Document Classification Methods with a Small-Size Training Set

Gendoh Kumoi[1], Takashi Ishida[2], Masayuki Goto[3], Shigeichi Hirasawa[1]

[1]Waseda Research Institute for Science and Engineering, Tokyo 169-8555, Japan
(m.kumoi@kurenai.waseda.jp)
[2]Media Network Center, Waseda University, Tokyo 169-8050, Japan,
[3]Faculty of Science and Engineering, Waseda University, Tokyo 169-8555, Japan

ABSTRACT

Document classification methods for a small number of training documents are discussed using mathematical models for information retrieval systems. First, we compare a method by the vector space model with cosine similarity measure and that by a statistical decision model using Bayesian statistics. As a result, it is shown that the former has a smaller probability of the classification error than the latter for a range less than one hundred training data. This suggests that the latter can be improved by a latent model, since it would be expected to be asymptotically the optimum if the model is true. Then we propose a Bayesian decision method using the matrix compression such as by singular value decomposition. It takes a mixture model in the compressed dimension with a prior probability. Although experimental results are not obtained at the present, the proposed method will demonstrate higher or equal performance in the classification error compared to the conventional methods, since it contains them as special cases.

Keywords: document, classification, small number of training data, Bayesian statistics

1. INTRODUCTION

In the research field of information retrieval, there are many mathematical models for retrieval systems such as the Boolean model, the vector space model (VSM), the probabilistic model, and their extended or modified models. Automatic document classification and document clustering systems are also realized by using these models. Hereafter we shall focus upon a document classification method.

For the automatic classification problem, we usually assume that a large number of training (supervised learning) documents can be sufficiently given so that the representative document for each class (e.g., pseudo document vectors calculated by the center of gravity of the training documents in a given class) can be precisely obtained. Then test documents are classified according to the similarity measure between the representative documents and the test document. In this technique, the performance of classification will be improved by extracting terms which contribute the classification method depending on the mutual information between the term and the classes, or on the result of document co-clustering. The present authors have proposed classification and clustering algorithms based on probabilistic latent semantic indexing (PLSI) model, and applied it to student questionnaire analysis for the purpose of faculty development [HC03] [HC04] [HSY07]. Since the number of documents, which corresponds to the number of students in a class, is usually 30-150, a classification algorithm with high performance for a set of small number of training documents is highly required, although the proposed algorithm exhibits relatively good performance compared with the conventional algorithms such as those using VSM, using latent semantic indexing (LSI) model, and using naive Bayes model. For small training sets, a document classification method has been discussed constructing a hierarchical mixture model which uses the EM algorithm [TCPH01]. The method, however, depends on the structure of given document sets.

In this paper, we propose a new classification method based on Bayesian decision model. Bayesian decision theory is known to give asymptotically the optimum decision algorithm of probabilistic models for infinite training data guaranteeing the property of the consistency, if the model includes the true model. It is also known to exhibit good performance even for a finite number of training data. For example, the output code length of the Bayes code has always the Bayes optimum for any finite input sequence [MIH91]. In the area on information retrieval, the Bayes optimum estimation algorithm for classification has been proposed by treating a probabilistic model in PLSI model as a probabilistic model class [GIH03]. This method uses Bayesian decision theory for the parameter estimation, and guarantees the Bayes optimum in a sense that the average square error between the (true) probability of occurrence of the term in the document and its estimated probability is the minimum. Note that it cannot guarantee, however, that the probability of classification error is always the minimum, although it is still useful for document retrieval systems. On the contrast to this method, the optimum method which performs Bayes decision directly, and minimizes the Bayes risk by the all training document set without using the representative documents has been proposed for a telegraph message classification problem [MO02]. However, the performance has been examined in the range of a large number of training documents such as 10^5 and of a small number of terms appeared in the document set such as 500. It is known that if the number of training documents is infinite, then even the naïve Bayes classifier gives the minimum probability of the classification error [DP97]. Moreover, all terms appeared in documents are adopted and no extraction of terms is discussed.

