A Generalization of Independence in Naive Bayes Model

Yu Fujimoto[1] and Noboru Murata[2]

[1] Aoyama Gakuin University,
5-10-1 Fuchinobe, Chuo, Sagamihara, Kanagawa 252-5258, Japan
[2] Waseda University,
3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555, Japan

Abstract. In this paper, generalized statistical independence is proposed from the viewpoint of generalized multiplication characterized by a monotonically increasing function and its inverse function, and it is implemented in naive Bayes models. This paper also proposes an idea of their estimation method which directly uses empirical marginal distributions to retain simplicity of calculation. Our method is interpreted as an optimization of a rough approximation of the Bregman divergence so that it is expected to have a kind of robust property. Effectiveness of our proposed models is shown by numerical experiments on some benchmark data sets.

Keywords: naive Bayes, generalized independence, nonloglinear marginal model, copula, Bregman divergence.

1 Introduction

Statistical models based on some kind of independence, such as naive Bayes (NB) models, Bayesian networks [1] or pLSA [2], are broadly used in various situations; the assumption of independence is attractive in modeling between categorical variables with a lot of categories because the composed model may have a significantly smaller number of parameters than the model denoting dependence. Technically, the assumption of independence in these models should be introduced by analyzing a data set. However, in practical scenes, the models are casually used without rigorous analysis; e.g., in classification problems, it is known that the NB model shows good performance even if the assumption is violated [3]. In this paper, we introduce a generalization of independence and propose an extension of the NB model to express weak special dependence with the small number of parameters.

In the statistical inference, we naturally use arithmetic operators, such as multiplication or division, for probability values. For instance, statistical independence can be defined with multiplication of marginal probabilities. We can generalize these operators with an appropriate monotonically increasing function $u(\cdot)$ and its inverse function $\xi(\cdot)$. For example, multiplication between two positive values a and b are generalized as follows.

$$a \times b = \exp(\log(a) + \log(b)) \xrightarrow{\text{generalization}} u(\xi(a) + \xi(b)).$$

C. Fyfe et al. (Eds.): IDEAL 2010, LNCS 6283, pp. 153–161, 2010.

This type of generalization has been proposed in several contexts (for example, discussions from a perspective of density integration are given in [4,5]), and is closely related to nonloglinear marginal models [6] or the Archimedean copula [7]. In this paper, we generalize conditional independence in NB models by using generalized multiplication, and experimentally show the effectiveness of our proposed model.

This paper is composed as follows. In Section 2, we introduce an idea of generalized independence defined with a monotonically increasing function. Some properties of the generalization are also discussed in this section. In Section 3, the NB model is extended by implementing generalized independence in several ways. In Section 4, we numerically evaluate extended NB models by using benchmark data sets. Lastly, in Section 5, concluding remarks are given.

2 Generalized Independence

In this section, we introduce generalized multiplication denoted with a monotonically increasing function $u(\cdot)$ [8].

2.1 U-multiplication and Generalized Independence

Let $\boldsymbol{X} = \{X^1, \ldots, X^M\}$ be a set of M categorical variables where X^m has a domain $\mathcal{X}^m = \{x_i^m\}_{i=1}^{I_m}$, and $p_{\boldsymbol{\mathcal{X}}}(\boldsymbol{x})$ be a joint probability of $\boldsymbol{x} \in \boldsymbol{X}$. Given marginal probability density functions (pdfs), defined as

$$p_{\mathcal{X}^m}(x^m) = \sum_{l \neq m} \sum_{x^l \in X^l} p_{\boldsymbol{\mathcal{X}}}(x^1, \ldots, x^M), \tag{1}$$

then statistical independence is defined as follows.

