Classification of Consumer Goods Safety Cases Based on Improved Bayesian Model

Yingcheng Xu1, Wei Feng1, Fei Pei1,2,* and Haiyan Wang3

1 China National Institute of Standardization, Beijing 100191, China
2 China National Institute of Standardization, Beijing 100191, China
3 School of Management and E-Business, Zhejiang Gongshang University, Hangzhou 310018, China

*Corresponding author: springblue410@126.com

Abstract. Given the shortcomings of traditional multinomial naive Bayesian model, an improved multinomial Bayesian model based on feature weighting is first proposed. Second, in the weight setting of the model's feature selection items, the TF-IDF algorithm does not consider the distribution of feature items between and within classes, and a range of factors related to word distribution is proposed such as intra-class factors, inter-class factors, document set factors, and normalization factors, then an improved TF-IDF weighting algorithm is put forward. Finally, combined with the case of consumer product safety, the effectiveness and feasibility of the proposed model and algorithm are verified.

Keywords: Bayes, Consumer Goods, Classification

1 Introduction
Due to the frequent occurrence of consumer goods safety incidents, the governments around the world have focused more attention on the safety of consumer goods. As an important information carrier of consumer goods safety incidents, the text information of cases contains a range of incidents-related factors, such as time of occurrence, scene of action, source of hazards and severity of injury. A large number of words and higher dimensions not only increase the burden on classification learning algorithm, but also reduce the classification effect of classifier. The improved Bayesian model can effectively solve the classification problem of consumer goods safety cases, especially in calculating the weight of feature items. The improved TF-IDF algorithm for feature weight helps build a vector space model of text by considering the distribution of feature items between and within classes. At last, the proposed model and algorithm are used for classifying the text of toy-related safety cases, so as to verify their validity and feasibility.1-7.

2 Improved Multinomial Naive Bayesian Model Based on Feature Weighting

2.1 Multinomial Naive Bayesian Model
In the multinomial naive Bayesian model, a document is deemed as a collection of well-ordered words,
and many characteristic words in different sequences correspond to the same document. Suppose a
document belongs to class C, and the probability that the characteristic word $t_j$ appears once is
$P(t_j \mid C_i)$. If the characteristic word $t_j$ appears $x_j$ times in the document, its probability of
occurrence shall be $P(t_j \mid C_i)^{x_j}$. The probability of occurrence for a set of words ordered in the
sequence shall be $\prod_j P(t_j \mid C_i)^{x_j}$. Its multinomial formula shall be written as follows:

$$P(d \mid C_i) = \binom{n}{x_1, x_2, \ldots} \prod_j P(t_j \mid C_i)^{x_j} = n! \prod_j \frac{P(t_j \mid C_i)^{x_j}}{x_j}$$  (1)

After optimization, Equation (2) is obtained:

$$P(C_i \mid d) \propto P(C_i)n! \prod_j \frac{P(t_j \mid C_i)^{x_j}}{x_j}$$  (2)

Where $x_j$ is the times that the characteristic word $t_j$ appears in the document, and $n = \sum_j x_j$.

$$P(t_j \mid C_i) = \frac{\text{Count}(t_j, C_i)}{\sum_j \text{Count}(t_j, C_i)}$$  (3)

To avoid the occurrence of zero value, Laplacian smoothing is used to process Equation (3):

$$P(t_j \mid C_i) = \frac{1 + \text{Count}(t_j, C_i)}{n + \sum_j \text{Count}(t_j, C_i)}$$  (4)

Where $\text{Count}(t_j, C_i)$ represents the times that the characteristic word $t_j$ appears in the
document belonging to the $C_i$ class, $\sum_j \text{Count}(t_j, C_i)$ represents the total number of all feature
words appearing in the $C_i$ class document, and $n$ is the total number of features in the feature set.

### 2.2 Improved Multinomial Naive Bayesian Model Based on Feature Weighting

Add a weight to each characteristic word. Given that different characteristic words have different
contributions to the text, an improved naive Bayesian classification model can be obtained after weighting:

$$P(C_i \mid d) \propto P(C_i)n! \prod_j \frac{P(t_j \mid C_i)^{x_j}}{x_j}$$  (5)

For the convenience of calculation, the logarithm on both sides of the equation above is taken to
obtain the final improved naive Bayesian classification model:

$$P(C_i \mid d) = \max \{\ln P(C_i) \sum_j (w_jx_j \ln P(t_j \mid C_i) - \ln x_j)\}$$  (6)