First, we compare the method by the vector space model with cosine similarity measure and that by a statistical decision model using Bayesian statistics. The methods are applied to the test set BMIR-J2 of Japanese news papers[mainichi95]. As a result, it is shown that the former has a smaller probability of the classification error than the latter for a range less than one hundred training data. This suggests that the latter can be improved by using a latent model, since it would be expected to be asymptotically the optimum if the model is true. Then we propose a Bayesian decision method using the sparse matrix compression [Bishop06] [TB99] such as by singular value decomposition. It takes a mixture model in the compressed dimension with a prior probability, where the sparse matrix stands for a term-document matrix. Although experimental results are not obtained at the present, the proposed method will demonstrate higher or equal performance in the classification error compared to the conventional methods, since it contains them as special cases. By this smaller dimension of the term-document matrix brings the reduction of computational work and storage requirements, and that of noise, redundancy and ambiguity.

Throughout this paper, a vector $\vec{x}$ represents a column vector such as $\vec{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$, and otherwise noted, where T denotes a transpose of the vector (or matrix).

2. PRELIMINARY

2.1 Document Classification Model

Let us define a document classification model.

Let $t_i \in \mathbf{T}_L$ be the *i*-th term in the training document set $\mathbf{D}_L$, and $\vec{d}_j$ be a vector representing the *j*-th document in $\mathbf{D}_L$, where $\mathbf{T}_L$ is a set of terms appeared in $\mathbf{D}_L$, $i = 1,2,\cdots,T$ and $j = 1,2,\cdots,D$.

The *j*-th document vector $\vec{d}_j$ is represented by:

$$\vec{d}_j = (a_{1j}, a_{2j}, \cdots, a_{Tj})^{\mathrm{T}} \quad \in \mathbf{N}^{T\times 1} \tag{2.1}$$

where a_{ij} is the number of t_i appeared in $\vec{d}_j$, i.e., $a_{ij} = tf(t_i, \vec{d}_j)$ [1]. The term-document matrix A is given by:

$$A = [a_{ij}] \quad \in \mathbf{N}^{T\times D} \tag{2.2}$$

Similar to (2.1), a test document $\tilde{d}$ is also represented by:

$$\tilde{\vec{d}} = (\tilde{a}_1, \tilde{a}_2, \cdots, \tilde{a}_T)^{\mathrm{T}} \quad \in \mathbf{N}^{T\times 1} \tag{2.3}$$

where $\tilde{a}_i = tf(t_i, \tilde{\vec{d}})$.

We let the m-th class (category) be denoted by C_m in the class set $\mathbf{C}$, where $m = 1,2,\cdots,M$ and $|\mathbf{C}| = M$.

[Definition 2.1]

Let a set of training ducuments $\vec{d}_j$ and their classes c_j be given by:

$$\mathbf{D}_L(\mathbf{C}) = [(\vec{d}_j, c_j)] \qquad (\mathrm{j} = 1,2,\cdots,\mathrm{D}) \tag{2.4}$$

where (2.4) represents the *j*-th document is a member of class $c_j \in \mathbf{C} = \{C_1, C_2, \cdots, C_M\}$.

Then the document classification problem is written by:

$$\text{for } \tilde{\vec{d}}, \{\mathbf{D}_{\mathrm{L}}(\mathbf{C}), \tilde{\vec{d}}\} \to \tilde{c} \tag{2.5}$$

where $\tilde{c}$ is the estimated class to which $\tilde{\vec{d}}$ is classified.

2. 2 Vector Space Model

As one of the most basic method, we show an algorithm using the vector space model (VSM). From given training documents $\vec{d}_j (j = 1,2,\cdots,D)$, we compute representative document vector $\vec{d}_m^*$ for each class $C_m (m = 1,2,\cdots,M)$:

$$\begin{aligned}\vec{d}^*{}_m &= \frac{1}{N_m} \sum_{\{\vec{d}_j : c_j = C_m\}} \vec{d}_j \\ &= (a_{1m}^*, a_{2m}^*, \cdots, a_{Tm}^*)^{\mathrm{T}} \quad \in \mathbf{R}^{T\times 1}\end{aligned} \tag{2.6}$$

where $N_m = |\{d_j : c_j = C_m\}|$ is the number of training documents with the class C_m. Then we compute the similarity function $s(\vec{d}_m^*, \tilde{\vec{d}})$ between $\vec{d}_m^*$ and $\tilde{\vec{d}}$ using cosine measure:

$$s(\vec{d}_m^*, \tilde{\vec{d}}) = \frac{\left\langle \vec{d}_m^*, \tilde{\vec{d}} \right\rangle}{\left|\vec{d}_m^*\right| \left|\tilde{\vec{d}}\right|} \tag{2.7}$$

where $\langle \vec{x}, \vec{y} \rangle$ denotes the inner product of vectors $\vec{x}$ and $\vec{y}$, and $|\vec{x}|$, the norm of $\vec{x}$.