Definition 1 (Independence). *Let* $p_\times$ *be the joint pdf defined with marginal pdfs* $p_{\mathcal{X}^1}, \ldots, p_{\mathcal{X}^M}$ *as*

$$p_\times(\boldsymbol{X}; p_{\mathcal{X}^1}, \ldots, p_{\mathcal{X}^M}) = \prod_{m=1}^{M} p_{\mathcal{X}^m}(X^m) = \exp\left(\sum_{m=1}^{M} \log(p_{\mathcal{X}^m}(X^m))\right). \tag{2}$$

Variables $X^1, \ldots, X^M$ *are mutually independent if their joint pdf* $p_{\boldsymbol{\mathcal{X}}}$ *has the following property,*

$$p_{\boldsymbol{\mathcal{X}}}(\boldsymbol{X}) = p_\times(\boldsymbol{X}). \tag{3}$$

Equations (3) and (2) indicate that the sum of logarithmic marginal probabilities in the function $\exp(\cdot)$ defines statistical independence. By introducing a monotonically increasing function $u(\cdot)$, we can generalize Definition 1, as follows.

Table 1. Examples of functions $u(\cdot)$ and $\xi(\cdot)$

	$u(z)$	$\mathrm{dom}(u)$	$\mathrm{range}(u)$	$\xi(z) = u^{-1}(z)$	$\mathrm{dom}(\xi)$	$\mathrm{range}(\xi)$	$\mathrm{dom}(\pi)$
	$\exp(z)$	$(-\infty, \infty)$	$(0, \infty)$	$\log(z)$	$(0, \infty)$	$(-\infty, \infty)$	-
Ex.1	$(\pi z + 1)^{\frac{1}{\pi}}$	$\left[-\frac{1}{\pi}, \infty\right)$	$[0, \infty)$	$\frac{z^\pi - 1}{\pi}$	$[0, \infty)$	$\left[-\frac{1}{\pi}, \infty\right)$	$(0, \infty)$
		$\left(-\infty, -\frac{1}{\pi}\right)$	$(0, \infty)$		$(0, \infty)$	$\left(-\infty, -\frac{1}{\pi}\right)$	$(-\infty, 0)$
Ex.2	$\exp(z) + \pi$	$(-\infty, \infty)$	(π, ∞)	$\log(z - \pi)$	(π, ∞)	$(-\infty, \infty)$	$(-\infty, \inf(z))$
Ex.3	$\exp\left(\mathrm{sgn}(z)\lvert z\rvert^{\frac{1}{\pi}}\right)$	$(-\infty, \infty)$	$(0, \infty)$	$\mathrm{sgn}(\log(z))\lvert\log(z)\rvert^\pi$	$(0, \infty)$	$(-\infty, \infty)$	$(0, \infty)$
Ex.4	$\exp\left(\frac{1-\exp(-z)}{\pi}\right)$	$(-\infty, \infty)$	$\left(0, \exp(\frac{1}{\pi})\right)$	$-\log(1 - \pi\log(z))$	$\left(0, \exp\left(\frac{1}{\pi}\right)\right)$	$(-\infty, \infty)$	$\left(0, \frac{1}{\log(\sup(z))}\right)$

Definition 2 (U-independence [8]). *Let $u(\cdot)$ be a monotonically increasing function, $\xi(\cdot) = u^{-1}(\cdot)$ be its inverse function and $p_\otimes$ be the joint pdf defined by using $u(\cdot)$ and $\xi(\cdot)$ as*

$$p_\otimes(\boldsymbol{X}; p_{\mathcal{X}^1}, \ldots, p_{\mathcal{X}^M}, u) = u\left(\sum_{m=1}^{M} \xi(p_{\mathcal{X}^m}(X^m)) - c\right) \tag{4}$$

$$= p_{\mathcal{X}^1}(X^1) \otimes p_{\mathcal{X}^2}(X^2) \otimes \cdots \otimes p_{\mathcal{X}^m}(X^M)$$

$$= \bigotimes_{m=1}^{M} p_{\mathcal{X}^m}(X^m), \tag{5}$$

where c is a constant to satisfy $\sum_{\boldsymbol{x} \in \boldsymbol{X}} p_\otimes(\boldsymbol{x}) = 1$. Variables $X^1, \ldots, X^M$ are called mutually U-independent if their joint pdf $p_{\boldsymbol{\mathcal{X}}}$ has the following property,

$$p_{\boldsymbol{\mathcal{X}}}(\boldsymbol{X}) = p_\otimes(\boldsymbol{X}). \tag{6}$$

