributes for $T = 0.5$.

NB version	Enr1	Enr2	Enr3	Enr4	Enr5	Enr6
FB	0.61	0.23	1.72	0.54	0.48	0.34
MV Gauss	1.17	0.75	5.94	1.77	5.91	4.88
MN TF	2.17	1.38	1.02	0.61	1.70	1.22
MV Bernoulli	1.47	0.63	6.37	2.04	2.11	1.22
MN Boolean	0.53	0.68	0.10	0.48	1.36	2.17

Table 3: Maximum difference ($\times 100$) in ham recall across 500, 1000, 3000 attributes for $T = 0.5$.

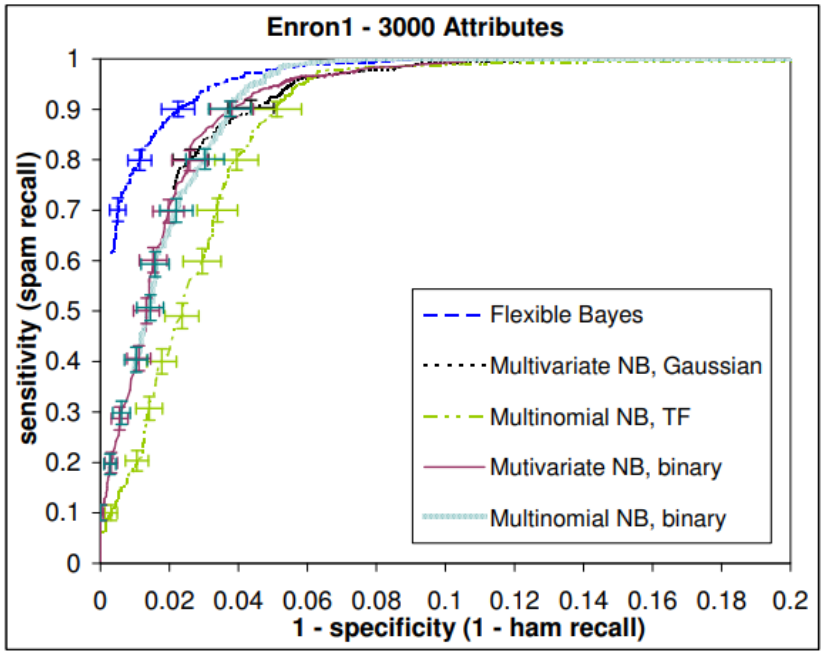
Table2, 3 shows the differences are small in all NB versions for $T = 0.5$, and similar results were obtained for all thresholds.

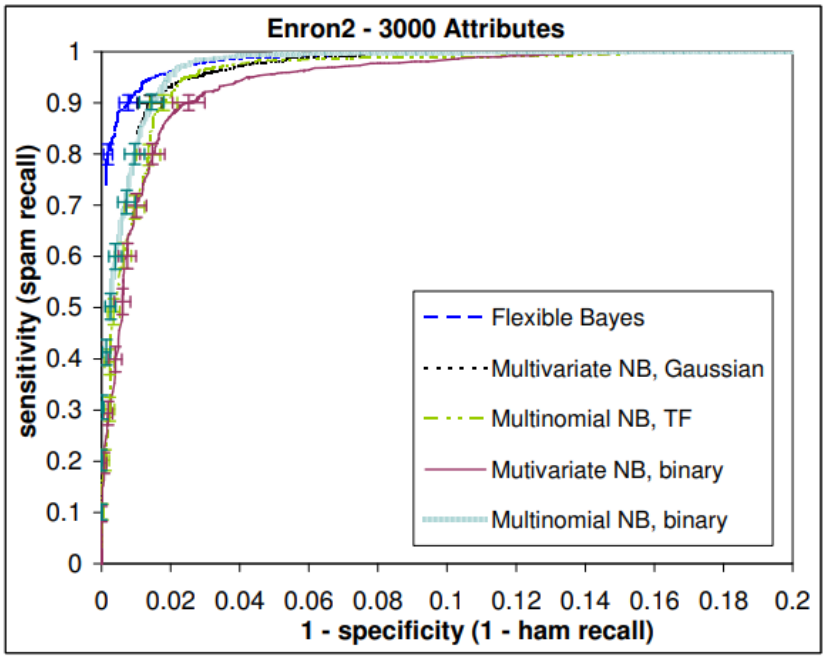
4.2.1 Experimental Results -Comparisons of NB-



multinomial NB with Boolean attributes perform best in 4 out of 6 datasets

4.2.2 Experimental Results -Comparisons of NB2-





- FB classifier performs with superiority in Enron 1, 2 datasets.
- But the ROC curves are shorter.
  This is because $p(c|X)$ are very close to 0, and FB cannot obtain smaller (1 - ham recall)

4.2.3 Experimental Results -Comparisons of NB3-

NB version	Enr1	Enr2	Enr3	Enr4	Enr5	Enr6	Avg.
FB	90.50	93.63	96.94	95.78	99.56	99.55	95.99
MV Gauss	93.08	95.80	97.55	80.14	95.42	91.95	92.32
MN TF	95.66	96.81	95.04	97.79	99.42	98.08	97.13
MV Bern.	97.08	91.05	97.42	97.70	97.95	97.92	96.52
MN Bool.	96.00	96.68	96.94	97.79	99.69	98.10	**97.53**

Table 4: Spam recall (%) for 3000 attributes, $T = 0.5$.

