Data Min Knowl Disc
DOI 10.1007/s10618-008-0101-6

Alternative prior assumptions for improving the performance of naïve Bayesian classifiers

Tzu-Tsung Wong

Received: 20 July 2007 / Accepted: 14 May 2008

Abstract The prior distribution of an attribute in a naïve Bayesian classifier is typically assumed to be a Dirichlet distribution, and this is called the Dirichlet assumption. The variables in a Dirichlet random vector can never be positively correlated and must have the same confidence level as measured by normalized variance. Both the generalized Dirichlet and the Liouville distributions include the Dirichlet distribution as a special case. These two multivariate distributions, also defined on the unit simplex, are employed to investigate the impact of the Dirichlet assumption in naïve Bayesian classifiers. We propose methods to construct appropriate generalized Dirichlet and Liouville priors for naïve Bayesian classifiers. Our experimental results on 18 data sets reveal that the generalized Dirichlet distribution has the best performance among the three distribution families. Not only is the Dirichlet assumption inappropriate, but also forcing the variables in a prior to be all positively correlated can deteriorate the performance of the naïve Bayesian classifier.

Keywords Conjugate · Dirichlet assumption · Generalized Dirichlet distribution · Liouville distribution · Naïve Bayesian classifier

1 Introduction

Naïve Bayesian classifiers are a widely used classification tool in many applications. There are two important assumptions for the functionality of naïve Bayesian

Responsible editor: Charles Elkan.

T.-T. Wong (✉)
Institute of Information Management, National Cheng Kung University, 1, Ta-Sheuh Road, Tainan City 701, Taiwan, ROC
e-mail: tzutsung@mail.ncku.edu.tw

classifiers. The first one is that the prior distribution of a discrete variable or a discretized continuous variable is implicitly or explicitly assumed to be a Dirichlet distribution. Since the Dirichlet distribution is conjugate to the multinomial distribution, the Bayesian updating in naïve Bayesian classifiers is simple. The second assumption specifies that attributes are independent of each other when the class value is given. This conditional independence assumption greatly increases the computational efficiency of naïve Bayesian classifiers.

Although the conditional independence assumption seems to be unreasonable, many researches have shown that this assumption is not as unrealistic as originally thought (Langley et al. 1992; Domingos and Pazzani 1997). The key factor for the feasibility of the conditional independence assumption is that the evaluation of predictions is based on a measure called zero-one loss. A prediction can be either correct, and the minimum zero-one loss is achieved, or wrong regardless of the process for generating the prediction. Thus, even though almost none of the real data sets satisfy the conditional independence assumption, the naïve Bayesian classifier still works well, and sometimes outperforms other classification tools.

The Dirichlet assumption is also essential for the operation of naïve Bayesian classifiers. The popular Laplace's estimate (Cestnik and Bratko 1991) and m-estimate approach (Mitchell 1997) for naïve Bayesian classifiers imply that the prior for an attribute is a Dirichlet distribution. The Dirichlet distribution has many advantages in being a prior for Bayesian analysis, such as the conjugate property, computational efficiency of its moments, and the arbitrariness of variables' order. However, the restrictions on the Dirichlet distribution are strenuous. In a Dirichlet random vector, all pairs of variables must be negatively correlated, and all variables must have the same normalized variance, as will be presented in Sect. 2. These restrictions can be vital for the Dirichlet distribution to be an appropriate prior, but no research has been done to investigate their impact on the performance of the naïve Bayesian classifier.

To study the impact of the Dirichlet assumption, we will pick two multivariate distributions, the generalized Dirichlet and Liouville distributions, that are also defined on the unit simplex as prior distributions to evaluate their classification performance. Both generalized Dirichlet and Liouville distributions allow variables to be positively correlated. Constructing a generalized Dirichlet distribution in which variables have the same mean but different normalized variances is possible. Both of the two distributions are different extensions of the Dirichlet distribution. Thus, naïve Bayesian classifiers will be assumed to have either generalized Dirichlet priors or Liouville priors to investigate the impact of the Dirichlet assumption. Our experimental results will show that the generalized Dirichlet distribution generally has the best performance, which suggests that the Dirichlet assumption is inappropriate for naïve Bayesian classifiers.

This article is organized as follows. Section 2 briefly introduces the basic properties of the Dirichlet distribution and the functionality of the two assumptions for naïve Bayesian classifiers. We will also point out the restrictions of the Dirichlet distribution in being a prior for Bayesian analysis. Some properties of the generalized Dirichlet and the Liouville distributions are presented in Sect. 3. The way to construct either a generalized Dirichlet or Liouville prior for a naïve Bayesian classifier is described in Sect. 4. We then analyze the performance of the naïve Bayesian classifier on 18 real

data sets when it has a Dirichlet prior, a generalized Dirichlet prior, or a Liouville prior, and discuss the experimental results in Sect. 5. The conclusions and future directions of this research are summarized in Sect. 6.

2 The Dirichlet assumption for naïve Bayesian classifiers

In this section, we will introduce some basic properties of the Dirichlet distribution and discuss its restrictions. We then review the operation of naïve Bayesian classifiers to explain the function of the two assumptions: the conditional independence assumption and the Dirichlet assumption.

2.1 Dirichlet distributions

Definition 1 A random vector $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ has a k-variate Dirichlet distribution with parameters $\alpha_j > 0$ for $j = 1, 2, \ldots, k+1$ if it has density

$$f(\Theta) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_{k+1})}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_{k+1})} \prod_{i=1}^{k} \theta_j^{\alpha_j - 1}(1 - \theta_1 - \cdots - \theta_k)^{\alpha_{k+1}-1}$$

for $\theta_1 + \theta_2 + \cdots + \theta_k \leq 1$ and $\theta_j \geq 0$ for $j = 1, 2, \ldots, k$. This distribution will be denoted $D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1})$.

The properties of the Dirichlet distribution can be found in Wilks (1962). The general moment function given in Lemma 1 below can be used to study the properties of the Dirichlet distribution.

Lemma 1 (Wilks 1962) *If* $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ *follows* $D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1})$, *then the general moment function of* Θ *is given by*

$$E(\theta_1^{r_1}\theta_2^{r_2}\cdots\theta_k^{r_k}) = \frac{\Pi_{j=1}^{k}\Gamma(\alpha_j + r_j)\Gamma\left(\Sigma_{j=1}^{k+1}\alpha_j\right)}{\Pi_{j=1}^{k}\Gamma(\alpha_j)\Gamma\left(\Sigma_{j=1}^{k+1}\alpha_j + \Sigma_{j=1}^{k} r_j\right)}.$$

When random vector $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ has a k-variate Dirichlet distribution $D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1})$, by Lemma 1, we have

$$E(\theta_j) = \frac{\alpha_j}{\alpha},$$
$$\mathrm{Var}(\theta_j) = \frac{\alpha_j(\alpha_j + 1)}{\alpha(\alpha + 1)} - \left(\frac{\alpha_j}{\alpha}\right)^2$$

for $j = 1, 2, \ldots, k+1$, where $\alpha = \alpha_1 + \alpha_2 + \cdots + \alpha_{k+1}$. For any $j \neq m$, the covariance between θ_j and θ_m is

$$\mathrm{Cov}(\theta_j, \theta_m) = \frac{\alpha_j\alpha_m}{\alpha(\alpha + 1)} - E(\theta_j)E(\theta_m) = -\frac{1}{\alpha + 1}E(\theta_j)E(\theta_m).$$

Hence, the variables in a Dirichlet random vector are all negatively correlated, and this is called a negative-correlation requirement. Note that the number of parameters in a k-variate Dirichlet distribution is k + 1. In constructing a Dirichlet prior, if the mean probabilities of the variables have been considered, there is only one degree of freedom (by selecting the value of α) that can be used to adjust the spread of the distribution. This implies that a Dirichlet prior is inappropriate for specifying positive correlations among variables.

Lemma 2 (Wong 2005) *Let random vector* $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ *have a k-variate Dirichlet distribution* $D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1})$. *Then*

(1) $E(\theta_j)/E(\theta_m) = Cov(\theta_i, \theta_j)/Cov(\theta_i, \theta_m)$ *for any* $i \neq j, m$.
(2) *If* $E(\theta_j)/E(\theta_m) = b$ *for some* $b \geq 1$, *we will have* $1 \leq Var(\theta_j)/Var(\theta_m) \leq b$.

Lemma 2 shows that in a Dirichlet random vector, $E(\theta_j)/E(\theta_m)$ must be exactly equal to $Cov(\theta_i, \theta_j)/Cov(\theta_i, \theta_m)$ for any i ≠ j, m. In addition, when $E(\theta_j) > E(\theta_m)$, $Var(\theta_j)/Var(\theta_m)$ cannot be larger than $E(\theta_j)/E(\theta_m)$. This implies that when we use the mean probabilities to solve the parameters of a Dirichlet distribution, we also set strenuous constraints on the variances and the covariances of the variables. Part 2 of Lemma 2 also indicates that in a Dirichlet random vector, variables with the same mean will have the same variance.

Bier and Yi (1995) define the normalized variance of a variable U defined on [0,1] as:

$$NV(U) = \frac{Var(U)}{E(U)[1 - E(U)]}.$$

For the variables in a random vector defined on the unit simplex, the normalized variance of each variable can be thought of as the analyst's relative uncertainty about that variable. A variable with a small normalized variance is less uncertain than a variable with a large normalized variance.

Theorem 1 *The variables in a Dirichlet random vector have the same normalized variance.*

Proof Suppose that the joint distribution of $(\theta_1, \theta_2, \ldots, \theta_k)$ is a Dirichlet distribution $D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1})$. By Lemma 1, variable θ_j has a beta distribution with parameters α_j and $\alpha - \alpha_j$, where $\alpha = \alpha_1 + \alpha_2 + \cdots + \alpha_{k+1}$. Thus, the normalized variance of θ_j is

$$NV(\theta_j) = \frac{Var(\theta_j)}{E(\theta_j)[1 - E(\theta_j)]} = \frac{1}{\alpha + 1},$$

which does not depend on the index j. □

Definition 2 Let $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$, and let $V_s = 1 - \theta_1 - \cdots - \theta_s$ for some s < k. Then $(\theta_1, \theta_2, \ldots, \theta_s)$ is said to be neutral if it is independent of $(\theta_{s+1}/V_s, \theta_{s+2}/V_s, \ldots, \theta_k/V_s)$. If $(\theta_1, \theta_2, \ldots, \theta_s)$ is neutral for all s < k, then Θ is said to be completely neutral.

Connor and Mosimann (1969) showed that every permutation of the variables in a Dirichlet random vector is completely neutral. This implies that the order of the variables in a Dirichlet random vector is arbitrary. For example, suppose that $(\theta_1, \theta_2, \theta_3)$ follows $D_3(\alpha_1, \alpha_2, \alpha_3; \alpha_4)$, and let $\theta_4 = 1 - \theta_1 - \theta_2 - \theta_3$. Then $(\theta_3, \theta_1, \theta_2)$ follows $D_3(\alpha_3, \alpha_1, \alpha_2; \alpha_4)$, $(\theta_2, \theta_4, \theta_1)$ follows $D_3(\alpha_2, \alpha_4, \alpha_1; \alpha_3)$, and so on. This symmetric property makes the construction of a Dirichlet prior much easier. Theorem 1 tells us that in selecting a Dirichlet prior, our confidence levels for all variables, measured by normalized variances, must be the same. This restriction is called the equal-confidence requirement.

For a training data set with n instances, let y_j be the number of occurrences of the jth possible outcome of an attribute for $j = 1, 2, \ldots, k+1$. Then the training data for this attribute can be represented as $\mathbf{y} = \{y_1, y_2, \ldots, y_{k+1}\}$. Suppose that the likelihood function of the data $L(\mathbf{y}|\Theta)$ follows a multinomial distribution. Since the posterior density $f(\Theta|\mathbf{y})$ is proportional to the product $L(\mathbf{y}|\Theta)f(\Theta)$, it is not difficult to show that the Dirichlet distribution is conjugate to the multinomial likelihood function, as in the result given in the following lemma.

Lemma 3 (Wong 2007) *When* $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ *follows* $D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1})$, *and* $L(\mathbf{y}|\Theta)$ *follows a multinomial distribution, the posterior density* $f(\Theta|\mathbf{y})$ *is* $D_k(\alpha'_1, \alpha'_2, \ldots, \alpha'_k; \alpha'_{k+1})$, *where* $\alpha'_j = \alpha_j + y_j$ *for* $j = 1, 2, \ldots, k+1$.

2.2 Naïve Bayesian classifiers

A classification problem usually involves an instance $\mathbf{x}$ with attribute values $x_1, x_2, \ldots, x_N$ for determining its class value c. The naïve Bayesian classifier calculates the conditional probability $p(c_j|\mathbf{x})$ for all classes c_j and picks the class with the largest conditional probability to be the predicted class of instance $\mathbf{x}$. By the Bayes' theorem, this conditional probability can be rewritten as

$$p(c_j|\mathbf{x}) = \frac{p(\mathbf{x}|c_j)p(c_j)}{p(\mathbf{x})} \propto p(c_j)p(x_1, x_2, \ldots, x_N|c_j).$$

The proportion of the instances with class c_j in the training data is an estimate of $p(c_j)$. If we can estimate $p(x_1, x_2, \ldots, x_N|c_j)$ for any given x_i and c_j, then the predicted class of $\mathbf{x}$ can be determined.