Note that we assume that $range(\sum_{m=1}^{M} \xi(p_{\mathcal{X}^m}) - c) \subseteq dom(u)$ and $range(u) \supseteq range(p_\otimes)$ hold.

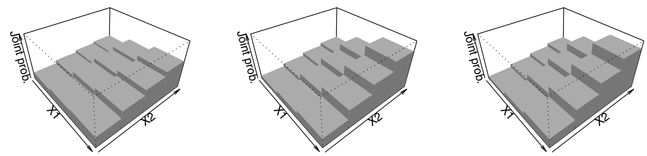
We use the operator $\otimes$ to derive a U-independent pdf, and the operation is called U-multiplication in this paper[1]. In Table 1, ranges and domains of several one-parameter family functions for $p_\otimes$ are shown. The U-independent model is a kind of nonloglinear marginal model [6] and naturally interpreted as a generalization of the independence expression in the loglinear model [9] by using $\xi(\cdot)$ instead of $\log(\cdot)$. Intuitively speaking, U-independence indicates that a sample set is observed in non-independent way and shows kind of weak dependence in the conventional term.

Figure 1 shows intuitive differences between conventional independence and U-independence. As shown in the figure, changing $u(\cdot)$ and $\xi(\cdot)$ in U-multiplication

[1] With Definition 2, we also obtain a marginal pdf from the given U-independent joint pdf as follows,

$$p_{\mathcal{X}^m}(X^m) = u\left(\xi(p_\otimes(\boldsymbol{X})) - \sum_{l \neq m} \xi(p_{\mathcal{X}^l}(X^l)) - c'\right),$$

where c' is a constant to satisfy $\sum_{x^m \in X^m} p_{\mathcal{X}^m}(x^m) = 1$.



(a) $\pi = -0.5$. (b) Conventional indep. (c) $\pi = 0.1$.

Fig. 1. Examples of $p_{\otimes} = p_{\mathcal{X}^1} \otimes p_{\mathcal{X}^2}$ constructed with Ex.1 in Table 1, where $p_{\mathcal{X}^1} = \{0.167, 0.333, 0.5\}$ and $p_{\mathcal{X}^2} = \{0.1, 0.2, 0.3, 0.4\}$. Note that conventional independence indicates $\pi \to 0.0$ in this case.

overstates, or understates, probabilities in marginal distributions. Therefore U-independence is interpreted as an expression of weak special dependence between variables. From another perspective, we can say that U-independence constructs a kind of copula [7] based on pdfs instead of cumulative distribution functions (cdfs). When the variables are continuous, this difference might be a drawback for U-independence because it is impossible to control dependency of upper and lower tails of marginals separately. However, this property is convenient to control these tails simultaneously, and it is handy to control categorical distributions.

2.2 Empirical Marginals

Consider a set of two discrete variables $\boldsymbol{X} = \{X^1, X^2\}$. Let $\mathcal{P}_\times$ and $\mathcal{P}_\otimes$ be sets of independent and U-independent distributions, given as

$$\begin{aligned} \mathcal{P}_\times &= \{p_\times(\boldsymbol{X}; p_{\mathcal{X}^1}, p_{\mathcal{X}^2})\} \\ \mathcal{P}_\otimes(\mathcal{U}) &= \{p_\otimes(\boldsymbol{X}; p_{\mathcal{X}^1}, p_{\mathcal{X}^2}, u) \mid u \in \mathcal{U}\}, \end{aligned}$$

