

# Classifying Two-Class Chinese Texts in Two Steps

**Xinghua Fan**

*Chongqing University of Posts and Telecommunications, China*

## INTRODUCTION

Text categorization (TC) is a task of assigning one or multiple predefined category labels to natural language texts. To deal with this sophisticated task, a variety of statistical classification methods and machine learning techniques have been exploited intensively (Sebastiani, 2002), including the Naïve Bayesian (NB) classifier (Lewis, 1998), the Vector Space Model (VSM)-based classifier (Salton, 1989), the example-based classifier (Mitchell, 1996), and the Support Vector Machine (Yang & Liu, 1999).

Text filtering is a basic type of text categorization (two-class TC). There are many real-life applications (Fan, 2004), a typical one of which is the ill information filtering, such as erotic information and garbage information filtering on the web, in e-mails and in short messages of mobile phones. It is obvious that this sort of information should be carefully controlled. On the other hand, the filtering performance using the existing methodologies is still not satisfactory in general. The reason lies in that there exist a number of documents with high degree of ambiguity, from the TC point of view, in a document collection, that is, there is a fuzzy area across the border of two classes (for the sake of expression, we call the class consisting of the ill information-related texts, or, the negative samples, the category of TARGET, and, the class consisting of the ill information-not-related texts, or, the positive samples, the category of Non-TARGET). Some documents in one category may have great similarities with some other documents in the other category, for example, a lot of words concerning love story and sex are likely appear in both negative samples and positive samples if the filtering target is erotic information.

## BACKGROUND

Fan et al observed a valuable phenomenon, that is, most of the classification errors result from the documents of falling into the fuzzy area between two categories, and presented a two-step TC method based on Naive Bayesian classifier (Fan, 2004; Fan, Sun, Choi & Zhang, 2005; Fan & Sun, 2006), in which the idea is inspired by the fuzzy area between categories. In the first step, the words with parts of speech verb, noun, adjective and adverb are regarded as candidate feature, a Naive Bayesian classifier is used to classify texts and fix the fuzzy area between categories. In the second step, bi-gram of words with parts of speech verb and noun as feature, a Naive Bayesian classifier same as that in the previous step is used to classify documents in the fuzzy area.

The two-step TC method described above has a shortcoming: its classification efficiency is not well. The reason lies in that it needs word segmentation to extract the features, and at currently, the speed of segmenting Chinese words is not high. To overcome the shortcoming, Fan et al presented an improved TC method that uses the bi-gram of character as feature at the first step in the two-step framework (Fan, Wan & Wang, 2006).

Fan presented a high performance prototype system for Chinese text categorization including a general two-step TC framework, in which the two-step TC method described above is regarded as an instance of the general framework, and then presents the experiments that are used to validate the assumption as the foundation of two-step TC method (Fan, 2006). Chen et al. has extended the two-step TC method to multi-class multi-label English (Chen et al., 2007).

**MAIN FOCUS**

**Fix the Fuzzy Area between Categories Using a Naïve Bayesian Classifier**

(Fan, 2004; Fan, Sun, Choi & Zhang, 2005; Fan & Sun 2006)

A Naïve Bayesian Classifier is used to fix the fuzzy area in the first step. For a document represented by a binary-valued vector  $d=(W_1, W_2, \dots, W_{|D|})$ , the two-class Naïve Bayesian Classifier is given as follows:

$$\begin{aligned}
 f(d) &= \log \frac{\Pr\{c_1 | d\}}{\Pr\{c_2 | d\}} \\
 &= \log \frac{\Pr\{c_1\}}{\Pr\{c_2\}} + \sum_{k=1}^{|D|} \log \frac{1-p_{k1}}{1-p_{k2}} + \sum_{k=1}^{|D|} W_k \log \frac{p_{k1}}{1-p_{k1}} - \\
 &\quad \sum_{k=1}^{|D|} W_k \log \frac{p_{k2}}{1-p_{k2}}
 \end{aligned} \tag{1}$$