We decide $\tilde{c}$ for given $\tilde{\vec{d}}$ by:

[1] $tf(\quad,\quad)$ stands for term frequency.

$$\tilde{c} = \arg \max_{C_m ; m=1,2,\cdots,M} s(\vec{d}_m^{\,*}, \vec{\tilde{d}}) \qquad (2.8)$$

In the VSM, an inverse document frequency $idf(t_i)$ is usually used, where $idf(t_i) = \log(D/df(t_i))$, and $df(t_i)$ is the number of the documents in $\mathbf{D}_L$ for which the term t_i appears. Hence a_{ij} in (2.1) and $\tilde{a}_i$ in (2.3) are replaced by

$$a_{ij} = tf(t_i, \vec{d}_j) idf(t_i) \qquad (2.9)$$

and

$$\tilde{a}_i = tf(t_i, \vec{\tilde{d}}) idf(t_i) \qquad (2.10)$$

respectively.

2.3 Statistical Decision Model

Assume an independence between terms t_i and $t_{i'}$ $(i \neq i')$. According to the previous work for a statistical decision model (SDM) [MO02], we can formulate this model as follows:

A statistical model is given by $p(\vec{d}_j \mid c_j, \theta)$, where θ is a parameter set. Letting the true parameter of θ be θ^*, the training set $\mathbf{D}_L(\mathbf{C})$ is generated depending on $p(c_j \mid \theta^*)$ and $p(t_i \mid c_j, \theta^*)$. A pair $(\vec{\tilde{d}}, \tilde{c})$ is also generated by such a way, where $\tilde{c}$ is the true class for given $\vec{\tilde{d}}$. Note that we do not know θ^* and $\tilde{c}$, and only observe $\mathbf{D}_L(\mathbf{C})$ and $\vec{\tilde{d}}$. By defining the 0-1 loss function $l(\tilde{c}, c)$ as

$$l(\tilde{c}, c) = \begin{cases} 0 & , \quad \tilde{c} = c \\ 1 & , \quad \tilde{c} \neq c \end{cases} \qquad (2.11).$$

for $\tilde{c}, c \in \mathbf{C}$, the average loss for unknown θ is given by:

$$L(\tilde{c}, \theta) = \sum_{c \in \mathbf{C}} P(c \mid \theta) l(\tilde{c}, c) \qquad (2.12)$$

Deriving a risk function $R(\tilde{c}, \theta)$ as an expectation of the loss function for $\tilde{c}$ which is the average classification error, and a Bayes risk $BR(\tilde{c})$ which is an expectation of $R(\tilde{c}, \theta)$ over the prior density function $p(\theta)$ of θ, we have the optimum decision algorithm for $\hat{\tilde{c}}$ as

$$\begin{aligned} \hat{\tilde{c}} &= \arg\max_{\tilde{c}\in\mathbf{C}} P(\tilde{c} \mid \mathbf{D}_L(\mathbf{C}), \vec{\tilde{d}}) \\ &= \arg\max_{\tilde{c}\in\mathbf{C}} P(\vec{\tilde{d}} \mid \tilde{c}, \mathbf{D}_L(\mathbf{C})) P(\tilde{c} \mid \mathbf{D}_L) \\ &= \arg\max_{\tilde{c}\in\mathbf{C}} \int_{\Theta} p(\theta \mid \mathbf{D}_L(\mathbf{C})) p(\tilde{c} \mid \theta) d\theta \\ &\quad \times \int_{\Theta} \prod_{i=1}^{|\mathbf{T}|} p(\theta \mid \mathbf{D}_L(\mathbf{C}), \tilde{c}, \tilde{a}^{(i-1)}) p(\tilde{a}_i \mid \tilde{c}, \theta) d\theta \end{aligned} \qquad (2.13)$$

by minimizing $BR(p(\theta))$ after a little manipulation, where $\tilde{a}^{(i)} = (\tilde{a}_1, \tilde{a}_2, \cdots, \tilde{a}_i)$ and $\tilde{a}^{(i)} = \phi$ for $i \leq 0$. The (2.13) gives the optimum decision in a sense that it guarantees the minimum probability of classification error under the Bayesian criteria for finite training documents.