where $\mathcal{U}$ is a set of monotonically increasing functions. Let $\tilde{p}_{\boldsymbol{\mathcal{X}}}$ and $\tilde{p}_{\mathcal{X}^m}$ be empirical distributions, given as

$$\tilde{p}_{\boldsymbol{\mathcal{X}}}(\boldsymbol{x}) = \frac{n_{\boldsymbol{x}}}{\sum_{\boldsymbol{x}' \in \boldsymbol{X}} n_{\boldsymbol{x}'}} \quad \text{and} \quad \tilde{p}_{\mathcal{X}^m}(x^m) = \frac{n_{x^m}}{\sum_{x'^m \in X^m} n_{x'^m}},$$

where $n_{\boldsymbol{x}}$ is the frequency of the observed event $\boldsymbol{x} \in \boldsymbol{X}$ and n_{x^m} is that of $x^m \in X^m$. The maximum likelihood (ML) estimate of the conventional independent model is given with empirical marginals, as follows,

$$\begin{aligned} \operatorname*{argmin}_{q \in \mathcal{P}_\times} \mathrm{D_{KL}}(\tilde{p}_{\boldsymbol{\mathcal{X}}}, q) &= \operatorname*{argmin}_{q \in \mathcal{P}_\times} \sum_{\boldsymbol{x} \in \boldsymbol{X}} \tilde{p}_{\boldsymbol{\mathcal{X}}}(\boldsymbol{x}) \log \frac{\tilde{p}_{\boldsymbol{\mathcal{X}}}(\boldsymbol{x})}{q(\boldsymbol{x})} \\ &= \tilde{p}_{\mathcal{X}^1} \times \tilde{p}_{\mathcal{X}^2}, \end{aligned} \tag{7}$$

where $\mathrm{D_{KL}}(\tilde{p}_{\boldsymbol{\mathcal{X}}}, q)$ is the KL divergence between $\tilde{p}_{\boldsymbol{\mathcal{X}}}$ and q. On the other hand, to obtain the ML estimate of the U-independent model, we need to solve a non-linear optimization problem, that is

$$\operatorname*{argmin}_{q \in \mathcal{P}_\otimes(\mathcal{U})} \mathrm{D_{KL}}(\tilde{p}_{\boldsymbol{\mathcal{X}}}, q) = \hat{u}(\hat{\xi}(\hat{p}_{\mathcal{X}^1}) + \hat{\xi}(\hat{p}_{\mathcal{X}^2}) - c), \tag{8}$$

with respect to marginals $\hat{p}_{\mathcal{X}^1}, \hat{p}_{\mathcal{X}^2}$ and the function $\hat{u}$ for multiplication operator. In this paper, we only focus on searching a function u from a one parameter family $\mathcal{U} = \{u(z;\pi)\}$ which is given as an example in Table 1 and use empirical marginals; i.e., we find the estimate

$$\operatorname*{argmin}_{q \in \tilde{\mathcal{P}}_{\otimes}(\mathcal{U})} \mathrm{D}_{\mathrm{KL}}(\tilde{p}_{\boldsymbol{\mathcal{X}}}, q) = \hat{u}(\hat{\xi}(\tilde{p}_{\mathcal{X}^1}) + \hat{\xi}(\tilde{p}_{\mathcal{X}^2}) - c) \tag{9}$$

where

$$\tilde{\mathcal{P}}_{\otimes}(\mathcal{U}) = \left\{p(\boldsymbol{X}; u) = \tilde{p}_{\mathcal{X}^1}(X^1) \otimes \tilde{p}_{\mathcal{X}^2}(X^2) \mid u \in \mathcal{U}\right\}.$$

Solving Eq.(9) with respect to $\hat{u}$ is much simpler than solving Eq.(8) with respect to $\hat{u}$, $\hat{p}_{\mathcal{X}^1}$ and $\hat{p}_{\mathcal{X}^2}$. We call the solution of Eq.(9) *empirical U-independent model.*