NB version	Enr1	Enr2	Enr3	Enr4	Enr5	Enr6	Avg.
FB	97.64	98.83	95.36	96.61	90.76	89.97	94.86
MV Gauss	94.83	96.97	88.81	99.39	97.28	95.87	95.53
MN TF	94.00	96.78	98.83	98.30	95.65	95.12	96.45
MV Bern.	93.19	97.22	75.41	95.86	90.08	82.52	89.05
MN Bool.	95.25	97.83	98.88	99.05	95.65	96.88	**97.26**

Table 5: Ham recall (%) for 3000 attributes, $T = 0.5$.

- Average scores both ham and spam recall become best with multinomial NB with Boolean attributes.
- Considering computional complexity, smoother ham-spam recall trade-off and run time, MN Bool is the best classifier after all.

4.3 Experimental Results -Learning Curves-

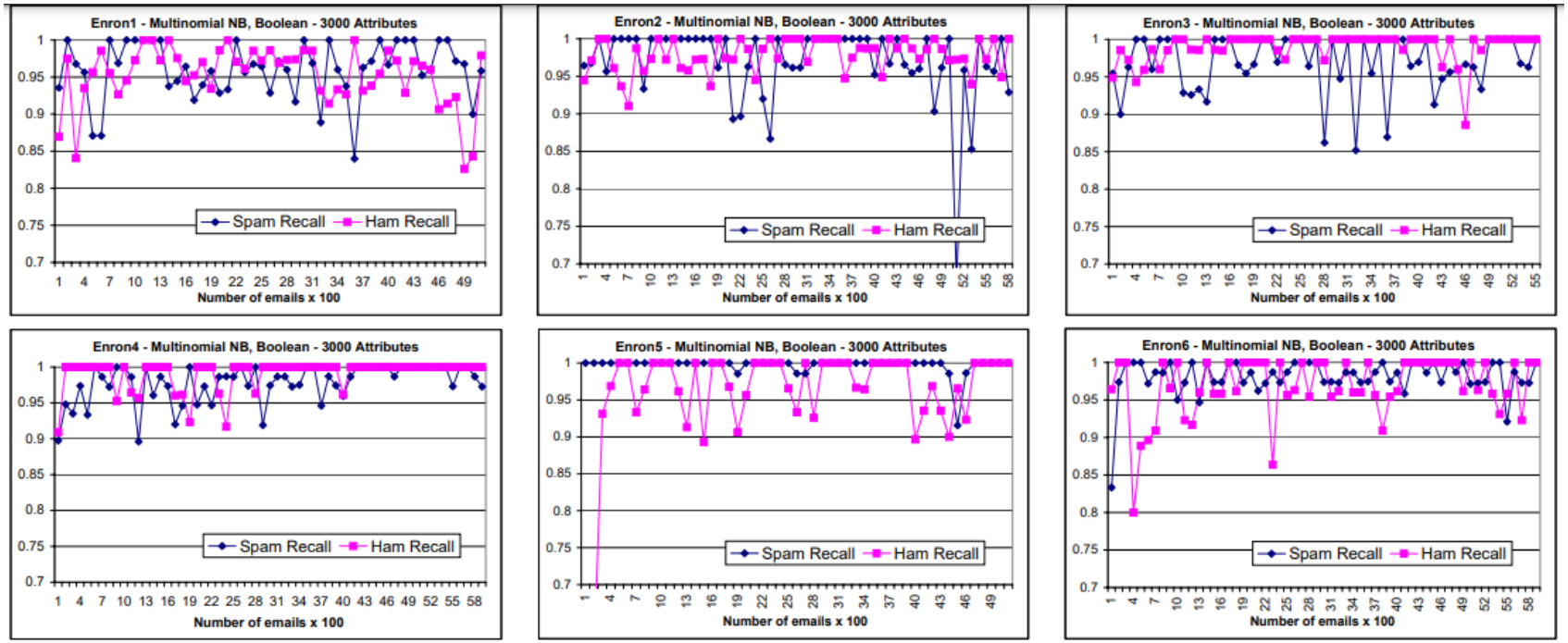


Figure 3: Learning curves for the multinomial NB with Boolean attributes and $T = 0.5$.

- The curves do not increase monotonically because of the unpredictable fluctuation of ham-spam ratio over time.
- The easier dataset (Enron4) reaches almost performances.
- The more difficult dataset (Enron1) shows the continuous fluctuation of ham, spam recall.

5 Conclusions and Further Work

- The best number of token attributes (m) is 3000, the maximum and the gain is insignificant
- FB and MN bool classifiers have not been used in spam filtering, but they perform best.
- Considering run time and smoother ham-spam recall trade-off, MN bool is preferred.

- To decide which NB classifier is better, further experiment is needed.
- Other learning algorithms on real mailboxes should be trained.