In (2.13), we have slightly deferent two algorithms. One is to let t_i be in $\mathbf{T}_L$ as shown (2.13), the other, t_i, only in $\tilde{\mathbf{T}}$, where $\tilde{\mathbf{T}}$ is the term set appeared in the test document $\vec{\tilde{d}}$. We call the former the normal case (NC), and the latter, MO [MO02]. Usually, $T >> \tilde{T}$, where $T = |\mathbf{T}_L|$ and $\tilde{T} = |\tilde{\mathbf{T}}|$, hence computational work can be reduced by using the MO. Computation methods for (2.13) are shown in Appendix A.

3. EXPERIMENTS

Let us discuss experiments for algorithms based on the VSM and the SDM

3.1 Document Sets

The 8 document sets are constructed by randomly and exclusively choosing documents from the Japanese Newspaper articles (=documents), BMIR-J2 [Mainichi95], where each set is composed of 3 classes (Economics, Sports, and Local), and each class has 50 documents, hence each set has 150 documents. Similarly, the other 8 document sets are constructed, where each set has 300 documents.

3.2 Evaluation Methods

The training documents use one document set among 8 sets, and the test documents, one set from the rest of 7 sets, and repeat it 7×8=56 times. As are seen, we have *M*=3, *D*=3,6,···, 150 (or 300).

3.3 Experimental Results

The results are shown in Figure 3.1 for (a) 150 training documents and Figure 3.2 for (b) 300 training documents The *x*-axis is *D/M* (=*D*/3), and *y*-axis, the average probability of classification error $\Pr(\varepsilon)$. The case of the VSM for tf with the NC is noted by the VSM(tfNC), and the VSM for tfidf with the NC, by the VSM(tfidfNC). Similarly, the cases of the the VSM with the MO, by the VSM(tfMO) and the VSM(tfidfMO), respectively. The case of the SDM with the NC is noted by the SDM(NC), and with the MO, by the SDM(MO), where as we have mentioned, the NC (normal case) uses the term t_i in the training set $\mathbf{T}_L$, and the MO [MO02], only in the test document $\tilde{\mathbf{T}}$. For a comparison, a document classification method by the support vector machine (SVM) for tfidf is also illustrated and noted by the SVM(tfidf).

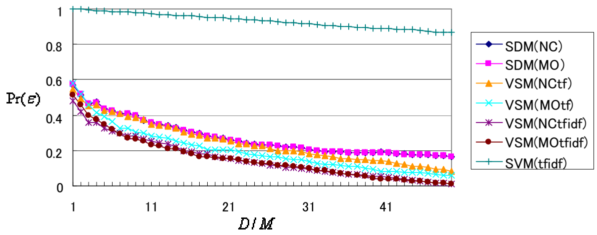

Figure 3.1: The probability of classification error $\Pr(\varepsilon)$ for the number of training documents per each class (D=150).

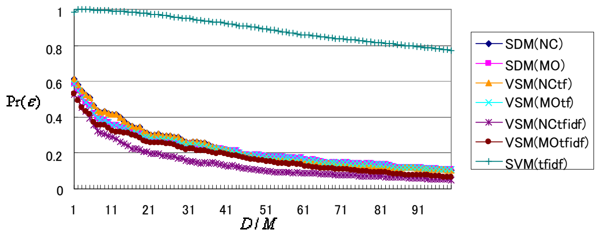

Figure 3.2: The probability of classification error $\Pr(\varepsilon)$ for the number of training documents per each class (*D*=300).

3.4 Remarks

From Figures 3.1 and 3.2, we see that the SVM(tfidf) has larger $\Pr(\varepsilon)$, which is well known for the case of a small number of training documents, since it acts as one to multiple classifier.

The differences between the performance of the SDM(NC) and that of the SDM(MO) is small. We can recognize that the SDM(NC) has slightly smaller $\Pr(\varepsilon)$ than the SDM(MO) in the range of a small number of training documents such as *D*/3<10.

