International Journal of Advanced Intelligence
Volume 5, Number 1, pp.1-13, July, 2013.

AIA International Advanced Information Institute

New Naive Bayes Methods using Data from All Classes

Kanako Komiya, Yusuke Ito, and Yoshiyuki Kotani
Tokyo University of Agriculture and Technology
2-24-16 Naka-cho, Koganei, Tokyo 184-8588 JAPAN
{kkomiya, kotani}@cc.tuat.ac.jp, ito21347@gmail.com

Received (18, December 2012)
Revised (29, May 2013)

The Naive Bayes (NB) is famous for text classification but its accuracy is sometimes not so high for skewed distribution data. On the other hand, the Negation Naive Bayes (NNB) tackled the non-uniformity of the data distribution but still used only one-sided data. This paper proposes the new two Bayesian approaches for text classification which use both the data of a class in question and its complement class. The Universal-set Naive Bayes (UNB) uses them simultaneously and the Selective Naive Bayes (SNB) selects them depending on the amount of the training data. These two methods are both derivable from the equation of the NB. The experiments show that the UNB is significantly better than the existing models which use only one-sided data, i.e., the NB or the UNB, in the 5-class classification. In addition, the results of the 3-class classification show that the UNB consistently gives us the best or the second best accuracy that is near to the best, comparing with the NB or the NNB.

Keywords: Naive Bayes; text classification; Universal-set Naive Bayes; Selective Naive Bayes.

1. Introduction

Many researchers have investigated text classification and the Naive Bayes (NB) is one of the most famous methods for it. There are some modifications of NB; Rennie et al. [1] proposed the complement naive Bayes (CNB) and Komiya et al. [2] proposed the Negation Naive Bayes (NNB). The NB estimates the parameters of a class using the data belong to it whereas the CNB and the NNB estimate them using the data from its complement class.

In this work, we combine ideas of the two methods, the multinomial model of the NB and the NNB, and propose two methods, the Universal-set Naive Bayes (UNB) and the Selective Naive Bayes (SNB), based on both a class in question and its complement class. The UNB uses the data from a class and its complement class simultaneously, and the SNB selects them depends on the amount of the training data.

This paper is organized as follows. Section 2 reviews related work on NB and Sections 3 explains the classification methods including the two proposed methods. Section 4 explains the experimental settings of the classification experiments. We

present the results and discuss them in Section 5. Finally, we conclude the paper in Section 6.

2. Related Work

Many works on text classification has been accomplished so far and Bayesian approach is often used in them [3]. Izutsu et al. [4] categorized the HTML documents and compared the NB classifier with discriminant analysis and the rule-based method. They suggested the simple implementation and the high scalability of the NB classifier. McCallum et al.[5] suggested the difference between multinomial model and multi-variate Bernoulli model of the NB classifier in text classification. Lewis [6] compared the difference of the effect between the types of features used for text classification: words, phrases, clustered words, clustered phrases and indexing terms. Church [7] used a concept called "Adaptation" as the weighting method to the words in substitution for IDF value, and defined the words related to contents but not included a document as "Neighbor". The feature terms were extracted depending on them. Wei et al. [8] reported multi-label classification algorithm.

The Naive Bayes (NB) is one of the most famous methods for text classification. However, its accuracy is sometimes not so high for skewed distribution data and that is why there are some modifications of NB; Rennie et al. [1] proposed the complement naive Bayes (CNB) and Komiya et al. [2] proposed the Negation Naive Bayes (NNB). Whereas the NB estimate the parameters of a class using the data belong to itself, the CNB and the NNB estimate them using the data from the complement class of it. Komiya et al. [2] reported that the equation of NNB is derivable from the equation of the NB unlike the CNB but it has the same advantages; it tackles the non-uniformity of the texts of each class, and they got the classification accuracies that exceed the NB and the CNB significantly when the data distribution is non-uniformly.

In this work, we focused on they still use only one-sided data and combine ideas of the two methods, the multinomial model of the NB and the NNB, and propose two methods, the Universal-set Naive Bayes (UNB) and the Selective Naive Bayes (SNB), based on both a class in question and its complement class. The UNB uses the data from a class and its complement class simultaneously, and the SNB selects them depending on the amount of the training data. These two methods are both derivable from the equation of the NB like NNB.