In general, estimating $p(x_1, x_2, \ldots, x_N|c_j)$ for all possible x_i and c_j from available data can be difficult, or some of the estimates can be unreliable, except when the data size is large. If the attributes are independent when the class value is given, the estimate of this probability can be simplified as $\Pi_{i=1}^{N} p(x_i|c_j)$, which is named the conditional independence assumption. Probability $p(x_i|c_j)$ represents the proportion of the instances with class c_j and attribute $X_i = x_i$ to the instances with class c_j. Estimating $p(x_i|c_j)$ for any given x_i and c_j will not be a problem, except when the number of instances with class c_j and attribute $X_i = x_i$ is zero. When $p(x_i|c_j) = 0$ for some i, the value of $p(c_j|\mathbf{x})$ will be zero regardless of the values of $p(x_m|c_j)$ for all $m \neq i$. This can greatly distort the classification result. Many researchers therefore choose

the Laplace's estimate $\alpha_j = 1$ for all j to compose a noninformative Dirichlet prior; i.e., all possible outcomes of an attribute have the same prior mean probability.

For each possible class value c_j, a (discretized) attribute X with $k+1$ possible outcomes has a random vector $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ to represent the occurring probabilities of its possible outcomes 1 through k. The Dirichlet assumption for naïve Bayesian classifiers sets a Dirichlet prior $D_k(1, 1, \ldots, 1; 1)$ for random vector $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ when the Laplace's estimate is used. Since every instance is collected independently, the likelihood function of available data $\mathbf{y}$ given Θ will follow a multinomial distribution. By Lemma 3, random vector $\Theta|\mathbf{y}$ will have a Dirichlet distribution $D_k(\alpha'_1, \alpha'_2, \ldots, \alpha'_k; \alpha'_{k+1})$, where $\alpha'_j = y_j + 1$ for $j = 1, 2, \ldots, k+1$. Let θ_m be the variable corresponding to possible outcome x_i of attribute X_i. Then the posterior mean $E(\theta_m|\mathbf{y}) = (y_m + 1)/(n + k + 1)$ is the estimate of $p(x_i|c_j)$ for calculating the classification probability $p(c_j|\mathbf{x})$. In this case, $p(x_i|c_j)$ will be positive even though the available data does not include any instance with class c_j and attribute $X_i = x_i$. The naïve Bayesian classifier can therefore work properly. The m-estimate approach analyzes the training data to set a positive value, not necessarily equal to one, for each parameter α_j in a Dirichlet prior.

3 Generalized Dirichlet and Liouville distributions

As pointed out in Sect. 2, there are some restrictions on the Dirichlet distribution. It should be of interest to know whether the Dirichlet assumption has a severe impact on the performance of the naïve Bayesian classifier. The Dirichlet distribution is a special multivariate distribution generated from a truncated stick-breaking process that recursively bipartitions a unit into positive fractions by using independent beta random variables (Ishwaran and James 2001). Since the beta random variables employed in the truncated stick-breaking process for a Dirichlet distribution are independent and in an arbitrary order, Aitchison (1986) pointed out that the variables in a Dirichlet random vector exhibit strong conditional independent relationships. However, the Dirichlet distribution is still popular for analyzing compositional data because of its conjugate property in Bayesian analysis and computational efficiency. In this section, we will introduce two multivariate distributions that are not as restrictive as the Dirichlet distribution in some aspects and can be generated from the truncated stick-breaking process. In addition, they can also be appropriate priors for the naïve Bayesian classifier.

3.1 Generalized Dirichlet distributions

Definition 3 A random vector $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ has a k-variate generalized Dirichlet distribution with parameters $\alpha_j > 0$ and $\beta_j > 0$ for $j = 1, 2, \ldots, k$ if it has density

$$f(\Theta) = \prod_{j=1}^{k} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \theta_j^{\alpha_j - 1} (1 - \theta_1 - \cdots - \theta_j)^{\lambda_j}$$

for $\theta_1 + \theta_2 + \cdots + \theta_k \leq 1$ and $\theta_j \geq 0$ for $j = 1, 2, \ldots, k$, where $\lambda_j = \beta_j - \alpha_{j+1} - \beta_{j+1}$ for $j = 1, 2, \ldots, k-1$ and $\lambda_k = \beta_k - 1$. This distribution will be denoted $GD_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$.

For a random vector $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$, let $Z_1 = \theta_1$ and $Z_j = \theta_j / V_{j-1}$ for $j = 2, 3, \ldots, k$, where $V_{j-1} = 1 - \theta_1 - \cdots - \theta_{j-1}$. If the Z_j are independent, then Θ is completely neutral. Connor and Mosimann (1969) assumed that each of the Z_j has a beta distribution with parameters α_j and β_j, and derived the density function for the generalized Dirichlet distribution as given in Definition 3. Wong (1998) used the concept of complete neutrality to derive the general moment function for the generalized Dirichlet distribution, as given in Lemma 4 below.

Lemma 4 (Wong 1998) *Let $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ be a vector random variable having a k-variate generalized Dirichlet distribution $GD_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$. Then the general moment function of $(\theta_1, \theta_2, \ldots, \theta_k)$ is*

$$E(\theta_1^{r_1}\theta_2^{r_2}\cdots\theta_k^{r_k}) = \prod_{j=1}^{k} \frac{\Gamma(\alpha_j + \beta_j)\Gamma(\alpha_j + r_j)\Gamma(\beta_j + \delta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)\Gamma(\alpha_j + \beta_j + r_j + \delta_j)}$$

where $\delta_j = r_{j+1} + r_{j+2} + \cdots + r_k$ for $j = 1, 2, \ldots, k-1$, and $\delta_k = 0$.

By Lemma 4, we have

$$E(\theta_j) = E\left[Z_j \prod_{i=1}^{j-1}(1 - Z_i)\right] = \frac{\alpha_j}{\alpha_j + \beta_j} \prod_{i=1}^{j-1} \frac{\beta_i}{\alpha_i + \beta_i},$$

$$\mathrm{Var}(\theta_j) = \frac{\alpha_j(\alpha_j + 1)}{(\alpha_j + \beta_j)(\alpha_j + \beta_j + 1)} \prod_{i=1}^{j-1} \frac{\beta_i(\beta_i + 1)}{(\alpha_i + \beta_i)(\alpha_i + \beta_i + 1)} - E(\theta_j)^2$$

for $j = 1, 2, \ldots, k$, and

$$E(\theta_{k+1}) = E\left[\prod_{i=1}^{k}(1 - Z_i)\right] = \prod_{i=1}^{k} \frac{\beta_i}{\alpha_i + \beta_i},$$

$$\mathrm{Var}(\theta_{k+1}) = \prod_{i=1}^{k} \frac{\beta_i(\beta_i + 1)}{(\alpha_i + \beta_i)(\alpha_i + \beta_i + 1)} - E(\theta_{k+1})^2.$$

Connor and Mosimann (1969) also showed that

$$\mathrm{Cov}(\theta_1, \theta_j) = -\frac{E(\theta_j)}{E(1 - \theta_1)}\mathrm{Var}(\theta_1) \text{ for } j = 2, 3, \ldots, k+1,$$

$$\mathrm{Cov}(\theta_j, \theta_{j+1}) = E(Z_{j+1})E[Z_j(1 - Z_j)]\prod_{i=1}^{j-1} E[(1 - Z_i)^2] - E(\theta_j)E(\theta_{j+1})$$
$$\text{for } j = 2, 3, \ldots, k-1,$$

and

$$\mathrm{Cov}(\theta_j, \theta_m) = \left[\frac{E(Z_m)}{E(Z_{j+1})}\right]\left[\prod_{i=j+1}^{m-1} E(1 - Z_i)\right]\mathrm{Cov}(\theta_j, \theta_{j+1}) \text{ for } 1 < j < m < k.$$

Thus, θ_1 is always negatively correlated with all other random variables. However, Lochner (1975) showed that $\mathrm{Cov}(\theta_j, \theta_m)$ can be positive for $j, m > 1$. If there exists some $m > j$ such that θ_j and θ_m are positively (negatively) correlated, then θ_j and θ_i will be positively (negatively) correlated for all $i > j$.

When $\beta_j = \alpha_{j+1} + \beta_{j+1}$ for $j = 1, 2, \ldots, k-1$, the generalized Dirichlet distribution $GD_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$ reduces to the Dirichlet distribution $D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_k)$. If $(\theta_1, \theta_2, \ldots, \theta_k)$ has a generalized Dirichlet distribution, then $(\theta_1, \theta_2, \ldots, \theta_k)$ is completely neutral. However, this does not mean that every permutation of $(\theta_1, \theta_2, \ldots, \theta_k)$ is also completely neutral. For example, if $(\theta_1, \theta_2, \theta_3)$ follows $GD_3(\alpha_1, \alpha_2, \alpha_3; \beta_1, \beta_2, \beta_3)$ and $\beta_1 \neq \alpha_2 + \beta_2$, then $(\theta_2, \theta_1, \theta_3)$ will not have a generalized Dirichlet distribution. So, when $(\theta_1, \theta_2, \ldots, \theta_k)$ has a generalized Dirichlet distribution, the order of the θ_j is generally not arbitrary.

Theorem 2 *A generalized Dirichlet distribution reduces to a Dirichlet distribution if and only if all variables in the generalized Dirichlet random vector have the same normalized variance.*

Proof Suppose that $(\theta_1, \theta_2, \ldots, \theta_k)$ follows the generalized Dirichlet distribution $GD_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$. If the generalized Dirichlet distribution reduces to a Dirichlet distribution, then by Theorem 1, the θ_j have the same normalized variance. For the necessity of this theorem, we will show by induction that if the θ_j have the same normalized variance, the parameters must satisfy $\beta_j = \alpha_{j+1} + \beta_{j+1}$ for $j = 1, 2, \ldots, k-1$. Let $Z_1 = \theta_1$ and $Z_j = \theta_j/(1 - \theta_1 - \cdots - \theta_{j-1})$ for $j = 2, 3, \ldots, k-1$. Then the normalized variances of θ_1 and θ_2 are

$$\mathrm{NV}(\theta_1) = \mathrm{NV}(Z_1) = \frac{1}{\alpha_1 + \beta_1 + 1}$$

and

$$\begin{aligned}\mathrm{NV}(\theta_2) &= \frac{\mathrm{Var}(\theta_2)}{\mathrm{E}(X_2)[1 - \mathrm{E}(X_2)]} = \frac{\mathrm{E}(Z_2^2)\mathrm{E}[(1 - Z_1)^2] - \mathrm{E}(Z_2)^2\mathrm{E}(1 - Z_1)^2}{\mathrm{E}(Z_2)\mathrm{E}(1 - Z_1)[1 - \mathrm{E}(Z_2)\mathrm{E}(1 - Z_1)]} \\ &= \frac{(\alpha_2 + 1)(\beta_1 + 1)(\alpha_2+\beta_2)(\alpha_1+\beta_1) - \alpha_2\beta_1(\alpha_2 + \beta_2 + 1)(\alpha_1 + \beta_1 + 1)}{(\alpha_2 + \beta_2 + 1)(\alpha_1 + \beta_1 + 1)[(\alpha_2 + \beta_2)(\alpha_1 + \beta_1) - \alpha_2\beta_1]}.\end{aligned}$$

By setting $\mathrm{NV}(\theta_1) = \mathrm{NV}(\theta_2)$, we have $\beta_1 = \alpha_2 + \beta_2$.