The VSM(tfidfNC) improves the the VSM(tfNC) in $\Pr(\varepsilon)$. And also the VSM(tfidfMO), the VSM(tfMO).

As a conclusion, the VSM is superior compared to the SDM. This suggests us that there is a possibility of the performance improvement of the SDM by introducing a latent model.

4. DISCUSSIONS

As stated in Section 2, the algorithm based on the statistical decision model (SDM) using Bayesian statistics can obtain the optimum decision in a sense that it guarantees the minimum probability of classification error under the Bayesian criteria for a finite number of training documents. If the model includes the true one, the algorithm has the consistency. Unfortunately, actual systems usually cannot be realized by simple mathematical models. Hence we should try to construct closer models to the true one if it exists. When we use Bayesian decision models (BDM), we should carefully choose a prior probability distribution function for parameters so that the computation of posterior probability can be easily performed and be effectively converged [NKM06]. However, it is quite difficult to analyze the behavior of the algorithm using Bayesian statistics for a finite number of training documents, especially in the range of a small number of training data. In other words, we can state nothing with regard to the behavior of the Bayes decision analytically for a finite number of training documents.In the following, we propose an algorithm which will be expected to have good performance for a small number of training data.

First, we use the sparse matrix compression for the term-document matrix A by singular value decomposition (SVD) or probabilistic principal component analysis (P-PCA) [Bishop06] [TB99]. Next, we take a mixture model for the value of the compressed dimension K with a prior probability $p(K)$, which is a kind of Bayesian PCA [Bishop06].

A probabilistic model for a BDM is represented by $p(\tilde{\vec{d}} \mid \tilde{c},\theta,\mathbf{D}_L(\mathbf{C}))$, where the training document set is given by (2.4), and $\theta=(K,\vec{\mu})$, where $K=|C|-1,|C|,\cdots,K_M, K_M=\min(\mathrm{rank}[A],T,D)$

.

Suppose that the matrix $A\in\mathbf{N}^{T\times D}$ is compressed into a matrix $B\in\mathbf{R}^{K\times D}$ by:

$$B=YA \tag{4.1}$$

$$A=[\vec{d}_1,\vec{d}_2,\cdots,\vec{d}_D]=[a_{ij}]$$

$$\vec{d}_j=(a_{1j},a_{2j},\cdots,a_{Tj})^{\mathrm{T}}\in\mathbf{N}^{T\times 1} \tag{4.2}$$

and

$$B=[\vec{d}^{\,*}{}_1,\vec{d}^{\,*}{}_2,\cdots,\vec{d}^{\,*}{}_D]=[b_{ij}],$$

$$\vec{d}^{\,*}{}_j=(b_{1j},b_{2j},\cdots,b_{Kj})^{\mathrm{T}}\in\mathbf{R}^{K\times 1} \tag{4.3}$$

We can simply represent $\vec{d}^{\,*}_j(j=1,2,\cdots,D)$ by $\vec{d}^{\,*}$:

$$\vec{d}^{\,*}=\vec{z}+\vec{\mu}+\vec{\varepsilon} \tag{4.4}$$

where the probability distribution of the k-th latent variable z_k is given by the Gaussian distribution, and we cam assume that z axes are mutually independent.

The matrix $Y \in \mathbf{R}^{K \times T}$ is given by the U_K^T, $U_K = [\vec{u}_1, \vec{u}_2, \cdots, \vec{u}_K]$, where $\vec{u}_k$ is the left eigen vector corresponding to the eigen value λ_k of the AA^{T}, hence

$$\vec{d}_j^* = Y\vec{d}_j \tag{4.5}$$

and

$$\vec{\tilde{d}}^* = Y\vec{\tilde{d}} \tag{4.6}$$

hold. A strategy for choosing larger eigen values will be important to perform effective matrix compression.