3. Classification Method

3.1. *Naive Bayes Classifier*

We used the NB classifier to classify the texts as a baseline. Let $d = w_1, w_2, \ldots, w_n$ denote the text containing the words and let c denote a class. Here, let $\hat{c}$ denote the class that d belongs to, and $\hat{c}$ is as follows:

$$\hat{c} = \underset{c}{\operatorname{argmax}}\, P(c|d). \tag{1}$$

By substituting theorem of conditional probability into the equation, we obtain the following:

$$\begin{aligned}\hat{c} &= \underset{c}{\operatorname{argmax}} P(c|w_1, w_2, \ldots, w_n) \\ &= \underset{c}{\operatorname{argmax}} P(w_1, w_2, \ldots, w_n|c)P(c) \end{aligned} \tag{2}$$

We assume that w_i is conditionally independent of every other word. This means that under the above independence assumptions, $P(w_1, w_2, \ldots, w_n|c)$ is approximated by the following:

$$P(w_1, w_2, \ldots, w_n|c) \approx \prod_i P(w_i|c) \tag{3}$$

Finally, the class $\hat{c}$ that d belongs to is determined by following:

$$\hat{c} = \underset{c}{\operatorname{argmax}} P(c) \prod_i P(w_i|c) \tag{4}$$

3.2. *Negation Naive Bayes Classifier*

In the text classification, the NB uses "training data belong to the class c" to estimate the parameters of c. The CNB classifier is an modification of the NB but is not derivable from eq. (1). Therefore, Komiya et al. [2] proposed the NNB, which is derivable from eq. (1) but also have the advantage like the CNB: it tackles the non-uniformity of data of each class.

They transform eq. (1) into eq. (5), substituting Bayes' theorem into the equation.

$$\begin{aligned}\hat{c} &= \underset{c}{\operatorname{argmax}}(1 - P(\bar{c}|d)) = \underset{c}{\operatorname{argmin}} P(\bar{c}|d) \\ &= \underset{c}{\operatorname{argmin}} P(\bar{c})P(w_1, w_2, \ldots, w_n|\bar{c}) \end{aligned} \tag{5}$$

As with eq. (4), the class $\hat{c}$ that d belongs to is determined by following:

$$\hat{c} = \underset{c}{\operatorname{argmin}} P(\bar{c}) \prod_i P(w_i|\bar{c}) \tag{6}$$

3.3. *Universal-set Naive Bayes Classifier*

Now we introduce the derivation of UNB.

First, we got eq. (7) when we applied the Bayes' theorem to the sum of $P(c|d)$ and $P(\bar{c}|d)$, and obtained eq. (8) from eq. (7).

$$\frac{P(c)P(d|c)}{P(d)} + \frac{P(\bar{c})P(d|\bar{c})}{P(d)} = 1 \tag{7}$$

$$P(d) = P(c)P(d|c) + P(\bar{c})P(d|\bar{c}) \tag{8}$$

Then eq. (9) follows from eq. (8) and Bayes' theorem as follows:

$$\begin{aligned} P(c|d) &= \frac{P(c)P(d|c)}{P(c)P(d|c) + P(\bar{c})P(d|\bar{c})} \\ &= \frac{1}{1 + \dfrac{1}{\dfrac{P(c)P(d|c)}{P(\bar{c})P(d|\bar{c})}}} \end{aligned} \tag{9}$$

From eq. (9), it turns out that maximizing the left-hand side is equals to maximizing $\frac{P(c)P(d|c)}{P(\bar{c})P(d|\bar{c})}$ of the right-hand side.

Finally, the class $\hat{c}$ that d belongs to is determined by following:

$$\hat{c} = \underset{c}{\operatorname{argmax}} \frac{P(c)\prod_i P(w_i|c)}{P(\bar{c})\prod_i P(w_i|\bar{c})}. \tag{10}$$

The numerator and the denominator of eq. (10) correspond to a class in question and its complement class respectively. Algorithm 1 shows the algorithm of the UNB.