Note that the set $\mathcal{P}_{\otimes}(u)$ defined with any u functions have the following property. Let us define uniform distributions on respective domains as

$$\bar{p}_{\mathcal{X}^1}(X^1) = \frac{1}{I^1}, \quad , \bar{p}_{\mathcal{X}^2}(X^2) = \frac{1}{I^2}, \quad \bar{p}_{\boldsymbol{\mathcal{X}}}(\boldsymbol{X}) = \frac{1}{I^1 \times I^2},$$

where I^1 and I^2 are the number of elements in $\mathcal{X}^1$ and $\mathcal{X}^2$. Then, the following property holds with any types of $u(\cdot)$,

$$\bar{p}_{\boldsymbol{\mathcal{X}}}(\boldsymbol{X}) = \bar{p}_{\mathcal{X}^1}(X^1) \otimes \bar{p}_{\mathcal{X}^2}(X^2). \tag{10}$$

As denoted in the previous subsection, the joint expression of the U-independent distribution is affected by the form of $u(\cdot)$, however it is reduced to $\bar{p}_{\boldsymbol{\mathcal{X}}}$ for any $u(\cdot)$ in the case that all the marginals are uniform distributions. This fact indicates that the space of empirical U-independence $\tilde{P}_{\otimes}$ is not a rich subspace in $\mathcal{P}_{\boldsymbol{\mathcal{X}}}$ if the empirical marginals are close to uniform. On the other hand, when the marginals are far from uniform and have extremely high (or low) probabilities because of small sample sets or outliers, empirical U-independent models can be flexible and convenient candidates.

The ML estimation of the empirical U-independent model is interpreted from the viewpoint of an approximated Bregman divergence [8]. Therefore, with an appropriate function u, the empirical U-independent model enjoys robustness which the Bregman divergence essentially has.

3 Extension of Naive Bayes Model

Let Y be a categorical variable, and $\boldsymbol{X} = \{X^m, \ldots, X^M\}$ be a set of categorical variables. Then, the naive Bayes (NB) model is defined as follows,

$$\begin{aligned} p_{\mathrm{NB}}(\boldsymbol{X}, Y) &= p(Y)p(\boldsymbol{X}|Y) \\ &= p(Y) \prod_{m=1}^{M} p(X^m|Y). \end{aligned} \tag{11}$$

The NB has some convenient properties, such as simple structure, easy estimation and scalability. And it is also known as a simple but robust classification tool [3]. With the empirical joint distribution $\tilde{p}(\boldsymbol{X}, Y)$, the ML estimate is given by

$$\hat{p}_{\mathrm{NB}} = \operatorname*{argmin}_{p_{\mathrm{NB}}} \mathrm{D}_{\mathrm{KL}}(\tilde{p}, p_{\mathrm{NB}}). \tag{12}$$

The concrete form of $\hat{p}_{\mathrm{NB}}(\boldsymbol{X}, Y) = \hat{p}(Y) \prod_{m=1}^{M} \hat{p}(X^m|Y)$ is composed of

$$\hat{p}(y) = \frac{n(y)}{\sum_{y' \in Y} n(y')} \tag{13}$$

$$p(x^m|y) = \frac{n(x^m, y) + \alpha}{\sum_{x'^m \in X^m} (n(x'^m, y) + \alpha)}, \tag{14}$$

where $n(y)$ and $n(x^m, y)$ are the numbers of observations of events y and (x^m, y) respectively, and $\alpha \in [0, 1]$ is a Laplace smoother for estimation of $p(X^m|Y)$.