Now, suppose that for some $s < k$, if the normalized variances of θ_1 through θ_{s-1} are the same, then $\beta_j = \alpha_{j+1} + \beta_{j+1}$ for $j = 1, 2, \ldots, s-2$. The normalized variance of θ_s is

$$\mathrm{NV}(\theta_s) = \frac{\mathrm{Var}(\theta_s)}{\mathrm{E}(\theta_s)[1 - \mathrm{E}(\theta_s)]} = \frac{\mathrm{E}(Z_s^2)\prod_{j=1}^{s-1}\mathrm{E}[(1 - Z_j)^2] - \left[\mathrm{E}(Z_s)\prod_{j=1}^{s-1}\mathrm{E}(1 - Z_j)\right]^2}{\mathrm{E}(Z_s)\prod_{j=1}^{s-1}\mathrm{E}(1 - Z_j)\left[1 - \mathrm{E}(Z_s)\prod_{j=1}^{s-1}\mathrm{E}(1 - Z_j)\right]}.$$

Since $\beta_j = \alpha_{j+1} + \beta_{j+1}$ for $j = 1, 2, \ldots, s-2$, we have

$$\prod_{j=1}^{s-1} E[(1-Z_j)^2] = \prod_{j=1}^{s-1} \frac{\beta_j(\beta_j+1)}{(\alpha_j+\beta_j)(\alpha_j+\beta_j+1)} = \frac{\beta_{s-1}(\beta_{s-1}+1)}{(\alpha_1+\beta_1)(\alpha_1+\beta_1+1)}$$

and

$$\prod_{j=1}^{s-1} E(1-Z_j) = \prod_{j=1}^{s-1} \frac{\beta_j}{\alpha_j+\beta_j} = \frac{\beta_{s-1}}{\alpha_1+\beta_1}.$$

Hence, the normalized variance of θ_s will be

$$\begin{aligned} NV(\theta_s) &= \frac{E(Z_s^2)\Pi_{j=1}^{s-1}E[(1-Z_j)^2] - \left[E(Z_s)\Pi_{j=1}^{s-1}E(1-Z_j)\right]^2}{E(Z_s)\Pi_{j=1}^{s-1}E(1-Z_j)\left[1-E(Z_s)\Pi_{j=1}^{s-1}E(1-Z_j)\right]} \\ &= \frac{(\alpha_s+1)(\beta_{s-1}+1)(\alpha_s+\beta_s)(\alpha_1+\beta_1) - \alpha_s\beta_{s-1}(\alpha_s+\beta_s+1)(\alpha_1+\beta_1+1)}{(\alpha_s+\beta_s+1)(\alpha_1+\beta_1+1)[(\alpha_s+\beta_s)(\alpha_1+\beta_1) - \alpha_s\beta_{s-1}]}. \end{aligned}$$

By setting $NV(\theta_1) = NV(\theta_s)$, we have $\beta_{s-1} = \alpha_s + \beta_s$. By induction, if the normalized variances are all the same, then $\beta_j = \alpha_{j+1} + \beta_{j+1}$ for $j = 1, 2, \ldots, k-1$. □

Theorem 2 indicates that the generalized Dirichlet distribution is a suitable prior for the variables with different confidence levels. It can be a prior for the variables with the same confidence level only when it reduces to a Dirichlet distribution.

Lemma 5 (Wong 1998) *When $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ follows $GD_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$, and likelihood function $L(\mathbf{y}|\Theta)$ follows a multinomial distribution, the posterior density $f(\Theta|\mathbf{y})$ is $GD_k(\alpha'_1, \alpha'_2, \ldots, \alpha'_k; \beta'_1, \beta'_2, \ldots, \beta'_k)$, where $\alpha'_j = \alpha_j + y_j$ and $\beta'_j = \beta_j + y_{j+1} + \cdots + y_{k+1}$ for $j = 1, 2, \ldots, k$.*

Lemma 5 shows that the generalized Dirichlet distribution is also conjugate to the multinomial sampling. When an attribute given class value c_j is assumed to have a generalized Dirichlet prior, the posterior mean $E(\theta_m|y) = \frac{\alpha'_m}{\alpha'_m+\beta'_m}\prod_{i=1}^{m-1}\frac{\alpha'_i}{\alpha'_i+\beta'_i}$ is an estimate of $p(x_i|c_j)$ for calculating the classification probability $p(c_j|\mathbf{x})$. Thus, the generalized Dirichlet distribution can be an appropriate prior for the naïve Bayesian classifier.

3.2 Liouville distributions

Definition 4 A random vector $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ has a k-variate Liouville distribution with parameters α_j for $j = 1, 2, \ldots, k$ and density generator $g(u)$ if it has density

$$f(\Theta) = C_0 g(u) \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

for $\theta_1+\theta_2+\cdots+\theta_k \le 1$ and $\theta_j \ge 0$ for $j = 1, 2, \ldots, k$, where $u = \theta_1+\theta_2+\cdots+\theta_k$ and C_0 is a normalizing constant. This distribution will be denoted $L_k(g(u); \alpha_1, \alpha_2, \ldots, \alpha_k)$.

Let $(Z_1, Z_2, \ldots, Z_{k-1})$ follow $D_{k-1}(\alpha_1, \alpha_2, \ldots, \alpha_{k-1}; \alpha_k)$, and let U defined on [0,1] be an independent random variable with probability density function f(u). Set $Z_k = 1 - Z_1 - \cdots - Z_{k-1}$ and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_k)$. Fang et al. (1990) showed that $\Theta = U\mathbf{Z}$ has a Liouville distribution $L_k(g(u); \alpha_1, \alpha_2, \ldots, \alpha_k)$, where $g(u) \propto u^{-(\alpha-1)}f(u)$ and $\alpha = \alpha_1+\alpha_2+\cdots+\alpha_k$. This result can be used to derive the general moment function of the Liouville distribution, as given in Lemma 6 below.

Lemma 6 (Wong 2007) *Let μ_r be the rth moment of U; i.e., $\mu_r = E(U^r)$. If Θ has a Liouville distribution $L_k(g(u); \alpha_1, \alpha_2, \ldots, \alpha_k)$, then the general moment function of Θ is*

$$E(\theta_1^{r_1}\theta_2^{r_2}\cdots\theta_k^{r_k}) = \mu_r \frac{\Pi_{j=1}^k \Gamma(\alpha_j + r_j)\Gamma\left(\Sigma_{j=1}^k \alpha_j\right)}{\Pi_{j=1}^k \Gamma(\alpha_j)\Gamma\left(\Sigma_{j=1}^k \alpha_j + r\right)},$$

where $r = r_1 + r_2 + \cdots + r_k$.

When the density generating variate U has a beta distribution with parameters γ and ω such that $\gamma = \alpha_1+\alpha_2+\cdots+\alpha_k$, by Lemma 6, we have $g(u) \propto (1-u)^{\omega-1}$ and

$$\begin{aligned} E(\theta_1^{r_1}\theta_2^{r_2}\cdots\theta_k^{r_k}) &= \frac{\Gamma(\gamma+\omega)\Gamma(\gamma+r)}{\Gamma(\gamma+\omega+r)\Gamma(\gamma)} \cdot \frac{\Pi_{j=1}^k \Gamma(\alpha_j + r_j)\Gamma\left(\Sigma_{j=1}^k \alpha_j\right)}{\Pi_{j=1}^k \Gamma(\alpha_j)\Gamma\left(\Sigma_{j=1}^k \alpha_j + r\right)} \\ &= \frac{\Pi_{j=1}^k \Gamma(\alpha_1 + r_1)\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_k+\omega)}{\Pi_{j=1}^k \Gamma(\alpha_j)\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_k+\omega+r)} \end{aligned}$$

which is the general moment function of the Dirichlet distribution $D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \omega)$. This means when the density generating variate U has a beta distribution with parameters γ and ω, the Liouville distribution will reduce to a Dirichlet distribution if $\gamma = \alpha_1+\alpha_2+\cdots+\alpha_k$.

Let μ_1 and μ_2 be the first and second moments of U. Then by Lemma 6, we have

$$E(\theta_j) = \mu_1 \frac{\alpha_j}{\alpha},$$

$$\mathrm{Var}(\theta_j) = \mu_2 \frac{\alpha_j(\alpha_j+1)}{\alpha(\alpha+1)} - \mu_1^2 \frac{\alpha_j^2}{\alpha^2},$$

and

$$\mathrm{Cov}(\theta_i, \theta_j) = \frac{\alpha_i\alpha_j}{\alpha}\left(\frac{\mu_2}{\alpha+1} - \frac{\mu_1^2}{\alpha}\right) \text{ for } i \ne j,$$

where $\alpha = \alpha_1+\alpha_2+\cdots+\alpha_k$. Since α, α_i, and α_j are all positive, we have

$$\begin{aligned}\mathrm{Cov}(\theta_\mathrm{i}, \theta_\mathrm{j}) &= \frac{\alpha_\mathrm{i}\alpha_\mathrm{j}}{\alpha}\left(\frac{\mu_2}{\alpha+1} - \frac{\mu_1^2}{\alpha}\right) > 0 \\ &\Rightarrow \sigma_\mathrm{U}/\mu_1 > 1/\sqrt{\alpha},\end{aligned}$$

where $\sigma_\mathrm{U}^2 = \mathrm{Var(U)}$. Thus, for any $\mathrm{i} \neq \mathrm{j}$, θ_i and θ_j will be positively correlated if variable U has a coefficient of variation greater than $1/\sqrt{\alpha}$. Note that if there exist $\mathrm{i} \neq \mathrm{j}$ such that θ_i and θ_j are positively (negatively) correlated, then θ_m and θ_q must be positively (negatively) correlated for any $\mathrm{m} \neq \mathrm{q}$.

Theorem 3 *The Liouville distribution reduces to a Dirichlet distribution if and only if all variables in the Liouville random vector have the same normalized variance and the density generator variate U has a beta distribution.*

Proof Suppose that the joint distribution of $(\theta_1, \theta_2, \ldots, \theta_\mathrm{k})$ is the Liouville distribution $\mathrm{L_k(g(u); \alpha_1, \alpha_2, \ldots, \alpha_k)}$, and let $\alpha = \alpha_1 + \alpha_2 + \cdots + \alpha_\mathrm{k}$. By Theorem 1, when the Liouville distribution reduces to a Dirichlet distribution, the θ_j will have the same normalized variance. Moreover, if distributions $\mathrm{D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1})}$ and $\mathrm{L_k(g(u); \alpha_1, \alpha_2, \ldots, \alpha_k)}$ are identical, by comparing their general moment functions, the rth moment of U must be

$$\mathrm{E(U^r)} = \frac{\Gamma(\alpha + \alpha_{\mathrm{k}+1})\,\Gamma(\alpha + \mathrm{r})}{\Gamma(\alpha + \alpha_{\mathrm{k}+1} + \mathrm{r})\,\Gamma(\alpha)},$$

which is the rth moment of a beta distribution with parameters α and $\alpha_{\mathrm{k}+1}$. Since U is defined on a compact support, U must have a beta distribution. For the necessity of this theorem, suppose that U has a beta distribution with parameters γ and ω, and let $\theta_{\mathrm{k}+1} = 1 - \mathrm{U}$. Then for any $\mathrm{j} < \mathrm{k} + 1$, since θ_j and $\theta_{\mathrm{k}+1}$ have the same normalized variance, we have

$$\begin{aligned}\mathrm{NV}(\theta_\mathrm{j}) = \mathrm{NV}(\theta_{\mathrm{k}+1}) &= \frac{\mathrm{Var}(\theta_{\mathrm{k}+1})}{\mathrm{E}(\theta_{\mathrm{k}+1})[1 - \mathrm{E}(\theta_{\mathrm{k}+1})]} = \mathrm{NV(U)} = \frac{1}{\gamma + \omega + 1} \\ \Rightarrow \frac{\mathrm{Var}(\theta_\mathrm{j})}{\mathrm{E}(\theta_\mathrm{j})[1 - \mathrm{E}(\theta_\mathrm{j})]} &= \frac{\mathrm{E(U^2)}\frac{\alpha_\mathrm{j}(\alpha_\mathrm{j}+1)}{\alpha(\alpha+1)} - \left(\frac{\alpha_\mathrm{j}}{\alpha}\right)^2 \mathrm{E(U)}^2}{\frac{\alpha_\mathrm{j}}{\alpha}\mathrm{E(U)}\left[1 - \frac{\alpha_\mathrm{j}}{\alpha}\mathrm{E(U)}\right]} \\ &= \frac{\alpha(\gamma+\omega)(\alpha_\mathrm{j}+1)(\gamma+1) - \alpha_\mathrm{j}\gamma(\alpha+1)(\gamma+\omega+1)}{(\alpha+1)(\gamma+\omega+1)[\alpha(\gamma+\omega) - \alpha_\mathrm{j}\gamma]} \\ &= \frac{1}{\gamma+\omega+1} \\ \Rightarrow (\alpha - \gamma)(\alpha_\mathrm{j} - \alpha) &= 0.\end{aligned}$$

Since $\alpha_\mathrm{j} < \alpha$, we have $\alpha = \gamma$. This means that the Liouville distribution reduces to a Dirichlet distribution. □

Note the fact that all variables having the same normalized variance is not enough to ensure that the Liouville distribution will reduce to a Dirichlet distribution. However,

if variables θ_1 through θ_k have the same mean, they will have the same normalized variance.

Lemma 7 (Wong 2007) *When* $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ *follows* $L_k(g(u); \alpha_1, \alpha_2, \ldots, \alpha_k)$, *and* $L(\mathbf{y}|\Theta)$ *follows a multinomial distribution, the posterior density* $f(\Theta|\mathbf{y})$ *is* $L_k(h(u); \alpha'_1, \alpha'_2, \ldots, \alpha'_k)$, *where* $\alpha'_j = \alpha_j + y_j$ *for* $j = 1, 2, \ldots, k$, *and* $h(u) = g(u)(1-u)^{y_{k+1}}$. *In particular, when U has a beta distribution with parameters* γ *and* ω, $U|\mathbf{y}$ *will have a beta distribution with parameters* $\gamma + y_1 + y_2 + \cdots + y_k$ *and* $\omega + y_{k+1}$.