Algorithm 1 UNB

Require: Classify instance i
  $score(i, C_j) \Leftarrow 1$
  for each class C_j **do**
    $score(i, C_j) \Leftarrow score(i, C_j) * P(C_j)$
    $score(i, C_j) \Leftarrow score(i, C_j)/(1 - P(C_j))$
    for each word w in instance i **do**
      $score(i, C_j) \Leftarrow score(i, C_j) * P(w|C_j)$
      $score(i, C_j) \Leftarrow score(i, C_j)/P(w|\bar{C_j})$
    end for
  end for
  $C_{max} = C_0$
  for each class C_j **do**
    if $score(i, C_{max}) < score(i, C_j)$ **then**
      $C_{max} \Leftarrow C_j$
    end if
  end for
  return C_{max}

3.4. *Selective Naive Bayes Classifier*

The SNB selects the probability to be maximized according to the amount of the training data for each class from the following two choices: (1) the probability that a document d belongs to a class c and (2) the probability that d does not belong to

its complement class $\bar{c}$. If the amount of c is greater than $\bar{c}$, (1) is maximized and otherwise, (2) is maximized. Therefore we got eq. (11) .

$$P(c|d) = \begin{cases} \dfrac{P(c)P(d|c)}{P(d)} \\ 1 - \dfrac{P(\bar{c})P(d|\bar{c})}{P(d)} \end{cases} \tag{11}$$

Note that the first and second case of eq. (11) is the original expression of the NB and the NNB respectively. Thus, the SNB selects a model from the NB-like and the NNB-like models for each class. Here, P(c) is set as a threshold value to select the models. If P(c) is greater than 0.5, the amount of the training data of $\frac{P(c)P(d|c)}{P(d)}$ will be greater than that of $\frac{P(\bar{c})P(d|\bar{c})}{P(d)}$ and vice versa. Then we got eq. (12) from eq. (11), eq. (4), and eq. (6).

$$\begin{aligned} \hat{c} &= \underset{c}{\operatorname{argmax}}\, P(c|d) \\ &= \underset{c}{\operatorname{argmax}} \begin{cases} \dfrac{P(c)\prod_i P(w_i|c)}{P(d)} & (P(c) \geq 0.5) \\ 1 - \dfrac{P(\bar{c})\prod_i P(w_i|\bar{c})}{P(d)} & (otherwise) \end{cases} \end{aligned} \tag{12}$$

It is necessary to calculate $P(d)$ for the SNB unlike the NB or the NNB. $P(d)$ has two different derivation of the eq. (13) and (14) as shown below. $P(d)$ will be like the normalization constant of the numerator of eq. (12)

For the first derivation, we consider the sum of each $P(c|d)$ of the first case in eq. (11), and $\sum_c P(c|d) = 1$.

$$\sum_c \frac{P(c)P(d|c)}{P(d)} = 1 \tag{13}$$

For the second derivation, we consider the sum of each $P(c|d)$ of the second case in eq. (11).

$$\begin{aligned} \sum_c 1 - \sum_c \frac{P(\bar{c})P(d|\bar{c})}{P(d)} &= 1 \\ \sum_c \frac{P(\bar{c})P(d|\bar{c})}{P(d)} &= \left(\sum_c 1\right) - 1 \end{aligned} \tag{14}$$

We get $P(d)$ from eq. (13) and (14) as follows:

$$P(d) = \sum_c P(c) \prod_{i=1}^{n} P(w_i|c) \tag{15}$$

$$= \frac{1}{|C| - 1} \sum_c P(\bar{c}) \prod_{i=1}^{n} P(w_i|\bar{c}). \tag{16}$$

where $|C| = (\sum_c 1)$ represent the number of classes.

Finally, the class $\hat{c}$ that d belongs to is determined by following:

$$\hat{c} = \underset{c}{\text{argmax}} \begin{cases} \dfrac{P(c)\prod_i P(w_i|c)}{\sum_c P(c)\prod_i P(w_i|c)} & (P(c) \geq 0.5) \\ 1 - \dfrac{(|C|-1)\,P(\bar{c})\prod_i P(w_i|\bar{c})}{\sum_c P(\bar{c})\prod_i P(w_i|\bar{c})} & (otherwise)\,. \end{cases} \tag{17}$$

Algorithm 2 shows the algorithm of the UNB.

4. Classification Experiments

We carried out the classification experiments and compared our proposed methods to the NB and the NNB. We used the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [9] which has five classes: (1) Q&A site on the WWW, (2) white papers, (3) publications, (4) periodicals, and (5) newspapers. The bag-of-words were used for the features.

Table 1 and Figure 1 show the number of documents and word tokens of each class in BCCWJ. The third class, publications, has more than 9,000 files and the rest have around 1,000 files for each.

