UDC 004.93

A COMPARISON OF NAIVE BAYES MODELS FOR TEXT CLASSIFICATION

Babomuradov O.J., Boboev L.B., Otaxonova B.I.
bobomurodov@edu.uz; repytolochin@gmail.com
Department of Software Engineering, Tashkent University of Information Technology;

The article deals with the task of classifying text documents in the Uzbek language, based on methods and models of data mining. For preliminary data processing, the word portfolio method is used, on the basis of which the characteristic space is formed in the form of an alphabet of words from text. For the classification of text documents, naive Bayes approaches are used — the Bernoulli model and the multi-nominal model. Text documents used in the article are formed from state official information sources of the National Information Agency of Uzbekistan. When comparing probabilistic classification methods, 600 documents were used, which consist of 169,205 words belonging to 6 classes. The result of a comparative analysis of experimental studies showed that with large dimensions of text documents, it is effective to use multi-nominal classification models, and to use the Bernoulli model on small text volumes.

Keywords: text, classification, probability model, Bayes, Bernoulli, multi-nominal.

Citation: Babomuradov O.J., Boboev L.B., Otaxonova B.I. 2019. A comparison of naive bayes models for text classification. *Problems of Computational and Applied Mathematics.* 1(19): 39–43.

1 Introduction

Nowadays intellectual information technologies are widely being utilized in many branches of our everyday life. One of these is intellectual analysis technology of textual information. With the text based on this technology, prediction and basic notions, and also the relations between them, will be studied and determined.

The point of the intellectual analysis of information is classifying it, including the text classification and search for information.

Text classification is done manually, with the help of expert instructions and machine learning methods [1-3]. Automatic classification of texts is mostly based on the concept of "similarity." Normally, such texts store similar words and word phrases in them.

One of the widespread methods of pre-processing of texts is Bag of Words [4]. In this model, firstly we create vocabulary V out of the words from the preset of texts. A histogram vector is created based on the number of repetitions of the words in the texts that match the vocabulary. Some methods look to shorten the vocabulary [5], and some improve the histogram by using the weight scheme. For ex-ample: TF-IDF (term frequency – inverse document frequency) method [4,6].

In some cases of text classification, based on intellectual information technologies, naive Bayes classifier could be helpful, but it will be problematic when we try to classify a natural language automatically. In order to solve these issues, parameters are normalized.

2 Proposed approach

Assume we have a V set of words of a language. Usually V set is called vocabulary. The validity of the V $N(N = |V|)$ is equal to the number of words in it. Based on the V

set, a vector of $S = (S_1, S_2, ..., S_N)$ words is formed. The $K = \bigcup_{i=1}^{m} K_i$ set of texts, say, is categories.

Say, we have a D_j text of K_i category:$(i = \overline{1,m}; j = \overline{1,p})$. The probability of the D_j text lying in the K_i set, according to the Bayes theorem $P(K_i|D_j)$, is equal to:

$$P(K_i|D_j) = \frac{P(D_j|K_i)P(K_i)}{P(D_j)} \Rightarrow P(D_j|K_i)P(K_i) \quad (1)$$

With the given D_j text, G_j set of words is formed and $W_j = (w_{j1}, w_{j2}, ..., w_{jr})$ vector of words is created, that matches the G_j set.
Based on the vector of S words, ${X_j}^i$ Boolean vector, with N dimension, is formed:

$$x_{jt} = \begin{cases} 1, \quad if \quad s_t = w_{je} & j = \overline{1,p}; \\ 0, \quad\quad otherwise & t = \overline{1,N}; e = \overline{1,r}; \end{cases}$$

If the probability of the s_t word is in the K_i is $P(s_t|K_i)$, then the probability of s_t is not in K_i equals to x $(1 - P(s_t|K_i))$. Then ac-cording to (1), the probability of D_j text belongs to K_i will be determined thus:

$$[P(D_j|K_i) = P(S|K_i) = \prod_{t=1}^{|V|} [x_t P(s_t|K_i) + (1 - x_t)(1 - P(s_t|K_i))]] \quad (2)$$

Say the number of documents (that have s_t words from K_i) is $\eta_{K_i}(s_t)$ and the number of documents that belong to that category is N_{K_i}; Then the probability of s_t word is equal to:

$$\hat{P}(s_t|K_i) = \frac{\eta_{K_i}(s_t)}{N_{K_i}} \quad (3)$$

If the total number of the learning documents is N, then the proba-bility of documents belonging to K_i is:

$$\hat{P}(K_i) = \frac{N_{K_i}}{N} \quad (4)$$

The Bernoulli model of classification of the set of learning docu-ments and the texts of K_i category is carried out through the following steps:

1. Vocabulary is created.
2. Learning.
3. Classification To determine a category of a non-classified document D, the combinations (1) and (2) are used:

$$P(K_i|S) \Rightarrow P(S|K_i)P(K_i) \Rightarrow P(K_i) \prod_{t=1}^{|V|} [x_t P(s_t|K_i) + (1 - x_t)(1 - P(s_t|K_i))] \quad (5)$$

In order to classify texts of greater magnitude, usually multi-nominal model is used, which is more effective than the Bernoulli model. Below is a detailed explanation of it.