In a naïve Bayesian classifier, when an attribute given class value c_j has a Liouville prior, the estimate of $p(x_i|c_j)$ is the posterior mean $E(\theta_m|\mathbf{y}) = E(U|\mathbf{y})\frac{\alpha'_m}{\alpha'}$, where $\alpha' = \alpha'_1 + \alpha'_2 + \cdots + \alpha'_k$. Thus, the Liouville distribution can also be a prior for naïve Bayesian classifiers.

4 Prior construction

Both the generalized Dirichlet and the Liouville distributions include the Dirichlet distribution as a special case, and they can overcome some restrictions of the Dirichlet distribution. We can use them to replace Dirichlet distributions as priors to investigate the impact of the Dirichlet assumption. In this section, we will introduce how to construct either a generalized Dirichlet prior or a Liouville prior for naïve Bayesian classifiers.

The main restrictions of the Dirichlet distribution are the negative-correlation and the equal-confidence requirements. Note that there are two implications about the Laplace's estimate $\alpha_j = 1$ for all j in a Dirichlet prior. The first one is that all variables will have the same prior mean, which means it is a noninformative prior. The Laplace's estimate also implies that the confidence levels about the mean values of the variables are low. To contrast the Dirichlet priors with the Laplace's estimate, we will explore two types of generalized Dirichlet and Liouville priors. The first type will satisfy the two implications of the Laplace's estimate: noninformative and unconfident. For the sake of ease of use, the noninformative implication should not be violated. Thus, the second type of generalized Dirichlet and Liouville priors will be allowed to show a high confidence level about some estimates. Note that noninformative generalized Dirichlet priors can release both requirements of the Dirichlet prior, while noninformative Liouville priors can release only the negative-correlation requirement. In particular, the variables in a Liouville random vector are either all positively or all negatively correlated.

4.1 Noninformative and unconfident priors

The variables in either a univariate or a bivariate generalized Dirichlet random vector cannot be positively correlated. Thus, an attribute with two or three possible outcomes will be assumed to have a Dirichlet prior. Only attributes with more than three possible outcomes will be set to have a generalized Dirichlet or Liouville prior. The number of parameters in a k-variate generalized Dirichlet distribution is 2k. When all mean values

must be considered in constructing a generalized Dirichlet distribution, the remaining degrees of freedom that can be used to adjust the spread of the distribution is k. If the mean values are all equal, we will have $E(\theta_j) = 1/(k+1)$ for $j = 1, 2, \ldots, k+1$, or equivalently $\alpha_j/\beta_j = \alpha_{j+1}/(\alpha_{j+1} + \beta_{j+1})$ for $j = 1, 2, \ldots, k-1$. When $Cov(\theta_2, \theta_3)$ is positive, we have

$$Cov(\theta_2, \theta_3) = E(\theta_2\theta_3) - E(\theta_2)E(\theta_3) > 0 \Rightarrow \frac{\alpha_1}{\beta_1} > \frac{\alpha_1 + \beta_1 + 1}{\alpha_2 + \beta_2}.$$

Note that the Dirichlet distribution $D_k(1, 1, \ldots, 1; 1)$ is equivalent to the generalized Dirichlet distribution $GD_k(1, 1, \ldots, 1; k, k-1, \ldots, 1)$. If we set $\alpha_1 = 1$ and $\beta_1 = k$ for a generalized Dirichlet prior, then $Cov(\theta_2, \theta_3)$ will be positive when $\alpha_2 + \beta_2 > k(k+2)$. The value of $\alpha_2 + \beta_2$ will be large if k is large, and the posterior mean of a variable with an index larger than one is no longer primarily determined by the training data.

All variables in a Dirichlet random vector have the same normalized variance, or equivalently the same confidence level. Since we also want to know the impact of the variables' confidence levels on the performance of naïve Bayesian classifiers, the variables following a generalized Dirichlet distribution will be divided into two sets $\{\theta_1, \theta_2\}$ and $\{\theta_3, \theta_4, \ldots, \theta_{k+1}\}$ such that the variables in the second set will have the same confidence level. This results in $\beta_j = \alpha_{j+1} + \beta_{j+1}$ for $j = 3, 4, \ldots, k$. Following from the above, an attribute with $k+1$ possible outcomes will be assumed to have a generalized Dirichlet prior $GD_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$ in one of the following three groups:

1. Generalized Dirichlet group 1 (GDG1): Prior Gm with parameters $\alpha_1 = 0.01$, $\alpha_j = 1$ for $j = 3, 4, \ldots, k$, $\beta_1 = k/100$, $\beta_j = k - j + 1$ for $j = 3, 4, \ldots, k$, and $\alpha_2 = 0.1 \times (m + 5)$ and $\beta_2 = \alpha_2 \times (k - 1)$ for $m = 1, 2, \ldots, 10$.
2. Generalized Dirichlet group 2 (GDG2): Prior Gm with parameters $\alpha_1 = 0.01$, $\alpha_j = 1.25$ for $j = 3, 4, \ldots, k$, $\beta_1 = k/100$, $\beta_j = 1.25 \times (k - j + 1)$ for $j = 3, 4, \ldots, k$, and $\alpha_2 = 0.1 \times (m-5)$ and $\beta_2 = \alpha_2 \times (k-1)$ for $m = 11, 12, \ldots, 20$.
3. Generalized Dirichlet group 3 (GDG3): Prior Gm with parameters $\alpha_1 = 0.01$, $\alpha_j = 2$ for $j = 3, 4, \ldots, k$, $\beta_1 = k/100$, $\beta_j = 2 \times (k - j + 1)$ for $j = 3, 4, \ldots, k$, and $\alpha_2 = 0.1 \times (m - 15)$ and $\beta_2 = \alpha_2 \times (k - 1)$ for $m = 21, 22, \ldots, 30$.

The values of k and m in prior Gm can be used to derive its covariance matrix. In each prior, since $\alpha_2/\beta_2 = 1/(k-1)$ and $\beta_j = \alpha_{j+1} + \beta_{j+1}$ for $j = 3, 4, \ldots, k$, variables θ_3 through θ_k have the same normalized variance. Note also that $Cov(\theta_j, \theta_{j+1})$ for $j \geq 3$ are identical.

Tables 1 through 3 show the values of $NV(\theta_2)$, $NV(\theta_3)$, $Cov(\theta_2, \theta_3)$, and $Cov(\theta_3, \theta_4)$ for the 9-variate generalized Dirichlet priors of these three groups, respectively. The value of $Cov(\theta_2, \theta_3)$ is gradually changed from negative to positive in each group. This design could help us to know whether the correlations among variables will affect the classification accuracy. The values of $NV(\theta_2)$ and $Cov(\theta_2, \theta_3)$ in the original Dirichlet prior $D_k(1, 1, \ldots, 1; 1)$ for $k = 9$ are 0.090909 and -0.000909, respectively. Note that the values of $NV(\theta_3)$ for the priors in GDG1 are all larger than 0.090909, and that the values of $NV(\theta_3)$ for the priors in GDG3 are all smaller than 0.090909. In GDG2,

Table 1 The values of NV(θ_2), NV(θ_3), Cov(θ_2, θ_3), and Cov(θ_3, θ_4) for GD$_9$(0.01, α_2, 1, 1, . . ., 1; 0.09, β_2, 7, 6, . . ., 1) with various values of α_2 and β_2

	α_2	β_2	NV(θ_2)	NV(θ_3)	Cov(θ_2, θ_3)	Cov(θ_3, θ_4)
G1	0.6	4.8	0.164141	0.110620	−0.000710	−0.000022
G2	0.7	5.6	0.145289	0.110096	−0.000498	−0.000046
G3	0.8	6.4	0.130574	0.109688	−0.000333	−0.000064
G4	0.9	7.2	0.118770	0.109360	−0.000200	−0.000079
G5	1.0	8.0	0.109091	0.109091	−0.000091	−0.000091
G6	1.1	8.8	0.101010	0.108866	0.000000	−0.000101
G7	1.2	9.6	0.094162	0.108676	0.000077	−0.000110
G8	1.3	10.4	0.088284	0.108513	0.000143	−0.000117
G9	1.4	11.2	0.083185	0.108371	0.000201	−0.000123
G10	1.5	12.0	0.078718	0.108247	0.000251	−0.000129

Table 2 The values of NV(θ_2), NV(θ_3), Cov(θ_2, θ_3), and Cov(θ_3, θ_4) for GD$_9$(0.01, α_2, 1.25, 1.25, . . ., 1.25; 0.09, β_2, 8.75, 7.50, . . ., 1.25) with various values of α_2 and β_2

	α_2	β_2	NV(θ_2)	NV(θ_3)	Cov(θ_2, θ_3)	Cov(θ_3, θ_4)
G11	0.6	4.8	0.164141	0.092982	−0.000710	0.000205
G12	0.7	5.6	0.145289	0.092500	−0.000498	0.000181
G13	0.8	6.4	0.130574	0.092124	−0.000333	0.000162
G14	0.9	7.2	0.118770	0.091822	−0.000200	0.000147
G15	1.0	8.0	0.109091	0.091575	−0.000091	0.000134
G16	1.1	8.8	0.101010	0.091368	0.000000	0.000124
G17	1.2	9.6	0.094162	0.091193	0.000077	0.000115
G18	1.3	10.4	0.088284	0.091043	0.000143	0.000108
G19	1.4	11.2	0.083185	0.090912	0.000201	0.000101
G20	1.5	12.0	0.078718	0.090798	0.000251	0.000095

Table 3 The values of NV(θ_2), NV(θ_3), Cov(θ_2, θ_3), and Cov(θ_3, θ_4) for GD$_9$(0.01, α_2, 2, 2, . . ., 2; 0.09, β_2, 14, 12, . . ., 2) with various values of α_2 and β_2

	α_2	β_2	NV(θ_2)	NV(θ_3)	Cov(θ_2, θ_3)	Cov(θ_3, θ_4)
G21	0.6	4.8	0.164141	0.064970	−0.000710	0.000565
G22	0.7	5.6	0.145289	0.064554	−0.000498	0.000540
G23	0.8	6.4	0.130574	0.064229	−0.000333	0.000520
G24	0.9	7.2	0.118770	0.063969	−0.000200	0.000505
G25	1.0	8.0	0.109091	0.063755	−0.000091	0.000492
G26	1.1	8.8	0.101010	0.063577	0.000000	0.000481
G27	1.2	9.6	0.094162	0.063426	0.000077	0.000472
G28	1.3	10.4	0.088284	0.063296	0.000143	0.000464
G29	1.4	11.2	0.083185	0.063184	0.000201	0.000458
G30	1.5	12.0	0.078718	0.063085	0.000251	0.000452

the values of NV(θ_3) are different but all close to 0.090909. This setting could reveal whether the value of the normalized variance does have an impact on the classification accuracy.

As introduced in Sect. 3.2, a Liouville random vector is composed by multiplying a Dirichlet random vector by an independent density generating variate U defined on

[0,1]. For the sake of computational efficiency, assume that U has a beta distribution with parameters γ and ω. As pointed out in Sect. 3.2, the variables with the same mean in a Liouville random vector will have the same normalized variance. So, let the Dirichlet distribution used to generate a noninformative and unconfident Liouville prior for an attribute with k+1 possible outcomes be $D_{k-1}(d, d, \ldots, d; d)$ for some positive constant d, and let $E(U) = k/(k+1) = \gamma/(\gamma+\omega)$. When two variables θ_i and θ_j in the Liouville prior are positively correlated, we will have

$$\alpha > \frac{E(U)^2}{Var(U)} = \frac{\gamma(\gamma+\omega+1)}{\omega} \Rightarrow d > \gamma+\omega+1,$$

where α is the sum of the parameters in the Dirichlet distribution for generating the Liouville distribution. Since both γ and ω are positive, the variables following a Liouville distribution cannot be positively correlated when $d \leq 1$.

When the density generating variate U has a beta distribution with parameters γ and ω, a k-variate noninformative Liouville prior for $k > 2$ is assumed to be in one of the following three groups:

1. Liouville group 1 (LG1): Prior Lm with parameters $\omega = 0.001 \times m$, $\gamma = k \times \omega$, and $d = 1.05$ for $m = 1, 2, \ldots, 10$.
2. Liouville group 2 (LG2): Prior Lm with parameters $\omega = 0.005 \times (m - 10)$, $\gamma = k \times \omega$, and $d = 1.25$ for $m = 11, 12, \ldots, 20$.
3. Liouville group 3 (LG3): Prior Lm with parameters $\omega = 0.02 \times (m - 20)$, $\gamma = k \times \omega$, and $d = 2$ for $m = 21, 22, \ldots, 30$.