### 2.3 Calculate Feature Weight

In the improved multinomial naive Bayesian model, the feature items are generally calculated by the
TF-IDF algorithm. However, the TF-IDF algorithm relies too heavily on word frequency and ignores
the imbalance between classes. A valid feature item for classification shall not only reflect the content
of the class it belongs to, but also distinguish its class from other classes. Thus, an improved algorithm
based on the distribution of word class is proposed to improve the shortcomings of traditional weight
algorithm.
The distribution information of feature items, also known as the discrimination of feature frequency distribution, can be divided into two types: inter-class discrimination $DI_{ac}$ and intra-class discrimination $DI_{ac}$. Where,

$$DI_{ac} = \sqrt{\frac{\sum_{i=1}^{m} (tf(t, C_i) - \overline{tf(t, C)}) / (m-1)}{\overline{tf(t, C)}}}$$ \hspace{1cm} (7)$$

In Equation (7), $tf(t, C_i)$ represents the frequency at which the characteristic word $t$ appears in $C_i$, while $m$ is the total number of classes. $tf(t, C)$ is the total frequency at which $t$ appears on the text set, and $\overline{tf(t, C)} = \sum_{i=1}^{m} tf(t, C_i) / m$. $\overline{tf(t, C)}$ is the average frequency at which $t$ appears in each class, and

$$DI_{ac} = \sqrt{\frac{\sum_{j=1}^{n} (tf(t, d_j) - \overline{tf(t, C_i)}) / (n-1)}{\overline{tf(t, C_i)}}}$$ \hspace{1cm} (8)$$

In Equation (8), $tf(t, d_j)$ represents the frequency at which $t$ appears in the $j^{th}$ document, and $n$ is the total number of documents in the class. $tf(t, C_i)$ represents the total frequency at which $t$ appears in the class, and $\overline{tf(t, C_i)} = \sum_{j=1}^{n} tf(t, d_j) / n$. $\overline{tf(t, C_i)}$ represents the average frequency at which $t$ appears in each class, and $\overline{tf(t, C_i)} = \frac{1}{n} \sum_{j=1}^{n} tf(t, d_j)$.

Let $DI(t_i) = DI_{ac}(t_i) \times (1 - DI_{in}(t_i))$ represent the weighting that combines the distribution information of class based on traditional TF-IDF algorithm. The improved TF-IDF formula is as follows:

$$w_i = \frac{tf_i \times \log(N / n_i + L)}{\sqrt{\sum_{j=1}^{n} (tf_j^2 \times \log(N / n_j + L)^2)}} DI(t_i)$$ \hspace{1cm} (9)$$

3 Experimental Results and Analysis

3.1 Experimental Data
To verify the validity of the improved Bayesian model and TF-IDF feature weight algorithm, a total of 1,759 children's toy cases were selected as the object of study. Among them, 1,000 cases were randomly selected as the training set and 759 as the test set. First, label toy cases according to the type of injury, the type of injury is poisoning hazard, the mark is 0, the type of injury is suffocation or suffocation, the mark is 1, the type of chemical hazard is 2, the type of injury is cut The risk of injury is marked 3.

3.2 Results and Discussion
Given that different characteristic words have different contributions to text classification in traditional multinomial naive Bayesian model, an improved multinomial naive Bayesian model is proposed. In terms of the weight setting of feature selection items, since TF-IDF does not take into account the distribution of feature items between and within classes, and a range of factors related to word distribution is proposed such as intra-class factors, inter-class factors, document set factors, and normalization factors, then an improved IF-IDF algorithm is given. At last, the proposed model and

| Table 1. Classification results of toy cases with the risk of poisoning and asphyxia |
|-----------------------------------------------|
| Algorithms | Poisoning | Asphyxia |
|            | Recall ratio r(%) | Precision ratio p(%) | F1(%) | Recall ratio r(%) | Precision ratio p(%) | F1(%) |
| LDA       | 73.62      | 75.71      | 74.65 | 74.55 | 76.48      | 75.51 |
| QDA       | 79.33      | 79.07      | 79.20 | 80.62 | 79         | 79.81 |
| NBC       | 83.51      | 83.35      | 83.43 | 85.17 | 84.98      | 85.07 |
| KNN       | 85.07      | 87.11      | 86.08 | 85.89 | 87.26      | 86.57 |
| ANN       | 88.43      | 87.6       | 88.01 | 88.43 | 89.97      | 89.20 |
| Improved NBC | 88.31  | 88.73      | 88.52 | 90.31 | 90.71      | 90.51 |

The improved Bayesian model was compared with multiple methods in the experiment, such as Fisher’s Linear Discriminant Analysis (LDA), Fisher’s Quadratic Discriminant Analysis (QDA), Naive Bayes Classification (NBC), K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN).