Now, we consider an extension of the NB model by using U-independence. For example, assume that all the elements in the variable set $\boldsymbol{X}$ are mutually conditional U-independent given Y, we can derive the following expression,

$$p_U(\boldsymbol{X}, Y) = p(Y) \left(\bigotimes_{m=1}^{M} p(X^m|Y) \right). \tag{15}$$

For another example, assume that only some of the elements in $\boldsymbol{X}$ are conditionally U-independent, given as

$$p_U(\boldsymbol{X}, Y) = p(Y) \left(\bigotimes_{m \in \bar{S}_I} p(X^m|Y) \right) \left(\prod_{m' \in S_I} p(X^{m'}|Y) \right), \tag{16}$$

where $\bar{S}_I$ is an index set of weakly dependent variables in $X^1, \ldots, X^M$ and S_I is an index set of the rest.

Given $\tilde{p}$, the ML estimates of Eqs.(15) and (16) are derived by

$$\hat{p}_U = \operatorname*{argmin}_{p_U} \mathrm{D}_{\mathrm{KL}}(\tilde{p}, p_U). \tag{17}$$

Exact solution of Eq.(17) is derived by solving nonlinear optimization problem. Alternatively, we directly use empirical distributions Eqs.(13) and (14) and select an appropriate function u in an analogous way as the discussion of the empirical U-independent model introduced in the previous section. The extended NB models with empirical marginals are called empirical U-NB models in this paper.

4 Numerical Experiments

In this section, we experimentally evaluate the empirical U-NB by using some benchmark data sets. As given in Eqs.(15) and (16), there are some implementation manners of U-independence for the NB model. Therefore, in the first

Table 2. Data sets used in experiments

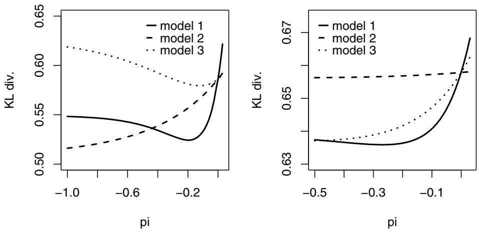
data set	M	# train data	# test data	note
MONKS1	6	124	432	$Y = 1$ when $(X^1 = X^2)$ or $X^5 = 1$
MONKS2	6	169	432	$Y = 1$ when exactly two of $\{X^m\}$ are 1
CAR	6	300	1728	
NUR	8	300	12960	

experiment, we compare some U-NBs extended with different manners. In addition, as denoted at the end of Section 2, the empirical U-independent model is expected to be a good candidate in the case that the sample set is small. In the second experiment, we try to tune the function u in the U-NB by using a small data set. In both experiments, we find an optimal u with respect to π in a one-parameter family Ex.1 in Table 1.

Here, we compare some extended NBs by using data sets "MONK1" and "MONK2" distributed in UCI ML repository [10]. These data sets are composed of a binary class variable Y and a discrete variable set $\boldsymbol{X} = \{X^1, \ldots, X^6\}$. And there is kind of dependence in $\boldsymbol{X}$ to determine Y as shown in Table 2. We compared three models given by Eq.(15) (model 1), Eq.(16) with $\bar{S}_I = \{1, 2\}$ (model 2) and Eq.(16) with $\bar{S}_I = \{3, 4, 5, 6\}$ (model 3); all the models are reduced to Eq.(11) at $\pi = 1$.

At first, we obtained the estimate $\hat{p}_U$ by using an empirical distribution $\tilde{p}$ of the training data set. Then we evaluated it by the KL divergence $\mathrm{D}_{\mathrm{KL}}(p^*, \hat{p}_U)$ where p^* is an empirical distribution of the test data set. In this experiment, we set $\alpha = 0$ in Eq.(14). Figure 2(a) shows the estimation result of "MONK1" data set for various π values. The figure indicates that models 1 and 2 which use U-multiplication for representation of relation between X^1 and X^2 show improvement. On the other hand, model 3 does not improve the result. In a similar way, Figure 2(b) shows the result of "MONK2" which has dependence in all the elements in $\boldsymbol{X}$. In this case, model 2 which uses U-multiplication only between X^1 and X^2 does not show improvement. However, model 3 (which applies U-multiplication for over half of the variables) and model 1 (which applies U-multiplication for all the variables in $\boldsymbol{X}$) show improvement. The results indicate that if there is no knowledge about dependence between variables, then a generalization like model 1 has a possibility to improve the NB model.