Table 1. The Number of Documents and Word Tokens of Each Class in BCCWJ

index	Genre	Number of docments	Number of tokens
1	Q&A site on the WWW	938	61,674
2	White papers	1,500	875,919
3	Publications	9,121	6,583,798
4	Periodicals	1,049	32,691
5	Newspapers	1,237	478,446
	Total	**13,845**	**8,732,528**

We carried out the 5-class classification. Five-fold cross validation was used in the experiments.

In addition, we also carried out the 3-class classification to see the property of the UNB because the 3-class classification is the simplest classification which gives us the different answers according to the approach. For the 3-class classification, we made ${}_5C_3 = 10$ datasets from five classes. Table 2 shows the index of the included classes of genre and the number of the files of each test set. For the 3-class classification, two experiments were carried out according to the type of the features.

Algorithm 2 SNB

Require: Classify instance i
  $Count \Leftarrow 0$
  $NBscore(i, C_j) \Leftarrow 1$
  $NNBscore(i, C_j) \Leftarrow 1$
  $NBSUM \Leftarrow 0$
  $NNBSUM \Leftarrow 0$
  for each class C_j **do**
    $NBscore(i, C_j) \Leftarrow NBscore(i, C_j) * P(C_j)$
    $NNBscore(i, C_j) \Leftarrow NNBscore(i, C_j) * (1 - P(C_j))$
    for each word w in instance i **do**
      $NBscore(i, C_j) \Leftarrow NBscore(i, C_j) * P(w|C_j)$
      $NNBscore(i, C_j) \Leftarrow NNBscore(i, C_j)/P(w|\bar{C}_j)$
    end for
    $NBSUM \Leftarrow NBSUM + NBscore(i, C_j)$
    $NNBSUM \Leftarrow NNBSUM + NNBscore(i, C_j)$
    $Count \Leftarrow Count + 1$
  end for
  $C_{max} = C_0$
  for each class C_j **do**
    if $P(C_j) \geq 0.5$ **then**
      $score(i, C_j) \Leftarrow NBscore(i, C_j)/NBSUM$
    else
      $score(i, C_j) \Leftarrow 1 - ((Count - 1) * NNBscore(i, C_j)/NNBSUM)$
    end if
    if $score(i, C_{max}) < score(i, C_j)$ **then**
      $C_{max} \Leftarrow C_j$
    end if
  end for
  return C_{max}

All the bag-of-words were used in the first experiment like 5-class classification and ten percent of the features were cut off according to TF/IDF [10] in the second experiment.

We will see the difference between the accuracies of the UNB and existing methods by performing the 3-class classification experiments which is simpler than 5-class in these two settings. Again, five-fold cross validation was used in the experiments.

Laplace smoothing [11] and Jeffreys Perks smoothing [12] were tested in the preliminary experiments and Jeffreys Perks smoothing was employed because it was better.

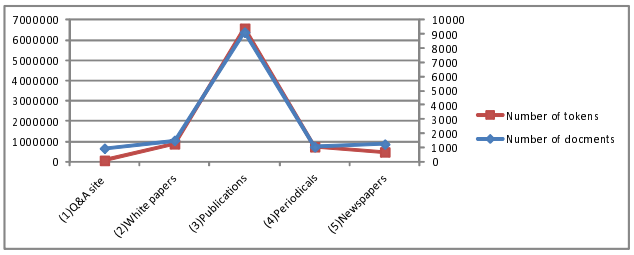

Fig. 1. The Number of Documents and Word Tokens of Each Class in BCCWJ

Table 2. The Index of the Included Classes and the Number of the Files of Each Test Set

The Index of the Set	The Index of the Classes	N of the Files
1	1,2,3	11,559
2	1,2,4	3,487
3	1,2,5	3,675
4	1,3,4	11,108
5	1,3,5	11,296
6	1,4,5	3,224
7	2,3,4	11,670
8	2,3,5	11,858
9	2,4,5	3,786
10	3,4,5	11,407
Total		83,070

5. Evaluation

5.1. *Five-class Classification*

Table 3 lists the performances of each method in the 5-class classification. (They are micro averaged over instances.) The differences of performances were confirmed to be significant according to a chi square test. The level of significance in the test was 0.05.

Table 3. The performances of each method in the 5-class classification

	NB	NNB	UNB	SNB
Precision	0.704	0.557	**0.744**	0.251
Recall	**0.831**	0.399	0.757	0.288
F-measure	**0.749**	0.336	0.738	0.264
Accuracy	78.0%	66.8%	**80.5%**	45.0%

Table 3 shows that the UNB is superior but the SNB gives us the worst results. We think this is because the accuracies of UNB did not decrease even when the NB or the NNB gave us low accuracies; it used the data from a class in question and its complement class simultaneously. However, the SNB selected the equation which gave us lower accuracy in such cases.