These classification models and the improved naive Bayesian classification model were trained by the training set. The trained models were then used to test the test set and calculate the index of classification results. The recall ratio (r), precision ratio (p) and F1 were used to compare the performance of all classification models. The experimental results are shown in Table 1 and Table 2.

| Table 2. Classification results of toy cases with chemical hazards and cutting risk |
|-----------------------------------------------|
| Algorithms | Chemicals | Cutting |
|            | Recall ratio r(%) | Precision ratio p(%) | F1(%) | Recall ratio r(%) | Precision ratio p(%) | F1(%) |
| LDA       | 73.06      | 71.22      | 72.13 | 69.91 | 71.06      | 70.48 |
| QDA       | 80.35      | 81.26      | 80.80 | 73.63 | 77.62      | 75.57 |
| NBC       | 83.01      | 82.18      | 82.59 | 83.94 | 85.08      | 84.51 |
| KNN       | 84.93      | 83.99      | 84.46 | 85.52 | 84.93      | 85.22 |
| ANN       | 85.86      | 88.69      | 87.25 | 86.64 | 85.36      | 86.00 |
| Improved NBC | 87.97  | 88.37 | 88.17 | 91.55 | 89.52      | 90.52 |

It can be seen from the table that regarding the toy cases with a poisoning risk, the improved naive Bayesian model has a recall ratio (r) higher than LDA, QDA, NBC and KNN, and only second to ANN, but its precision ratio (p) and F1 are both higher than other models; regarding the toy cases with the risk of asphyxia and cutting, the improved naive Bayesian model has higher recall ratio (r), precision ratio (p) and F1 than all other models; regarding the toy cases with chemical hazards, the improved naive Bayesian model is only slightly inferior to ANN in terms of precision ratio (p), but superior to all other models in terms of recall ratio (r) and F1.

The empirical results show that the naive Bayesian model is effectively improved. Compared with the methods including LDA, QDA, NBC, KNN and ANN, the improved naive Bayesian model has advantages in classifying toy-related cases.

Conclusions

Given that different characteristic words have different contributions to text classification in traditional multinomial naive Bayesian model, an improved multinomial naive Bayesian model is proposed. In terms of the weight setting of feature selection items, since TF-IDF does not take into account the distribution of feature items between and within classes, and a range of factors related to word distribution is proposed such as intra-class factors, inter-class factors, document set factors, and normalization factors, then an improved IF-IDF algorithm is given. At last, the proposed model and
algorithm are applied to the text classification of toy-related safety cases. Experiments show that the proposed methods play an effective role in classifying the text of consumer goods safety cases.

**Acknowledgements**

We would like to acknowledge that this research is supported and funded by the Market Supervision Technology Assurance Project "Research on Evolution and Response Strategy of Product Quality and Safety Emergencies Based on Domain Ontology", the the Basic Scientific Research Business Projects 552019Y-6770-1, and National Science Foundation of China under Grant No. 91646122, 91746202.

**References**

[1] Harry Z, Sheng S, Learning Weighted Naive Bayes with Accurate Banking. In: Fourth IIE International Conference on Date Mining (ICDM'04).2004:567-570.

[2] Jorge Enrique Rodríguez, Jenny P. Ortiz. Bayesian methods for classification inappropriate web pages. Vision electric, 2017, 1:10-10.

[3] Lindsey Harrison, Emilie O'Connor, Chunfa Jie et al. Potentially inappropriate medication prescribing in the elderly: Is the Beers Criteria relevant in the Emergency Department today? American Journal of Emergency Medicine, 2019, 37(9):45-58.

[4] Oliver Mitchell, Samantha Cottam, Sanjay Sharma Analyzing the reasons behind inappropriate referrals on the ‘two week wait’ head and neck cancer pathway British Journal of Oral & Maxillofacial Surgery, 2019, 57(10):124-132.

[5] Wang Z H, Webb G I, Zheng F. Adjusting Dependence Relations for Semi-Lazy TAN Classifier. In: Advances in Artificial Intelligence, LNAI 2903.Berlin Heidelberg:Spring-Verlag,2003:453-456.

[6] Wang Z H, Webb G I.A Heuristic Lazy Bayesian Rule Algorithm.In: Sendoff S William G J,Hegland M. Proceedings of Australian Data Mining WorkshopSydne , Australia: Sydney University of Technology Press,2002:57-63.

[7] Wenbiao Hu,Rebecca A O’Leary,Kerrie Mengersen,Samantha Low Choy. Bayesian classification and regression trees for predicting incidence of cryptosporidiosis.[J]. PLoS ONE,2017,6(8):22-32.