[Theorem 4.1] Let A in (4.2) be compressed into B in (4.3) by the P-PCA. Then the Bayesian decision algorithm with a mixture model for K and μ gives the optimum decision $\hat{\tilde{c}}$ for $\vec{\tilde{d}}^*$ as:

$$\hat{\tilde{c}} = \arg \max_{\tilde{c} \in \mathbf{C}} \sum_{K=|\mathbf{C}|-1}^{K_M} \int_{\vec{\mu} \in \mathbf{R}^{K \times 1}} p(\vec{\tilde{d}}^* \mid \tilde{c}, K, \vec{\mu}) \, p(\vec{\mu} \mid \mathbf{D}^*{}_L(\mathbf{C})) p(\tilde{c}) p(K) d\vec{\mu} \tag{4.7}$$

where

$$\mathbf{D}_L^*(\mathbf{C}) = \left[\left(\vec{d}_j^*, c_j\right)\right] \quad (j = 1, 2, \cdots, D) \tag{4.8}$$

Although experimental results are not obtained at the present, the proposed algorithm will demonstrate higher or equal performance in the classification error, since it takes a mixture for K, hence it includes conventional methods as special cases. By this smaller dimension of the term-document matrix brings the reduction of computational work and storage requirements, and that of noise, redundancy and ambiguity.

The above discussions are dependent on our experiences in the field of source coding by the Bayes codes, the LZ codes, the LZW codes and so on, which are usually called the universal source codes [HK02]. The problems on the document classification are quite similar to those on source coding, since actual models for the both problems (document set and information source) are not known exactly and would not be represented by simple mathematical models. Therefore the results discussed on source coding give us suggestive advice much more to the document classification [RJ91].

In this paper, we have proposed relaxed and moderate models for the document classification rather than the complete and full mixture model for parameters so that computational work can be reduced. This implies that we intend to compute the estimated values by intermediate between the model selection and the mixture.

5. CONCLUDING REMARKS

As shown in Section III, the algorithm based on the VSM with cosine similarity measure exhibits fairly good performance for a small number of training document sets compared with that on the SDM. We have extended the algorithm into the case of a mixture model for the value of compressed dimension K. In Theorem 4.1, we should take a mixture model for $\vec{\mu}$ and Y, if the complete mixture model is adopted.

As further works, another document sets such as the English Reuters should be applied.

Acknowledgements

The authors would like to thank Dr. T. Suko for his fruitful comments and discussions. The research leading to this paper was partially supported by Waseda University Grant for Special Research Project No:2008 B-198.

Appendix

Appendix A: Computation methods for (2.13).

Letting a prior probability distribution of θ be assumed to be Dirichlet distribution, we can reduce the computational work, since it is a conjugate prior distribution of multinomial distribution[Ferguson67]. We then have

$$\hat{\tilde{c}} = \arg \max_{C_m : m=1,2,\cdots,M} p(\vec{\tilde{d}} \mid C_m, \mathbf{D}_L(\mathbf{C})) p(C_m \mid \mathbf{D}_L(\mathbf{C})) \tag{A.1}$$

Denoting the number of t_i appeared in $\mathbf{D}_L$ on condition the class C_m be $N(t_i \mid C_m, \mathbf{D}_L(\mathbf{C}))$, i.e.,:

$$N(t_i \mid C_m, \mathbf{D}_L(\mathbf{C})) = \sum_{\{d_j | c_j = C_m\}} tf(t_i, \vec{d}_j) \; (i = 1, 2, \cdots, T) \tag{A.2}$$

equations to compute (A.1) are given by:

(i) for the NC

$$p(\vec{\tilde{d}} \mid \tilde{c}, \mathbf{D}_L(\mathbf{C})) = \frac{\prod_{i=1}^{T} \prod_{u=0}^{\tilde{a}_i - 1} \left(\left(N(t_i \mid \tilde{c}, \mathbf{D}_L(\mathbf{C}) \right) + u + \frac{1}{2} \right)}{\prod_{v=0}^{W-1} \left(\sum_{i=1}^{T} \{ N(t_i \mid \tilde{c}, \mathbf{D}_L(\mathbf{C})) \} + v + \frac{T}{2} \right)}$$

$$= \frac{\prod_{\{i | t_i \in \tilde{\mathbf{T}}\}} \prod_{u=0}^{\tilde{a}_i - 1} \left(N(t_i \mid \tilde{c}, \mathbf{D}_L(\mathbf{C})) + u + \frac{1}{2} \right)}{\prod_{v=0}^{W-1} \left(\sum_{i=1}^{T} \{ N(t_i \mid \tilde{c}, \mathbf{D}_L(\mathbf{C})) \} + v + \frac{T}{2} \right)}$$