In the multi-nominal model, a vector of signs is created based on the repetition of a word in a vocabulary-based text.

Here, n_i is the amount of repetition i word from the given vocabulary. Multi-nominal division of words based on the multi-nominal coeffi-cient is calculated with the following formula:

$$M_k = \frac{N!}{n_1! n_2! ... n_2!}$$

Here, n_i is the amount of repetition i word from the given vocabulary.
Multi-nominal division of words based on the multi-nominal coeffi-cient is calculated with the following formula:

$$P(N) = \frac{N!}{n_1!n_2!\dots n_N!} p_1^{n_1} p_2^{n_2} \dots p_N^{n_N} = \frac{N!}{\prod_{t=1}^{N} n_t!} \prod_{t=1}^{N} p_t^{n_t} \tag{6}$$

Here, the probability of words' sequence is divided by the $\prod_{t=1}^{N} p_t^{n_t}$ multiplication, and classify the target.

Say, is the frequency of s_t word in a D_j document. In that case, the probability of s_t is in the K_i equals to: $P(s_t|K_i)$. Then, the probability of D_j text belongs to K_i, i.e. the probability of S words belong to K_i is:

$$P(D_j|K_i) = P(S|K_i) = \frac{N!}{\prod_{t=1}^{|V|} n_t!} \prod_{t=1}^{|V|} P(s_t|K_i)^{n_t} \Rightarrow \prod_{t=1}^{|V|} P(s_t|K_i)^{n_t} \tag{7}$$

Due to the fact that the normalization doesn't concern whether the s_t word is the property of any class, it is not necessary to conduct a normalization.
In the multi-nominal model, the probability of the $P(s_t|K_i)$ category document and $P(K_i)$ category will develop parameters for the model. Whether D_j document belongs to K_i category, is created by evaluating the parameters of a set of learning docu-ments, and valued with 1 or 0. When the total number of documents is N, $P(s_t|K_i)$ probability is determined through the below formula:

$$\hat{P}(s_t|K_i) = \frac{\sum_{j=1}^{N} n_{jt} z_{fi}}{\sum_{f=1}^{|V|} \sum_{j=1}^{N} n_{jf} z_{ji}} = \frac{n_i(s_t)}{\sum_{f=1}^{|V|} n_i(s_f)} \tag{8}$$

$\{Y_1, Y_2, ..., Y_l\}$ is formed based on the set of learning documents, that is, if Y_t belongs to K_i category, z_{ti} variable is 1, otherwise it is 0.

Say, Y set of learning documents and K set of categories are given, the algorithm of text classification based on multi-nominal model would be as follows:

1. V vocabulary is developed;
2. The followings will be calculated:

– N – total number of documents

– N_k – the number of documents, that belong to category k, is determined $k = \overline{1, K}$

– n_{it} the frequency of the word s_t in D_i document, for each word in V, is calculated; simultaneously, the $n_i(s_t)$ frequency of s_t words in K_i category documents is determined;

3. Using (8), $P(s_t|K_i)$ probability is cal-culated.
4. Using (4), $P(K_i)$ probability is calculated.
5. Whether a text belongs to K_i category is found out thus.

When classifying the D_j document, the category proba-bility is calculated through the combinations of (1) and (7):

$$P(K_i|D_j) = P(K_i|S) \Rightarrow P(S|K_i)P(K_i) \Rightarrow P(K_i) \prod_{t=1}^{|V|} P(s_t|K_i)^{n_t} \tag{9}$$

Unlike the Bernoulli model, in the multi-nominal model, words that don't exist ($s_t =$ $= 0$) in a document don't affect the proba-bility ($p^0 = 1$).

If the words in a document are symbolized as u, the probability is calculated as follows:

$$P(K_i|D_j) \Rightarrow P(K_i) \prod_{t=1}^{len(D)} P(u_t|K_i) \tag{10}$$

Here, u_t is the t-nth word in document D_j. In experimental procedure, the change in time of transformation was observed. TfidfVectorizer and HashingVectorizer transformation approaches were used to verify the reliability of results, as shown in Figure 1.