Similarly, the values of k and m in prior Lm can be used to derive its covariance matrix. Tables 4 through 6 show the values of $NV(\theta_1)$ and $Cov(\theta_1, \theta_2)$ for the 9-variate Liouville priors with various values of γ and ω. Similar to constructing the noninformative and unconfident generalized Dirichlet priors, $Cov(\theta_1, \theta_2)$ is gradually changed from positive to negative as $\gamma + \omega$ is getting larger in each Liouville group. We also make the normalized variances for the priors in LG1, LG2, and LG3 be larger than, approximately equal to, and smaller than 0.090909.

Table 4 The values of $NV(\theta_1)$ and $Cov(\theta_1, \theta_2)$ for the Liouville priors $L_9(g(u); 1.05, 1.05, \ldots, 1.05)$ with various values of γ and ω

	γ	ω	$NV(\theta_1)$	$Cov(\theta_1, \theta_2)$
L1	0.009	0.001	0.106642	0.000038
L2	0.018	0.002	0.106431	0.000028
L3	0.027	0.003	0.106223	0.000019
L4	0.036	0.004	0.106020	0.000009
L5	0.045	0.005	0.105820	0.000000
L6	0.054	0.006	0.105624	−0.000009
L7	0.063	0.007	0.105432	−0.000018
L8	0.072	0.008	0.105243	−0.000027
L9	0.081	0.009	0.105080	−0.000035
L10	0.090	0.010	0.104877	−0.000043

Table 5 The values of $NV(\theta_1)$ and $Cov(\theta_1, \theta_2)$ for the Liouville priors $L_9(g(u); 1.25, 1.25, \ldots, 1.25)$ with various values of γ and ω

	γ	ω	$NV(\theta_1)$	$Cov(\theta_1, \theta_2)$
L11	0.045	0.005	0.091999	0.000155
L12	0.090	0.010	0.091115	0.000111
L13	0.135	0.015	0.090309	0.000071
L14	0.180	0.020	0.089569	0.000034
L15	0.225	0.025	0.088889	0.000000
L16	0.270	0.030	0.088261	−0.000031
L17	0.315	0.035	0.087680	−0.000060
L18	0.360	0.040	0.087140	−0.000087
L19	0.405	0.045	0.086637	−0.000113
L20	0.450	0.050	0.086168	−0.000136

Table 6 The values of $NV(\theta_1)$ and $Cov(\theta_1, \theta_2)$ for the Liouville priors $L_9(g(u); 2, 2, \ldots, 2)$ with various values of γ and ω

	γ	ω	$NV(\theta_1)$	$Cov(\theta_1, \theta_2)$
L21	0.18	0.02	0.061404	0.000351
L22	0.36	0.04	0.059315	0.000226
L23	0.54	0.06	0.057749	0.000132
L24	0.72	0.08	0.056530	0.000058
L25	0.90	0.10	0.055556	0.000000
L26	1.08	0.12	0.054758	−0.000048
L27	1.26	0.14	0.054094	−0.000088
L28	1.44	0.16	0.053531	−0.000121
L29	1.62	0.18	0.053049	−0.000150
L30	1.80	0.20	0.052632	−0.000175

4.2 Noninformative priors

In a Dirichlet prior, the sum of its parameters represents the total confidence level about this prior. For example, person A with noninformative prior $D_4(1, 1, 1, 1; 1)$ is less confident than person B with noninformative prior $D_4(20, 20, 20, 20; 20)$, because the variables in the former prior have a larger normalized variance. When $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ follows $D_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \alpha_{k+1})$, and $L(\mathbf{y}|\Theta)$ follows a multinomial distribution, by Lemma 3, we have

$$E(\theta_j|\mathbf{y}) = \frac{\alpha_j + y_j}{\alpha + n} = \frac{\alpha}{\alpha + n} \cdot \frac{\alpha_j}{\alpha} + \frac{n}{\alpha + n} \cdot \frac{y_j}{n} = \frac{\alpha}{\alpha + n} E(\theta_j) + \frac{n}{\alpha + n} \cdot \frac{y_j}{n},$$

where $\alpha = \alpha_1 + \alpha_2 + \cdots + \alpha_k$ and $n = y_1 + y_2 + \cdots + y_{k+1}$. This expression means the posterior mean $E(\theta_j|\mathbf{y})$ is the sum of the prior mean $E(\theta_j)$ multiplied by the prior weight $\alpha/(\alpha + n)$ and the data mean y_j/n multiplied by the data weight $n/(\alpha + n)$. So, the value of α represents the equivalent sample size for composing the Dirichlet prior. We will gradually increase the value of α_j from 1 to M with stepsize 1 to search for the best noninformative Dirichlet prior that has the highest classification accuracy. This means that for an attribute with $k + 1$ possible outcomes, the largest equivalent sample size for its noninformative Dirichlet priors will be kM.

The parameters in a generalized Dirichlet prior can also be used to adjust its confidence level. When a generalized Dirichlet prior is noninformative, all variables must have the same mean. So, an attribute with k+1 possible outcomes will be assumed to have a generalized Dirichlet prior $GD_k(\alpha_1, \alpha_2, \ldots, \alpha_k; \beta_1, \beta_2, \ldots, \beta_k)$ in which $\beta_j = (k-j+1) \times \alpha_j$ for $j = 1, 2, \ldots, k$. When all variables have the same mean, a variable with a smaller variance will have a smaller normalized variance. Thus, when all parameters in a noninformative generalized Dirichlet prior remain unchanged except for α_j and β_j, a larger $\alpha_j + \beta_j$ indicates that the confidence level about the estimate of $E(\theta_j)$ is higher.

In a generalized Dirichlet prior, every α_j can be increased independently from 1 to M with stepsize 1 to search for the best generalized Dirichlet prior. In this case, we will need to test M^k noninformative generalized Dirichlet priors to find the one with the highest classification accuracy. This can be tractable only when both M and k are small. For instance, when $M = 50$ and $k = 9$, $M^k = 50^9 \approx 1.95 \times 10^{15}$. Note that $Var(\theta_m)$ is independent of parameters α_j and β_j for all $j > m$. The procedure to find the best noninformative generalized Dirichlet prior was therefore designed as follows. The initial noninformative generalized Dirichlet prior will be $GD_k(1, 1, \ldots, 1; k, k-1, \ldots, 1)$, which is equivalent to $D_k(1, 1, \ldots, 1; 1)$. Then the value of α_1 is gradually increased from 1 to M with stepsize 1 to search for the value of α_1, say α_1^*, that achieves the highest classification accuracy. We then fixed $\alpha_1 = \alpha_1^*$ and $\beta_1 = k\alpha_1^*$ and gradually increased α_2 from 1 to M with stepsize 1 to find the value of α_2 that achieves the highest classification accuracy, and so on. The number of noninformative generalized Dirichlet priors considered in this procedure will be kM instead of M^k.

Since attributes can have different numbers of possible outcomes, we adopted the justifying-right policy to set the values of the parameters in their noninformative generalized Dirichlet priors. That is, when attributes A and B have k and m possible outcomes, respectively, and $m < k$, the value of parameter α_j in the noninformative generalized Dirichlet prior for attribute B is equal to the value of parameter α_{k-m+j} in the noninformative generalized Dirichlet prior for attribute A. For example, if the numbers of possible outcomes of attributes A and B are 10 and 5, respectively, the value of α_j for attribute B is always equal to the value of α_{5+j} for attribute A in searching for the best noninformative generalized Dirichlet prior. Thus, in determining the values of α_1^* through α_5^* for attribute A, the prior for attribute B remains $GD_4(1, 1, 1, 1; 4, 3, 2, 1)$, or equivalently $D_4(1, 1, 1, 1; 1)$.

In constructing a noninformative Liouville prior, let the Dirichlet distribution and the distribution for the density generating variate U be $D_{k-1}(d, d, \ldots, d; d)$ and beta(γ, ω), respectively. Since this Liouville prior is noninformative, we must have $\gamma = k\omega$ to ensure that the k+1 possible outcomes have the same mean probability. The parameters that can be used to adjust its confidence level are d and ω. When the values of d and ω get larger, the confidence level about the Liouville prior becomes higher. Thus, we will gradually and independently increase the values of d and ω from 1 to M with stepsize 1 to search for the best noninformative Liouville prior. In this way, the number of noninformative Liouville priors that will be tested is M^2. As discussed before, when $d > \gamma + \omega + 1 = (k+1)\omega + 1$, the variables in the Liouville distribution will be all positively correlated. If M is not too small, the variables in some of the M^2 noninformative Liouville priors will be positively correlated.

As presented in Sect. 3.2, when U has a beta distribution with parameters $\gamma = kd$ and $\omega = d$, the Liouville distribution will reduce to the Dirichlet distribution $D_k(d, d, \ldots, d; d)$. This means the M Dirichlet priors for searching the best noninformative Dirichlet prior will also be tested in searching for the best noninformative Liouville prior. This guarantees that the classification accuracy resulting from the best noninformative Liouville prior will not be smaller than the classification accuracy resulting from the best noninformative Dirichlet prior. However, except for the Dirichlet prior with the Laplace's estimate, the procedure for searching the best noninformative generalized Dirichlet prior will generally test none of the other $M - 1$ noninformative Dirichlet priors.

When the value of M is large with respect to the size of the training data, the prior will dominate the classification results. Since the priors are noninformative, too large a value of M should deteriorate the classification accuracy. We therefore set M to be 60, which is equivalent to a sample with 540 instances when an attribute has 10 possible outcomes.

5 Experimental results

We chose 18 data sets, as shown in Table 7, from the UCI machine learning repository (Blake and Merz 1998) to study the impact of the Dirichlet assumption in naïve Bayesian classifiers. In general, there are two ways to handle continuous attributes for the naïve Bayesian classifier: discretize the attributes or estimate normal distributions for them. Empirical studies (Dougherty et al. 1995; Kohavi and Sahami 1996) have shown that the discretizing approach can perform similar to or better than the other approach. As addressed by Hsu et al. (2003) for naïve Bayesian classifiers, different discretization methods used on continuous attributes will achieve similar classification accuracies. We therefore employed the ten-bin discretization method to discretize continuous attributes; i.e., the range of a continuous attribute was divided into ten equal-width intervals. Missing values were ignored in calculating classification probabilities. Stratified five-fold cross validation was the method used for evaluating classification accuracy.

The order of the variables in a Dirichlet random vector is arbitrary. When an attribute has a k-variate Liouville prior that does not reduce to a Dirichlet distribution, only the position of the variable corresponding to possible outcome k+1 of the attribute cannot be changed arbitrarily. However, in a generalized Dirichlet random vector, two consecutive variables θ_j and θ_{j+1} can be interchanged only when $\beta_j = \alpha_{j+1} + \beta_{j+1}$. Thus, when an attribute has a generalized Dirichlet prior, the way to assign the order of its possible outcomes is as follows. The possible outcomes of a discretized continuous attribute or a discrete attribute measured by an ordinal scale will be sorted in an ascending order. When an attribute in a data set is nominal, the order of its possible outcomes will be the same as the order they appeared in the description document for the data set.

There are two types of generalized Dirichlet and Liouville priors that will be tested on the 18 data sets. One is similar to the Dirichlet prior with the Laplace's estimate: noninformative and unconfident, and another is allowed to show high confidence levels

Table 7 The characteristics of testing data sets

Data set	No. of instances	No. of attributes	No. of classes
Australian	690	14	2
Balance-scale	625	4	3
Breast-w	699	9	2
Cleve	303	13	2
Ecoli	336	8	8
Glass	214	10	7
Haberman	306	3	2
Heart	270	13	2
Hepatitis	155	19	2
Iris	150	4	3
Liver	345	6	2
New-thyroid	215	5	3
Pima	768	8	2
Segment	2310	19	7
Tae	151	5	3
Vehicle	946	18	4
Wine	178	13	3
Yeast	1484	8	10

about estimates. In this section, we will present the experimental results of these two types of priors. The noninformative generalized Dirichlet distribution can release both the negative-correlation and the equal-confidence requirements, and the noninformative Liouville distribution allows variables to be either all positively or all negatively correlated. They are different extensions of the Dirichlet distribution. Thus, the experimental results can provide the necessary information for us to study the impact of the Dirichlet assumption in naïve Bayesian classifiers.

5.1 Noninformative and unconfident priors

The parameters in the noninformative and unconfident generalized Dirichlet and Liouville priors constructed in Sect. 4.1 are all between 0.01 and 2. So, we first set the value of α_j in the noninformative Dirichlet prior to be various values in this range to identify its impact on the performance of the naïve Bayesian classifier. The testing results for $\alpha_j = 0.01, 0.5, 1, 1.5$, and 2 are shown in Table 8. The last row of Table 8 shows the p-values of the paired-t tests for the other four settings of α_j with respect to the Laplace's estimate. Since the p-values are all larger than 0.05, we adopted the classification accuracy obtained from the Dirichlet prior with the Laplace's estimate to be the baseline for evaluating the impact of the noninformative and unconfident generalized Dirichlet and Liouville priors.