Next, we compared the U-NB (model 1) with the conventional NB with an appropriately tuned Laplace smoother. In addition to previously denoted data sets, we also use "car evaluation"(CAR) and "nursery"(NUR) from UCI ML repository in this experiment. At first, we find the optimal $\hat{\alpha}$ for $\hat{p}_{\mathrm{NB}}$ by using training data sets with 10-fold cross validation (CV). Secondly, we find the optimal $\hat{\pi}$ for $\hat{p}_U$ under $\hat{\alpha}$ with CV. Then, we evaluate $\mathrm{D}_{\mathrm{KL}}(p^*, \hat{p}_{\mathrm{NB}})$ and $\mathrm{D}_{\mathrm{KL}}(p^*, \hat{p}_U)$. The experimental results shown in Table 3 indicate that U-NBs outperform conventional NBs even if NBs have appropriately tuned Laplace smoothers. Thus, we see that the estimation of an empirical U-NB has a robust property even when the data set is not composed of a large number of samples.



(a) MONKS1 (b) MONKS2

Fig. 2. Results of experiment 1

Table 3. Results of experiment 2

	$D_{KL}(p^*, \hat{p}_U)$	
data set	NB	U-NB
MONKS1	0.5796	**0.5518** ($\hat{\pi} = -0.12$)
MONKS2	0.6520	**0.6385** ($\hat{\pi} = -1.05$)
CAR	0.6307	**0.6251** ($\hat{\pi} = -0.07$)
NUR	0.4321	**0.4204** ($\hat{\pi} = -0.03$)

5 Conclusion

In this paper, we introduced U-independence and proposed an extension of the NB model. To reduce computational cost, we also proposed the empirical U-NB model which has robust property attributable to the approximated Bregman divergence. In the same manner with the U-NB model, we can extend loglinear models and Bayesian networks. Compared with cross terms in loglinear models or link expression in Bayesian network, U-independence has some advantages in the number of parameters and in robustness; e.g., we expect simple description of graphical models by omitting some links with weak dependence by using U-independence.

It is interesting to use the U-NB as a classifier. Especially, we can combine it with the useful classification method, such as complement naive Bayes [11]. The selection of an appropriate u is another interesting topic though it remains as the future work.

Acknowledgment. This work was partially supported by JSPS, Grant-in-Aid for Young Scientists(B), 22700292.

References

1. Jensen, F.V.: Bayesian Networks and Decision Graphs. In: Statistics for Engineering and Information Science. Springer, Heidelberg (2001)
2. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177–196 (2001)
3. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29, 103–130 (1997)
4. Amari, S.: Integration of stochastic models by minimizing α-divergence. Neural Computation 19, 2780–2796 (2007)
5. Murata, N., Fujimoto, Y.: Bregman divergence and density integration. Journal of Math-for-Industry 1, 97–104 (2009)
6. Bergsma, W., Croon, M., Hagenaars, J.A.: Marginal Models for Dependent, Clustered, and Longitudinal Categorical Data. Springer, Heidelberg (2009)

7. Nelsen, R.B.: An Introduction to Copulas, 2nd edn. Springer Series in Statistics. Springer, Heidelberg (2006)
8. Fujimoto, Y., Murata, N.: A generalized product rule and weak independence based on Bregman divergence. In: Proc. 12th World Multi-Conference on Systemics, Cybernetics and Informatics (2008)
9. Agresti, A.: Categorical Data Analysis, 2nd edn. Wiley Inc., Chichester (2002)
10. Asuncion, A., Newman, D.: UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences (2007)
11. Rennie, J.D., Shih, L., Teevan, J., Karger, D.: Tackling the poor assumptions of naive Bayes text classifiers. In: Proceedings of the Twentieth International Conference on Machine Learning, ICML 2003 (2003)