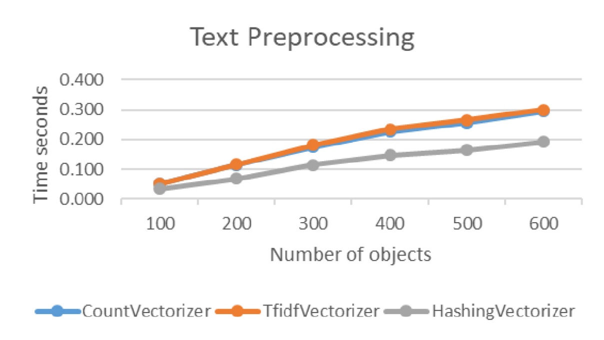

Figure 1 Time-consuming comparison of different types of transfor-mation

Here are results of algorithm based on Naive Bayes probability with CountVectorizer. Classification accuracy being increased from 78% to 88%.

When applying TfidfVectorizer approach, speed and accuracy were low: 78-88% accuracy was obtained.
The following table shows the result of a comparison models' accuracy and time consuming (Table 1).

Model	Precision	Time
BernoulliNB	0.65	0.017
Multinomi-alNB	0.86	0.009
LinearSVC	0.82	0.89
Perceptron	0.86	0.019

To assess the effectiveness of the classification models such as the Bernoulli and multinominal, 600 documents, with 169205 words of 6 categories in it, have been used and with the set of documents, a 28343-word vocabulary has been created. When testing the classi-fication of the selected texts with the Bernoulli model, average accu-racy was 65% and it took 17.28 milliseconds. As for the multi-nominal model, the accuracy came to about 86% and it took 9.79 milliseconds. Experimental research works have proven the multi-nominal model more accurate and faster than the Bernoulli.

3 Conclusion

Pre-processing of texts, with Bernoulli and multi-nominal methods, has been looked through. The space for symbols, which is the most important for the classification of texts in Uzbek language, and mathematical way of classification have been developed. In order to make them recognizable, texts of various themes were formed and classified into categories. The results show the effectiveness of the multi-nominal model, when classifying the texts of bigger size.

References

[1] S.B. Kim, K.S. Han, H.C. Rim, S.H. Myaeng. December 2006. "Some Effective Techniques for Naive Bayes Text Classification", IEEE Transactions on Knowledge and Data En-gineering.

[2] A. A. Alekseev, A. S. Katasev, A. E. Kirillov, A. P.. 2016. "Kirpichnikov. Classification of text documents based on text mining technology "Informatics, Computational Technique and Control, Bulletin of the Technological University.

[3] Polyakov I. V., Sokolova T. V., Chepovsky A. A., Chepovsky A. M. 2015. "The problem of text classification and differentiating features" Vestn. Novosib. state un. Series: Information Technology.55–63

[4] C.D. Manning, P. Raghavan, H. Schütze. 2008. *"Introduction to Information Retrieval", Cambridge University Press, Cambridge* . 2nd ed. Reading, MA: Addison-Wesley. 463 p.

[5] R.E. Madsen, S. Sigurdsson, L.K. Hansen, J. Larsen. 2004. Pruning the vocabulary for better context recognition", Proceedings of the International Conference on Pattern Recognition, 2.pp. 483-488.

[6] J.H. Paik. 2013. "A novel tf-idf weighting scheme for effective ranking", Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 343-352

Received December 08, 2018

УДК 004.93

СРАВНЕНИЕ НАИВНЫХ БАЙЕСОВСКИХ МОДЕЛЕЙ ДЛЯ КЛАССИФИКАЦИИ ТЕКСТА

Бабомурадов О.Ж., Бобоев Л.Б., Отахонова Б.И.

bobomurodov@edu.uz; repytolochin@gmail.com

Ташкентскый университет информационных технологий им. М. Ал-Хоразмий

В статье рассматривается задача классификации текстовых документов на узбекском языке. Для предварительной обработки данных используется метод портфеля слов, с помощи которого формируется признаковое пространство в виде алфавита слов из текста. Для классификации текстовых документов используются подходы наивного Байеса - модель Бернулли и мультиноминальная модель. Текстовые документы, используемые в статье, сформированы из государственных официальных информационных источников Национального Информационного Агентства Узбекистана. При сравнении вероятностных методов классификации использовано 600 документов, состоящих из 169205 слов и относящихся к 6 классам. Результат сравнительного анализа экспериментальных исследований показали, что при большой размерности текстовых документов эффективно использовать мультиноминальную модель классификации, а модель Бернулли использовать на малых объемах текста

Ключевые слова: текст, классификация, вероятностная модель, Байес, Бернулли, мультиноминал

Цитирование: *Бабомурадов О.Ж., Бобоев Л.Б., Отахонова Б.И.* Сравнение наивных байесовских моделей для классификации текста // Проблемы вычислительной и прикладной математики. — 2019. — № 1(19). — С. 39–43.