As described in Sect. 4.1, an attribute with either two or three possible outcomes will be assumed to have a Dirichlet prior for evaluating the estimate of $p(x_i|c_j)$. For any attribute with $k+1 > 3$ possible outcomes, its generalized Dirichlet and Liouville priors will be one of G1 through G30 and one of L1 through L30, respectively. The

Table 8 The testing results of various noninformative and unconfident Dirichlet priors

Data set	$\alpha_j = 0.01$	$\alpha_j = 0.5$	$\alpha_j = 1$	$\alpha_j = 1.5$	$\alpha_j = 2$
Australian	83.85	84.34	84.19	84.32	84.62
Balance-scale	90.79	90.58	90.58	90.58	90.58
Breast-w	97.43	97.45	97.45	97.45	97.31
Cleve	82.39	83.35	83.00	83.00	83.31
Ecoli	82.09	83.72	82.50	81.30	79.69
Glass	60.98	59.21	59.35	60.10	57.66
Haberman	75.23	75.88	75.54	75.50	74.78
Heart	81.77	83.59	84.21	83.86	83.86
Hepatitis	88.01	86.21	86.21	86.21	86.21
Iris	93.81	93.81	93.19	92.56	92.56
Liver	61.52	62.71	63.27	63.60	63.28
New-thyroid	94.13	93.21	91.86	91.86	91.42
Pima	75.34	75.44	76.01	76.22	76.04
Segment	90.43	89.56	88.93	88.11	87.69
Tae	53.18	53.31	52.63	53.28	52.63
Vehicle	61.96	61.78	61.82	61.76	61.51
Wine	95.65	96.70	96.14	96.70	97.18
Yeast	58.13	57.97	57.58	57.60	57.47
p-value	0.6672	0.0824		0.8345	0.0860

classification results for these generalized Dirichlet and Liouville priors are summarized in the 18 line charts in Fig. 1.

In each line chart, the baseline is the classification accuracy of the Dirichlet prior with the Laplace's estimate, and the horizontal axis represents the generalized Dirichlet and the Liouville priors for the naïve Bayesian classifier. For instance, the horizontal label 10 means the generalized Dirichlet and the Liouville priors for the naïve Bayesian classifier are G10 and L10, respectively. From the 18 charts, we can see that most values in the line curves corresponding to the noninformative and unconfident generalized Dirichlet priors are above or on the baselines, while most values in the line curves corresponding to the noninformative and unconfident Liouville priors are below the baselines. This phenomenon suggests that the generalized Dirichlet prior can often achieve a better or equal accuracy, and that the accuracy for the Liouville prior is usually worse. This also implies that among the three possible prior families for the naïve Bayesian classifier, the generalized Dirichlet distribution should be the best choice. However, the differences between their resulting accuracies are small, because the priors are all noninformative and unconfident.

Tables 9 and 10 summarize the best and the worst mean accuracies resulting from the generalized Dirichlet and the Liouville prior groups for various data sets, where "+" ("−") indicates that a mean accuracy is larger (smaller) than the mean accuracy resulting from the corresponding Dirichlet prior. The 2-tuples in the last row of both tables show the numbers of plus and minus signs in each column. The first two of the three generalized Dirichlet prior groups, even the worst case, have larger values for the plus sign. On the other hand, in the three Liouville prior groups, every value for the plus sign is smaller than the corresponding value for the minus sign. This again suggests that the generalized Dirichlet prior is the best among the three prior families. Note

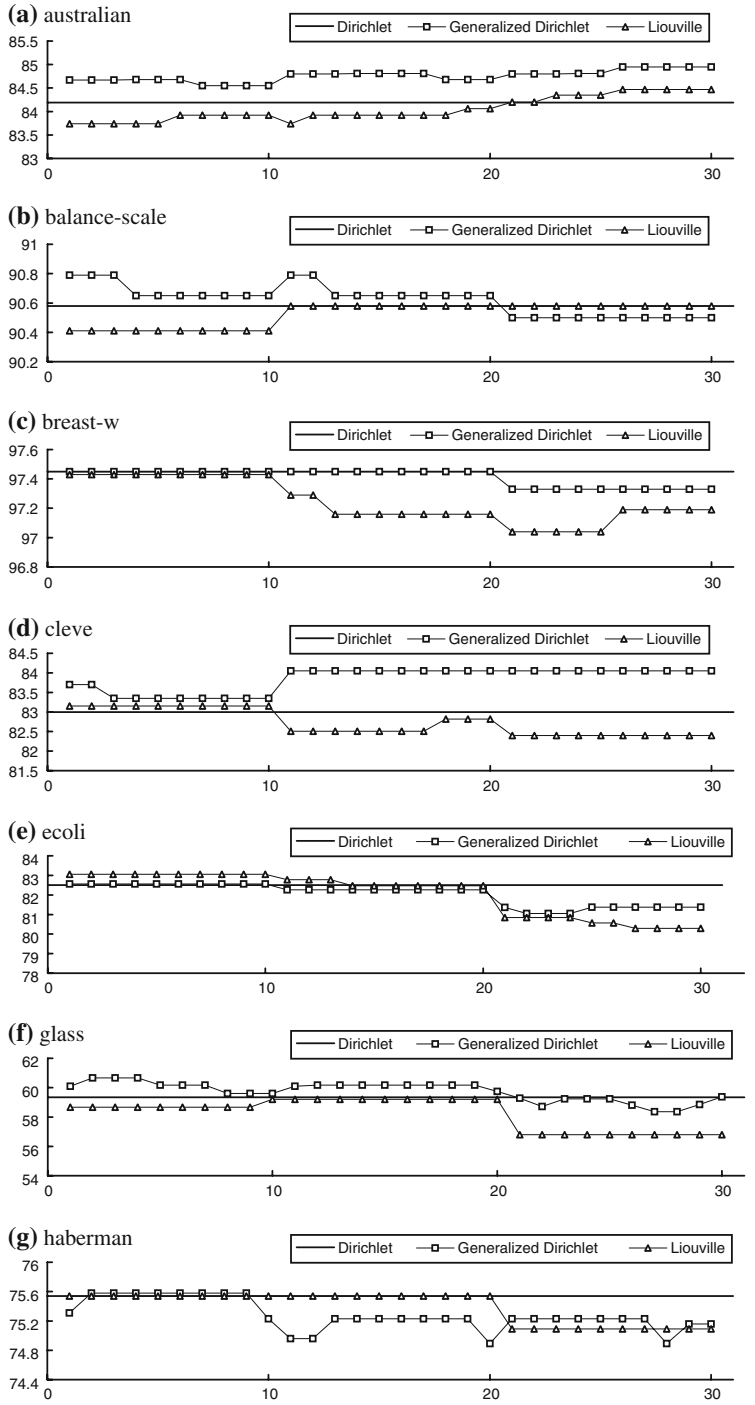

Fig. 1 The testing results for various noninformative and unconfident generalized Dirichlet and Liouville priors specified in Sect. 4.1

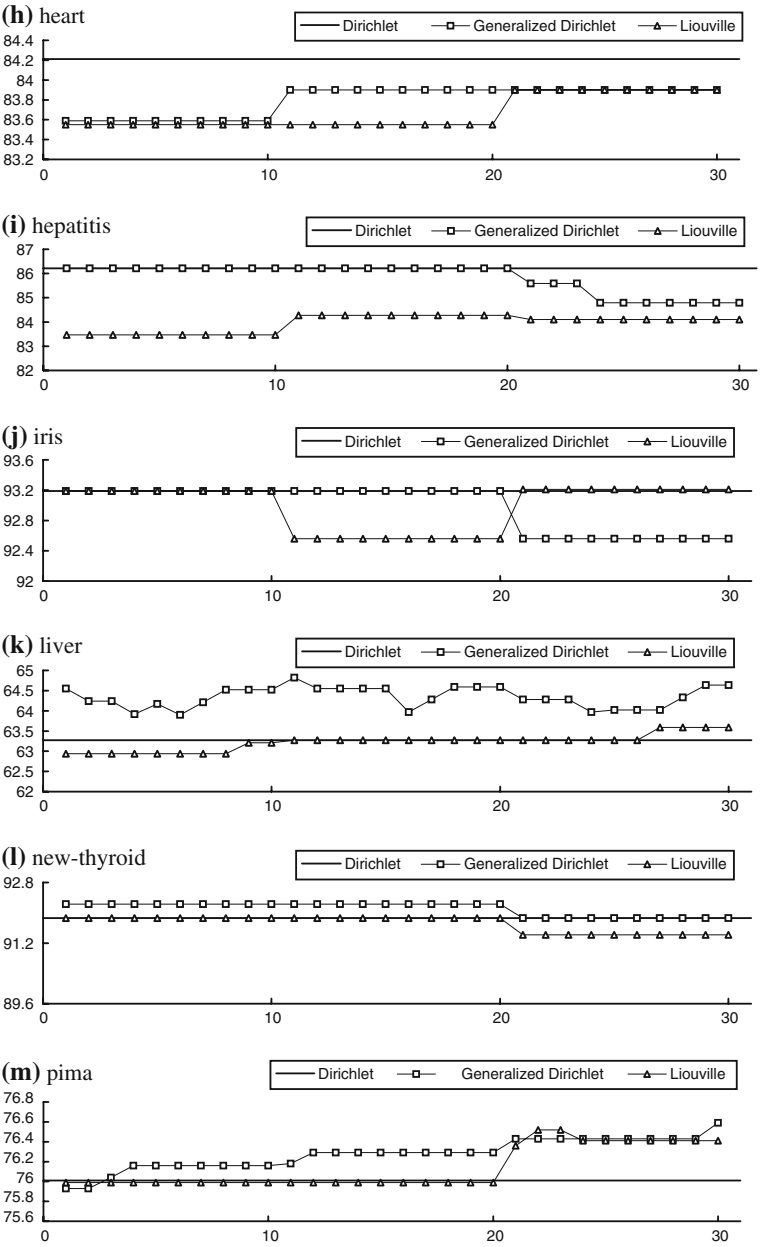

Fig. 1 continued

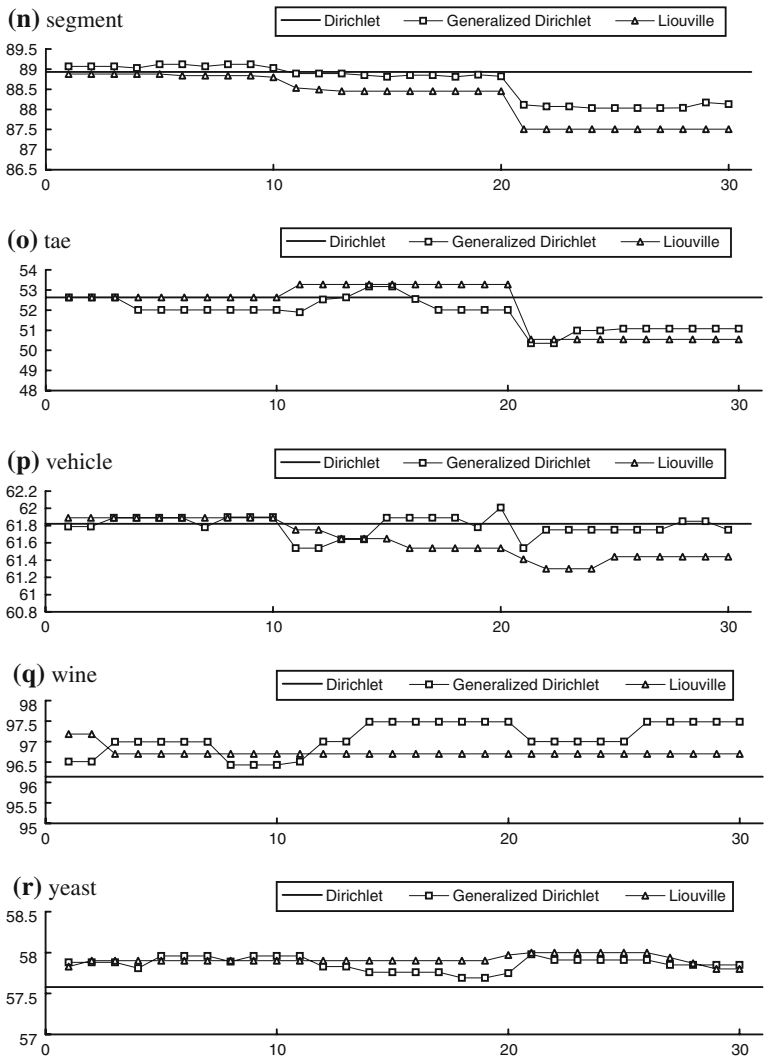

Fig. 1 continued

also that the values for the minus sign in the best or worst columns gradually increase for the three groups in both tables. The priors in GDG3 have the smallest normalized variances among the three generalized Dirichlet prior groups, and similarly for LG3. One reason for this phenomenon could be when priors are noninformative and unconfident, smaller normalized variances will make the values for parameters larger. When the number of instances in a data set is small, such as the data sets "ecoli", "heart", and "tae", improper and larger values for parameters may cause some misclassifications that greatly reduce its classification accuracy.