The precision, the recall, and the F-measure of each class are shown in Figure 2, Figure 3, and Figure 4.

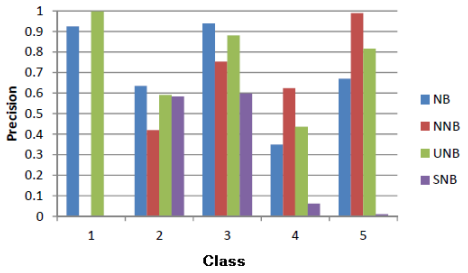

Fig. 2. The Precision of Each Class for 5-class Classification

Figure 2 shows that the value of precision varies greatly according to the methods and the classes. In particular, they are 0 for the NNB and the SNB when the method used only the complement class of the data in class 1. However, the UNB had the best or the second best precision in every class.

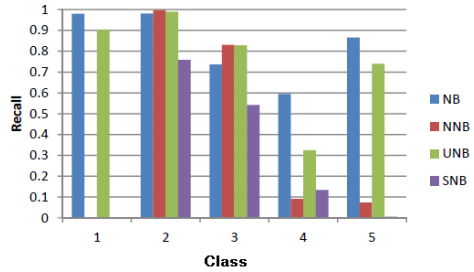

Fig. 3. The Recall of Each Class for 5-class Classification

Figure 3 shows that the recalls of the NNB and the SNB also became 0 like their precision. In addition, these of the UNB are always the second best and the difference from the NNB, which has the best recall, is very small for the third class.

We expected that the performance of the NNB and the SNB will be high when

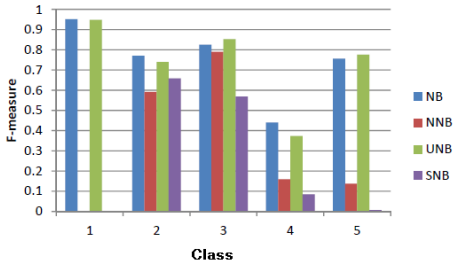

Fig. 4. The F-measure of Each Class for 5-class Classification

their number of the files is small, because their training data will increase; they use the complement set. However, the performance of them tended to be low in such cases.

The accuracies of the NNB were less than half of that of the NB in three classes (Q&A site, periodicals, newspapers) and they decreased the total accuracy.

These results revealed that the ideal threshold to select the equation which gives us higher accuracies should be investigated in the future.

5.2. *Three-class Classification*

Table 4 lists the accuracies of the NB, NNB, and UNB in the 3-class classification. The first row shows the way the accuracy is calculated, i.e., micro averaged or macro averaged, and the second row shows the features that are used for each experiment, i.e., all the features or 90% of the features. Macro in the table means that the micro averaged accuracies over instances are macro averaged over ten sets. This table shows that the NB was the best when all the features were used, and the UNB was the best for the macro averaged accuracy and the NNB was the best for the micro averaged accuracy, when 90% of the features were used. In addition, the UNB is the second best if it is not the best and the difference between the UNB and the best method is always less than one percent, although the differences between the best methods and the other methods are more.

Figures 5 and 6 show the accuracies of each method for each test set when all the features were used and when ten percent of the features were cut off according to the TF/IDF. (Only 90% of the features were used for the experiment.)

These two tables show that the accuracy of the UNB is the best or the second best except the tenth set in the Table 6. (The difference is very small.)

We think that these results support the discussion for the 5-class classification: the UNB is the best or the second best because the accuracies of the UNB did not decrease even when the NB or the NNB gave us low accuracies. These experiments show that the UNB consistently gives us the best or the second best accuracy that

Table 4. The Accuracies of each method in the 3-class classification

	Macro	Micro	Macro	Micro
Features	all	all	90%	90%
NB	**89.46%**	**87.25%**	55.77%	45.08%
NNB	79.94%	80.80%	55.87%	**49.86%**
UNB	88.48%	86.72%	**59.33%**	49.85%