(A.3)

where $W = \sum_{\{i|t_i \in \mathbf{T}_L\}} \tilde{a}_i$, and

(ⅱ) for the MO

$$p(\vec{\tilde{d}}\,|\,\tilde{c},\mathbf{D}_L(\mathbf{C})) = \frac{\prod_{i=1}^{T}\prod_{u=0}^{\tilde{a}_i-1}\left(\left(N(t_i\,|\,\tilde{c},\mathbf{D}_L(\mathbf{C})\right)+u+\frac{1}{2}\right)}{\prod_{v=0}^{W-1}\left(\sum_{\{i|t_i\in\tilde{\mathbf{T}}\}}\left\{N(t_i\,|\,\tilde{c},\mathbf{D}_L(\mathbf{C}))\right\}+v+\frac{\tilde{T}}{2}\right)}$$

$$= \frac{\prod_{\{i|t_i\in\tilde{\mathbf{T}}\}}\prod_{u=0}^{\tilde{a}_i-1}\left(N(t_i\,|\,\tilde{c},\mathbf{D}_L(\mathbf{C}))+u+\frac{1}{2}\right)}{\prod_{v=0}^{W-1}\left(\sum_{\{i|t_i\in\tilde{\mathbf{T}}\}}\left\{N(t_i\,|\,\tilde{c},\mathbf{D}_L(\mathbf{C}))\right\}+v+\frac{\tilde{T}}{2}\right)}$$

(A.4)

REFERENCES

[1] [Bishop06] C. M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag New York, 2006.

[2] [DP97] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine Learning, vol.29, pp.103-130, 1997.

[3] [Ferguson67] T. S. Ferguson, Mathematical Statistics, Academic Press, San Diego, U.S.A., 1967.

[4] [GIH03] M. Goto, T. Ishida, and S. Hirasawa, "Representation method for a set of documents from the viewpoint of Bayesian statistics," Proc. of 2003 IEEE Int. Conf. on System, Man and Cybernetics, pp.4637-4642, Washington DC, U.S.A., Oct. 2003.

[5] [HC02] S. Hirasawa and W. W. Chu, "Knowledge acquisition from documents with both fixed and free formats," Proc. of 2003IEEE Int. Conf. on System, Man, and Cybernetics, pp.4694-4699, Washington DC, U.S.A., Oct. 2003.

[6] [HC04] S. Hirasawa and W. W. Chu, "Classification methods for documents with both fixed and free formats by PLSI model," Proc. 2004 International Conference in Management Sciences and Decision Making, Tamkang University, Taipei, May 29, 2004.

[7] [HK02] T. S. Han and K. Kobayashi, Mathematics of Information and Coding (Translations of Mathematical Monographs), American Mathematical Society, 2002.

[8] [HSY07] S. Hirasawa, F-Y. Shih, and W-T. Yang, "Student questionnaire analysis for class management by text mining both in Japanese and in Chinese," Proc. 2007 IEEE Int. Conf. on System, Man and Cybernetics, pp.398-405, Montreal, Canada, Oct. 2007.

[9] [Mainichi95] Mainichi Newspaper CD '94, Naigai Associates, 1995.

[10] [MIH91] T. Matsushima, H. Inazumi, and S. Hirasawa, "A class of distortionless code designed by Bayes decision theory," IEEE Trans. Inform. Theory, vol.37, no.5, pp.1288-1293, 1991.

[11] [MO02] Y. Maeda and H. Ohara, "A note on telegram categorization algorithm based on statistical decision theory," (in Japanese) IPSJ Trans., vol.43, no.10, pp.3119-3126, Oct. 2002.

[12] [NKM06] A. Nakano, D. Koizumi, and T. Matushima, "Text data compression by using Bayes coding algorithms," (in Japanese) IPSJ, Technical report on algorithm, 2007-AL-110-(3), pp.15-22, 2006.

[13] [RJ91] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition :recommendations for practitioners," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, no. 3, Mar 1991.

[14] [TB99] M. E. Tipping and C. M. Bishop, "Probabilistic component analysis," Journal of the Royal Statistical Society, Series B, vol.61, no.3, pp.611-622, 1999.

[15] [TCPH01] K. Toutanova, F. Chen, K. Popat, and T. Hofmann, "Text classification in a hierarchical mixture model for small training sets," Proceedings 10th International Conference on Information and Knowledge Management, 2001.