Table 9 The best and the worst mean accuracies resulting from the generalized Dirichlet priors

Data set	Dirichlet	GDG1		GDG2		GDG3	
		Best	Worst	Best	Worst	Best	Worst
Australian	84.19	84.68^{+}	84.55^{+}	84.81^{+}	84.68^{+}	84.95^{+}	84.80^{+}
Balance-scale	90.58	90.79^{+}	90.65^{+}	90.79^{+}	90.65^{+}	90.50^{-}	90.50^{-}
Breast-w	97.45	97.45	97.45	97.45	97.45	97.33^{-}	97.33^{-}
Cleve	83.00	83.59^{+}	83.59^{+}	84.05^{+}	84.05^{+}	84.05^{+}	84.05^{+}
Ecoli	82.50	82.56^{+}	82.56^{+}	82.26^{-}	82.26^{-}	81.37^{-}	81.05^{-}
Glass	59.35	60.67^{+}	59.61^{+}	60.18^{+}	59.75^{+}	59.38^{+}	58.36^{-}
Haberman	75.54	75.58^{+}	75.23^{-}	75.23^{-}	74.89^{-}	75.23^{-}	74.89^{-}
Heart	84.21	83.59^{-}	83.59^{-}	83.90^{-}	83.90^{-}	83.90^{-}	83.90^{-}
Hepatitis	86.21	86.21	86.21	86.21	86.21	85.59^{-}	84.79^{-}
Iris	93.19	93.19	93.19	93.19	93.19	92.56^{-}	92.56^{-}
Liver	63.27	64.55^{+}	63.90^{+}	64.82^{+}	63.97^{+}	64.64^{+}	63.97^{+}
New-thyroid	91.86	92.23^{+}	92.23^{+}	92.23^{+}	92.23^{+}	91.86	91.86
Pima	76.01	76.16^{+}	75.93^{-}	76.29^{+}	76.18^{+}	76.59^{+}	76.43^{+}
Segment	88.93	89.12^{+}	89.03^{+}	88.89^{-}	88.81^{-}	88.17^{-}	88.03^{-}
Tae	52.63	52.63	52.01^{-}	53.17^{+}	51.89^{-}	51.07^{-}	50.35^{-}
Vehicle	61.82	61.90^{+}	61.78^{-}	62.01^{+}	61.54^{-}	61.85^{+}	61.54^{-}
Wine	96.14	96.99^{+}	96.43^{+}	97.48^{+}	96.51^{+}	97.48^{+}	97.00^{+}
Yeast	57.58	57.96^{+}	57.81^{+}	57.96^{+}	57.69^{+}	57.98^{+}	57.85^{+}
No. of data sets		(13, 1)	(10, 5)	(11, 4)	(9, 6)	(8, 9)	(6, 11)

Table 10 The best and the worst mean accuracies resulting from the Liouville priors

Data set	Dirichlet	LG1		LG2		LG3	
		Best	Worst	Best	Worst	Best	Worst
Australian	84.19	83.92^{-}	83.74^{-}	84.06^{-}	83.74^{-}	84.47^{+}	84.20^{+}
Balance-scale	90.58	90.41^{-}	90.41^{-}	90.58	90.58	90.58	90.58
Breast-w	97.45	97.43^{-}	97.43^{-}	97.29^{-}	97.16^{-}	97.19^{-}	97.04^{-}
Cleve	83.00	83.15^{+}	83.15^{+}	82.82^{-}	82.51^{-}	82.40^{-}	82.40^{-}
Ecoli	82.50	83.06^{+}	83.06^{+}	82.78^{+}	82.47^{-}	80.85^{-}	80.29^{-}
Glass	59.35	59.20^{-}	58.67^{-}	59.20^{-}	59.20^{-}	56.80^{-}	56.80^{-}
Haberman	75.54	75.54	75.54	75.54	75.54	75.09^{-}	75.09^{-}
Heart	84.21	83.55^{-}	83.55^{-}	83.55^{-}	83.55^{-}	83.90^{-}	83.90^{-}
Hepatitis	86.21	83.47^{-}	83.47^{-}	84.10^{-}	84.10^{-}	84.10^{-}	84.10^{-}
Iris	93.19	93.19	93.19	92.56^{-}	92.56^{-}	93.21^{+}	93.21^{+}
Liver	63.27	63.21^{-}	62.94^{-}	63.27	63.27	63.59^{+}	63.27
New-thyroid	91.86	91.86	91.86	91.86	91.86	91.42^{-}	91.42^{-}
Pima	76.01	75.99^{-}	75.99^{-}	75.99^{-}	75.99^{-}	76.52^{+}	76.36^{+}
Segment	88.93	88.88^{-}	88.80^{-}	88.53^{-}	88.45^{-}	87.51^{-}	87.51^{-}
Tae	52.63	52.63	52.63	53.28^{+}	53.28^{+}	50.55^{-}	50.55^{-}
Vehicle	61.82	61.89^{+}	61.89^{+}	61.75^{-}	61.54^{-}	61.44^{-}	61.30^{-}
Wine	96.14	97.18^{+}	96.70^{+}	96.70^{+}	96.70^{+}	96.70^{+}	96.70^{+}
Yeast	57.58	57.90^{+}	57.83^{+}	57.97^{+}	57.90^{+}	58.00^{+}	57.80^{+}
No. of data sets		(5, 9)	(5, 9)	(4, 10)	(3, 11)	(6, 11)	(5, 11)

Table 11 The percentage of instances classified differently by the naïve Bayesian classifiers with noninformative and unconfident generalized Dirichlet and Liouville priors

Data set (No. of instances)	GDG1 %	GDG2 %	GDG3 %	LG1 %	LG2 %	LG3 %
Australian (690)	0.70	0.83	0.86	0.36	0.28	0.39
Balance-scale (625)	0.59	0.61	0.80	0.16	0.32	0.32
Breast-w (699)	0.14	0.14	0.14	0.29	0.31	0.36
Cleve (303)	0.40	0.99	0.99	0.99	1.55	2.31
Ecoli (336)	0.89	0.86	3.27	1.19	1.28	3.81
Glass (214)	8.32	8.55	9.67	3.69	1.87	6.78
Haberman (306)	0.65	1.01	1.47	0.33	0.33	0.98
Heart (270)	0.74	0.37	0.37	0.74	0.74	1.11
Hepatitis (155)	0.65	0.65	1.10	5.16	4.52	4.52
Iris (150)	0	0	0.47	0	0.67	1.33
Liver (345)	2.96	3.07	4.29	0.81	1.16	3.01
New-thyroid (215)	0.93	0.93	0.47	0.47	0.47	0.47
Pima (768)	0.98	0.79	1.16	0.52	0.26	1.54
Segment (2,310)	1.03	1.31	2.33	0.23	0.53	1.65
Tae (151)	1.32	2.19	7.75	0.66	1.32	7.28
Vehicle (946)	0.58	0.94	1.34	0.39	0.30	0.88
Wine (178)	1.97	2.58	2.53	0.67	0.56	0.56
Yeast (1,484)	1.31	1.03	1.54	0.62	0.58	1.72
Average	1.34	1.51	2.25	0.96	0.95	2.17

According to the above analysis of the 18 data sets, when noninformative and unconfident priors are of the same family, the settings on the normalized variances of variables and the correlations among variables have a slight impact on the accuracy of the naïve Bayesian classifier. However, the performance of the generalized Dirichlet distribution is the best among the three distribution families, and the performance of the Liouville distribution is the worst.

We summarize the percentage of instances classified differently by the naïve Bayesian classifiers with noninformative and unconfident generalized Dirichlet and Liouville priors with respect to the naïve Bayesian classifiers that have Dirichlet priors with the Laplace's estimate in Table 11. On average, there are at most 2.3 percent of the instances that are predicted differently. The data set "glass" has only 214 instances and low classification accuracy, and the number of classes in this data set is 7. Thus, different priors have a relatively large impact on the accuracy of this data set.

5.2 Noninformative priors

In searching for the best noninformative Dirichlet prior, the value of α_j is gradually increased from 1 to 60 with stepsize 1. When a data set has continuous attributes, after the process of the ten-bin discretization, the equivalent sample size corresponding to a discretized continuous attribute can be as large as 540. The 18 data sets all have continuous attributes, except the data set "balance-scale". An attribute in the data set "tae" has 26 possible outcomes, hence the equivalent sample for this data set can be as large as 1,500. We first examine whether a value larger than 60 for the α_j is necessary to achieve a higher classification accuracy. The classification accuracies for various

values of α_j of the noninformative Dirichlet priors are depicted in Fig. 2. We can see that when α_j is between 30 and 60, all curves gradually decline except the data set "balance-scale". This data set has four discrete attributes, and every attribute has five possible outcomes. Every possible outcome combination of the four attributes appears in this data set, hence it has 625 instances. The value of α_j has a tiny impact on the classification accuracy of this data set, even when the value of α_j is 500. As shown in Table 12, the best noninformative Dirichlet prior for this data set is the Dirichlet distributions with the Laplace's estimate. We therefore conclude that the value of α_j in the best noninformative Dirichlet priors for these 18 data sets will not be larger than 60.

Table 12 also shows the best noninformative Liouville priors and the comparison with respect to the best noninformative Dirichlet priors, where the second, third, and fourth columns of the best noninformative Liouville prior represent the percentage of instances classified differently, the values of parameters d and ω, and the covariance between any pair of variables, respectively. As pointed out in Sect. 4.2, we will test 3,600 Liouville priors, which include the 60 Dirichlet priors for searching the best noninformative Dirichlet prior, to find the best noninformative Liouville prior for each data set. It is therefore not surprising that for every data set the accuracy resulting from the best noninformative Liouville prior is larger than or equal to the accuracy resulting from the best noninformative Dirichlet prior, but the differences are all smaller 0.011. An interesting finding is that only one of the 18 covariances is positive and close to zero. This suggests that to achieve a higher accuracy, setting all variables to be negatively correlated is generally better than setting all variables to be positively correlated. Since the sum of the variables in a multivariate distribution defined on the unit simplex has an upper bound one, for realistic cases, it is unlikely that all variables can be significantly positively correlated.

The best noninformative generalized Dirichlet priors are summarized in Table 13. We only show the values of parameters α_1 through α_k, because $\beta_j = (k-j+1) \times \alpha_j$ for $j = 1, 2, \ldots, k$. As stated in Sect. 4.2, the noninformative generalized Dirichlet priors tested for a data set generally include only one of the 60 noninformative Dirichlet priors for searching the best Dirichlet prior. Thus, it is possible that for a given data set, the accuracy resulting from the best noninformative generalized Dirichlet prior is smaller than the accuracy resulting from the best noninformative Dirichlet prior. However, the best noninformative generalized Dirichlet priors outperform the best noninformative Dirichlet priors and the best noninformative Liouville priors in 17 and 16 of the 18 data sets, respectively. The three best noninformative priors have the same performance only in the data set "wine", because the accuracy of this data set is very close to one and the number of instances in this data set is less than 200.

The NV_{max}, NV_{min}, Cov_{max}, and Cov_{min} in Table 13 represent the maximal normalized variance, the minimal normalized variance, the maximal covariance, and the minimal covariance resulting from the best noninformative generalized Dirichlet prior, respectively. An interesting result is that 16 of the 18 best noninformative generalized Dirichlet priors include positive correlations among variables, as indicated by the data given in column Cov_{max}. Thus, allowing some variables in a prior for the naïve Bayesian classifier to be positively correlated can be beneficial. Note also that the range of NV_{max}/NV_{min} is between 4 and 32. This demonstrates that a reasonable prior for the naïve Bayesian classifier should exhibit different confidence levels about the estimates.