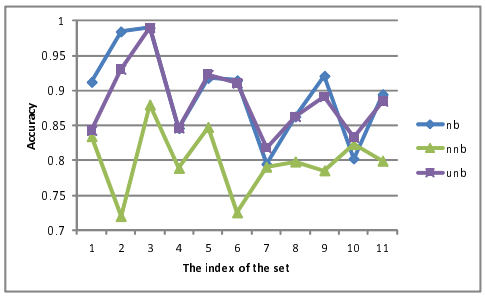

Fig. 5. The Accuracies of Each Set when All the Features Were Used

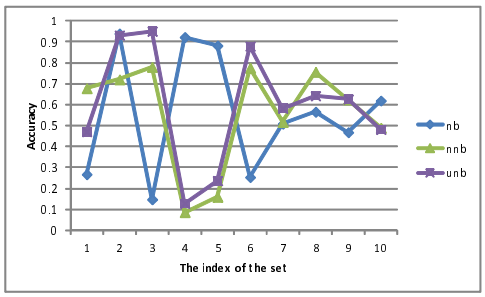

Fig. 6. The Accuracies of Each Set when the Ten Percent of the Features Were Cut Off According to the TF/IDF

is near to the best, comparing with the NB or the NNB.

6. Conclusion

In this paper, we proposed the new two Bayesian approaches for text classification, the UNB and the SNB, which use both the data of a class in question and its complement class, and compared them with the NB and the NNB. The experiments show that the UNB is the best in the four methods but the SNB is the worst in 5-

class classification. We think this is because the accuracies of UNB did not decrease even when the NB or the NNB gave us low accuracies; it used the data from a class in question and its complement class simultaneously. However, the SNB selected the equation which gave us lower accuracy in such cases. The ideal threshold to select the equation which gives us higher accuracies should be investigated in the future. In addition, the results of the 3-class classification show that the UNB consistently gives us the best or the second best accuracy that is near to the best, comparing with the NB or the NNB.

References

1. J.D.M.Rennie, L.Shih, J.Teevan, and D.R.Karger, "Tackling the poor assumptions of naive bayes text classification," in *ICML2003*, 2003, pp. 616–623.
2. K. Komiya, N. Sato, K. Fujimoto, and Y. Kotani, "Negation naive bayes for categorization of product pages on theweb," in *RANLP2011*, 2011, pp. 587–591.
3. D. Mochihashi, "Bayesian approaches in natural language processing," in *IEICE Technical Report. NC, Neurocomputing (In Japanese)*, 2006, pp. 25–30.
4. K. Izutsu, M. Yokozawa, and T. Shinohara, "Comparative evaluation and applications of automatic web-based document classification methods," in *IPSJ SIG Notes 2005(32) (In Japanese)*, 2005, pp. 25–32.
5. A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI/ICML-98 Workshop on Learning for Text Categorization*, 1998, pp. 41–48.
6. D. D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," in *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992, pp. 37–50.
7. K. W.Church, "Empirical estimates of adaptation: The chance of two noriegas is closer to $p/2$ than p^2," in *COLINGf00*, 2000, pp. 173–179.
8. Z. Wei, H. Zhang, Z. Zhang, W. Li, and D. Miao, "A naive bayesian multi-label classifcation algorithm with application to visualize text search results," *International Journal of Advanced Intelligence*, vol. Vol. 3, no. No. 2, pp. 173–188, 2011.
9. K. Maekawa, "Balanced corpus of contemporary written japanese," in *the 6th Workshop on Asian Language Resources (ALR)*, 2008, pp. 101–102.
10. G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. Vol. 18, no. No. 11, pp. 613 – 620, 1975.
11. P. S. marquis de Laplace, *Philosophical Essay On Probabilities.* New York: Springer-Verlag, 1995.
12. G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis.* Reading,MA:Addison-Wesley, 1973.

Kanako Komiya

She received the Ph.D degree in 2009 from Department of Electronic and Information Engineering, Graduate School of Engineering, the Tokyo University of Agriculture and Technology. She is currently an assistant professor at the Tokyo University of Agriculture and Technology. She is interested in Natural Language Processing. She is a member of IPSJ, JSAI, and ANLP.

Yusuke Ito

He was a student at the Department of Computer and Information Sciences, Faculty of Engineering, the Tokyo University of Agriculture and Technology, Japan until March, 2012. His thesis was about the Naive Bayes and text classification.

Yoshiyuki Kotani

He received the Ph.D degree from the University of Tokyo. He is currently a professor at the Tokyo University of Agriculture and Technology. His research interests include Artificial Intelligence, Game Programming, and Natural Language Processing. He is a member of IPSJ and JSAI.