Figure 1. Distance from point (x,y) to the separate line

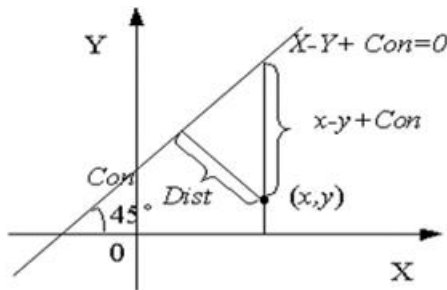
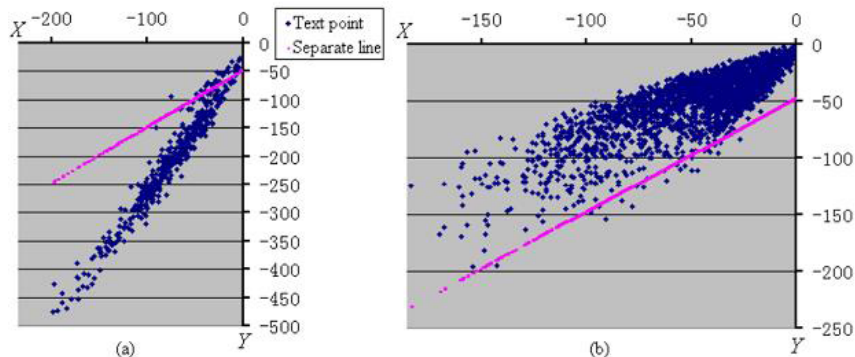


Figure 2. Distribution of the training set in the two-dimensional space



where  $\Pr\{\cdot\}$  is the probability that event  $\{\cdot\}$  occurs,  $c_i$  is category  $i$ , and  $p_{ki}=\Pr\{W_k=1|c_i\}$  ( $i=1,2$ ). If  $f(d) \geq 0$ , the document  $d$  will be assigned the category label  $c_1$ , otherwise,  $c_2$ .

Let:

$$Con = \log \frac{\Pr\{c_1\}}{\Pr\{c_2\}} + \sum_{k=1}^{|D|} \log \frac{1-p_{k1}}{1-p_{k2}} \tag{2}$$

$$X = \sum_{k=1}^{|D|} W_k \log \frac{p_{k1}}{1-p_{k1}} \tag{3}$$

$$Y = \sum_{k=1}^{|D|} W_k \log \frac{p_{k2}}{1-p_{k2}} \tag{4}$$

Where  $Con$  is a constant relevant only to the training set,  $X$  and  $Y$  are the measures that the document  $d$  belongs to categories  $c_1$  and  $c_2$  respectively. (1) is rewritten as:

$$f(d) + X - Y + Con \tag{5}$$

Apparently,  $f(d)=0$  is the separate line in a two-dimensional space with  $X$  and  $Y$  being X-coordinate and Y-coordinate respectively. In this space, a given document  $d$  can be viewed as a point  $(x, y)$ , in which the values of  $x$  and  $y$  are calculated according to (3) and (4). As shown in Figure1, the distance from the point  $(x, y)$  to the separate line will be:

$$Dist = \frac{1}{\sqrt{2}}(x - y + Con) \tag{6}$$

4 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/chapter/classifying-two-class-chinese-texts/10822](http://www.igi-global.com/chapter/classifying-two-class-chinese-texts/10822)

## Related Content

---

### Mining Email Data

Steffen Bickel (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1262-1267).  
[www.irma-international.org/chapter/mining-email-data/10984](http://www.irma-international.org/chapter/mining-email-data/10984)

### Data Pattern Tutor for AprioriAll and PrefixSpan

Mohammed Alshalalfa (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 531-537).  
[www.irma-international.org/chapter/data-pattern-tutor-aprioriall-prefixspan/10871](http://www.irma-international.org/chapter/data-pattern-tutor-aprioriall-prefixspan/10871)

### Statistical Data Editing

Claudio Conversano and Roberta Siciliano (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1835-1840).  
[www.irma-international.org/chapter/statistical-data-editing/11068](http://www.irma-international.org/chapter/statistical-data-editing/11068)

### Legal and Technical Issues of Privacy Preservation in Data Mining

Kirsten Wahlstrom, John F. Roddick, Rick Sarre, Vladimir Estivill-Castro and Denise de Vries (2009).  
*Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1158-1163).  
[www.irma-international.org/chapter/legal-technical-issues-privacy-preservation/10968](http://www.irma-international.org/chapter/legal-technical-issues-privacy-preservation/10968)

### Segmenting the Mature Travel Market with Data Mining Tools

Yawei Wang (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1759-1764).  
[www.irma-international.org/chapter/segmenting-mature-travel-market-data/11056](http://www.irma-international.org/chapter/segmenting-mature-travel-market-data/11056)