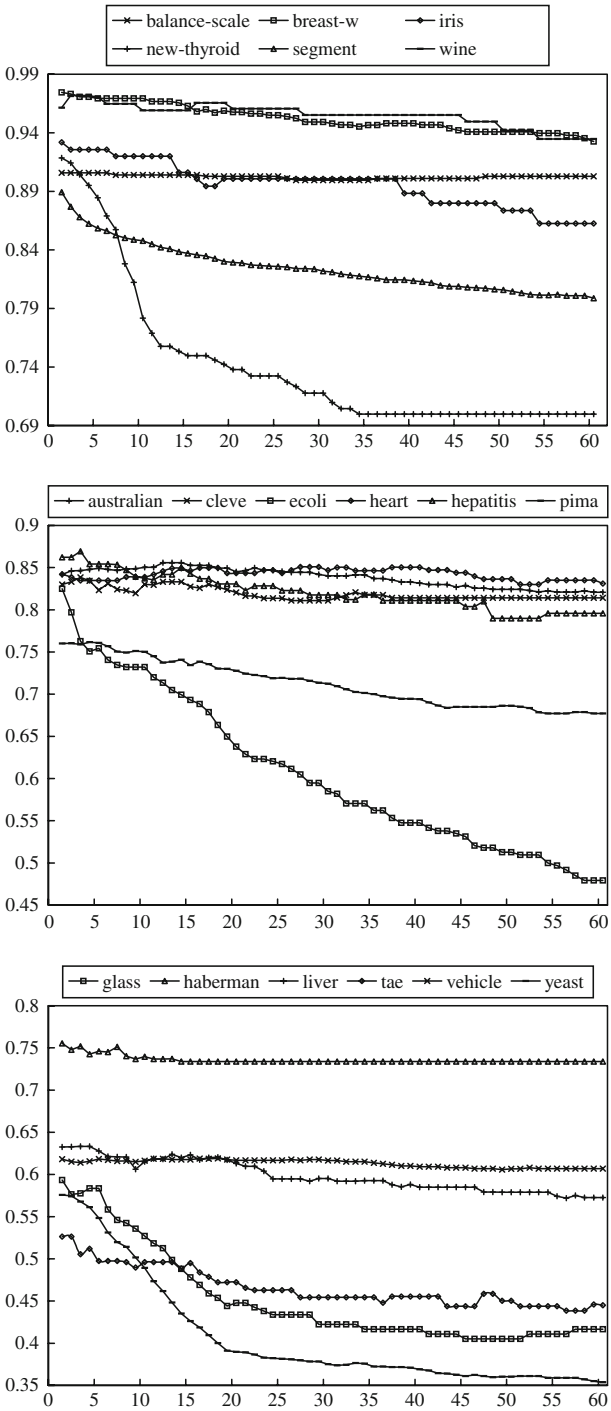

Fig. 2 The classification accuracies for various values of α_j of the noninformative Dirichlet priors

Table 12 The best noninformative Dirichlet priors and the best noninformative Liouville priors

Data set (No. of instances)	Best Dirichlet prior		Best Liouville prior			
	Accuracy	α_j	Accuracy	%	(d, ω)	Covariance
Australian (690)	85.58	13	85.80	1.45	(21, 7)	−0.000015
Balance-scale (625)	90.58	1	90.77	0.48	(2, 3)	−0.003889
Breast-w (699)	97.45	1	97.45	0	(1, 1)	−0.000909
Cleve (303)	83.85	3	83.96	9.24	(56, 5)	0.000002
Ecoli (336)	82.50	1	82.50	0	(1, 1)	−0.000909
Glass (214)	59.35	1	59.64	2.80	(1, 4)	−0.000976
Haberman (306)	75.54	1	75.84	1.63	(3, 1)	−0.000260
Heart (270)	85.08	27	85.37	1.11	(26, 31)	−0.000039
Hepatitis (155)	86.93	3	86.93	0	(3, 3)	−0.000323
Iris (150)	93.19	1	93.81	0.67	(1, 2)	−0.000952
Liver (345)	63.33	4	64.38	2.32	(3, 1)	−0.000260
New-thyroid (215)	91.86	1	92.23	0.47	(1, 23)	−0.000996
Pima (768)	76.16	4	76.33	0.65	(4, 1)	−0.000172
Segment (2,310)	88.93	1	88.93	0	(1, 1)	−0.000909
Tae (151)	52.63	1	52.63	0	(1, 1)	−0.000055
Vehicle (946)	61.90	18	62.13	0.42	(17, 10)	−0.000054
Wine (178)	97.18	2	97.18	0	(1, 34)	−0.000997
Yeast (1,484)	57.58	1	57.74	2.83	(2, 1)	−0.000431

Table 13 The best noninformative generalized Dirichlet priors

Data set	Accuracy	α_1 through α_k	%	NV_{max}	NV_{min}	Cov_{max}	Cov_{min}
Australian	86.17	4, 1, 1, 32, 26, 4, 28, 1, 24, 13, 1, 1, 1	2.32	0.066127	0.003314	0.000169	−0.000834
Balance-scale	90.97	1, 2, 8, 1	0.96	0.166667	0.041667	0.001481	−0.006667
Breast-w	97.57	1, 1, 1, 15, 1, 1, 1, 1, 1	0.14	0.090909	0.010863	0.000292	−0.001081
Cleve	84.01	1, 3, 15, 1, 1, 1, 1, 1, 13	2.64	0.090909	0.008145	0.000062	−0.001112
Ecoli	83.16	7, 1, 1, 1, 6, 1, 1, 1, 1	4.46	0.089202	0.014085	0.000232	−0.001189
Glass	62.71	1, 44, 1, 1, 1, 1, 4, 9, 2	7.01	0.090909	0.003384	0.001038	−0.001018
Haberman	76.40	5, 2, 1, 1, 3, 7, 4, 1, 1	2.94	0.088152	0.019401	0.000278	−0.001860
Heart	85.92	27, 24, 1, 32, 3, 1, 1, 1, 3	5.56	0.086615	0.003690	0.000124	−0.001430
Hepatitis	88.18	7, 3, 1, 1, 1, 1, 1, 59, 1	1.29	0.087612	0.014085	0.001856	−0.001057
Iris	96.68	1, 1, 1, 4, 17, 1, 1, 1, 1	3.33	0.090909	0.010629	0.000348	−0.001275
Liver	69.26	1, 28, 4, 29, 1, 42, 1, 48, 1	13.33	0.090909	0.004671	0.001081	−0.001631
New-thyroid	92.23	1, 5, 1, 1, 1, 1, 1, 1, 1	0.47	0.090909	0.020641	−0.000119	−0.000997
Pima	77.41	16, 1, 3, 1, 1, 1, 1, 8, 4	2.47	0.089027	0.006211	0.001413	−0.001084
Segment	89.18	6, 1, 1, 1, 1, 1, 1, 37, 4	2.81	0.089253	0.016393	0.001915	−0.000984
Tae	59.24	5, 1, 2, 1, 1, 2, 1, 1, 2, 1, 3, 3, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1	11.26	0.036947	0.007634	0.000018	−0.000077
Vehicle	64.79	1, 55, 58, 45, 5, 31, 1, 1, 1	9.09	0.090909	0.002841	0.000112	−0.001845
Wine	97.18	1, 2, 1, 23, 1, 1, 1, 1, 1	1.12	0.090909	0.007904	0.000265	−0.001138
Yeast	58.01	1, 1, 4, 1, 2, 1, 1, 1, 1	3.71	0.090909	0.026630	−0.000083	−0.001128

5.3 Discussion

When priors are noninformative and unconfident, the generalized Dirichlet distribution can generally perform at least as good as the Dirichlet distribution, while the Liouville distribution usually results in a lower classification accuracy with respect to the Dirichlet distribution. The reason for this could be that forcing the variables with the unit-sum constraint to be all positively correlated and have the same confidence level is inappropriate for most real data sets. The variables in a noninformative Liouville random vector must have the same normalized variance and be either all positively or all negatively correlated. That is why only one of the 18 best noninformative Liouville priors given in Table 12 allows variables to be all positively correlated. Note, however, that the value of that positive covariance is very close to zero.

When all priors are noninformative and unconfident, the accuracies resulting from the Dirichlet, the generalized Dirichlet, and the Liouville distributions are not significantly different. In many of these cases, even if the generalized Dirichlet distribution outperforms the Dirichlet and the Liouville distributions, we are not totally sure that releasing the two requirements of the Dirichlet assumption can assist in increasing the performance of the naïve Bayesian classifier. Thus, we attempted to increase the confidence levels about priors to search for the best noninformative Dirichlet, generalized Dirichlet, and Liouville priors. The experimental results again show that the generalized Dirichlet distribution is the best, and unlike the Liouville distribution, the generalized Dirichlet distribution greatly enhances the classification accuracy of the naïve Bayesian classifier in several data sets.

Setting all variables to be positively correlated can be useless to the performance of the naïve Bayesian classifier, as demonstrated by the noninformative Liouville priors. However, allowing some variables to be positively correlated can be beneficial to the operations of the naïve Bayesian classifier. Almost all of the best noninformative generalized Dirichlet priors given in Table 13 include some, but not all, positively correlated variables. Since an appropriate prior for a naïve Bayesian classifier must be defined on the unit simplex, an observation for some possible outcome j will increase the occurrence probability of this outcome and generally decrease the occurrence probabilities of some, but not all, other possible outcomes. Thus, not only is the negative-correlation requirement of the Dirichlet assumption inappropriate, but so is the requirement of totally positive correlation.

An observation can be an occurrence of either outcome j or some other possible outcome. From the viewpoint of outcome j, an observation is the result of a Bernoulli test. Let p_j be the probability for an observation to be outcome j, and let $\bar{p}_j$ be the estimate of p_j from a sample with n observations. When $np_j \geq 5$ and $n(1 - p_j) \geq 5$, by the central limit theorem, the sampling distribution of $\bar{p}_j$ is approximately a normal distribution with mean p_j and variance $p_j(1 - p_j)/n$ (Anderson et al. 2006). In this case, we have $NV(\bar{p}_j) = 1/n$ that does not depend on the index j. All estimates $\bar{p}_j$ therefore have the same normalized variance when the observations are independent and data size n is large. Note, however, that the minimal value of n for satisfying the two inequalities $np_j \geq 5$ and $n(1 - p_j) \geq 5$ will be different for different values of p_j. This implies that assuming the occurrence probabilities of all possible outcomes to have the same confidence level is not

reasonable. Our experimental results show that the naïve Bayesian classifiers with noninformative generalized Dirichlet priors usually achieve better prediction accuracies, and that all of the best noninformative generalized Dirichlet random vectors allow variables to have different normalized variances. These could be evidence of the inappropriateness of the equal-confidence requirement of the Dirichlet assumption.

6 Conclusions

The Dirichlet assumption is essential to the operation of naïve Bayesian classifiers. This assumption implies two requirements about variables: negative correlation and an equal confidence level measured by the normalized variance. In this article, we proposed two different multivariate distributions, generalized Dirichlet and Liouville distributions, defined on the unit simplex as the priors of naïve Bayesian classifiers to investigate the impact of the Dirichlet assumption. The noninformative generalized Dirichlet prior can release both requirements, while the noninformative Liouville prior can release only the negative-correlation requirement.

A Dirichlet prior with the Laplace's estimate has two implications: noninformative and unconfident. To test the appropriateness of the Dirichlet assumption, we first presented the ways to construct generalized Dirichlet and Liouville priors that are also noninformative and unconfident for naïve Bayesian classifiers. When the priors are all unconfident, the resulting classification accuracies can only be slightly different. We therefore introduced a way to study the performance of the naïve Bayesian classifier when the priors are noninformative but can have high confidence levels about some estimates.

With respect to the Dirichlet priors with the Laplace's estimate, our experimental results on 18 real data sets show that, on average, at most 2.3% of the instances are classified differently when naïve Bayesian classifiers have noninformative and unconfident generalized Dirichlet or Liouville priors. Since all priors are still unconfident, the prediction accuracies for various prior families are similar. In this case, the prediction accuracy resulting from a generalized Dirichlet prior is usually better, while a Liouville prior generally results in a relatively lower prediction accuracy. When priors allow high confidence levels about some estimates, the generalized Dirichlet distribution is still the best among the three distribution families, and the percentage of instances classified differently by the best noninformative Dirichlet priors and the best noninformative generalized Dirichlet priors can be larger than 10%. The best noninformative generalized Dirichlet priors indicate that a reasonable prior for the naïve Bayesian classifier should allow variables to be positively correlated and have different confidence levels about their mean values. Thus, not only is the Dirichlet assumption in naïve Bayesian classifiers inappropriate, but also forcing the variables to be all positively correlated can generally deteriorate the classification accuracy of such classifiers. Multivariate distributions that possess the conjugate property and allow different confidence levels and alternate correlation relations for variables should be more suitable priors.

Since generalized Dirichlet priors usually result in higher classification accuracies, it should be of interest to investigate how to derive the most appropriate generalized Dirichlet prior from data to improve the performance of the naïve Bayesian classifier.

An efficient way to determine whether more complicated priors are necessary for a data set can be the focus of future research.

The computational complexity of the approach proposed in this study for searching the best noninformative generalized Dirichlet prior is linearly proportional to the maximum of the numbers of possible outcomes of the attributes in a data set. When the maximum is not large, our approach can be executed efficiently. However, in applying the naïve Bayesian classifier to text classification, the number of variables or words in a generalized Dirichlet prior can be tens of thousands. This makes our approach time-consuming in searching for the best noninformative generalized Dirichlet prior. A more efficient way for this kind of search should be developed to enhance the applicability of the naïve Bayesian classifiers for text classification.

References

Aitchison J (1986) The statistical analysis of compositional data. John Wiley, New York

Anderson DR, Sweeney DJ, Williams TA, Chen JC (2006) Statistics for business and economics: a practical approach, Chap. 7, Thomson Learning

Bier VM, Yi W (1995) A Bayesian method for analyzing dependencies in precursor data. Int J Forecast 11:25–41

Blake C, Merz C (1998) UCI machine learning repository, http://www.ics.uci.edu/~mlearn/MLRepository.html.

Cestnik B, Bratko I (1991) On estimating probabilities in tree pruning. Machine Learning–EWSL-91, European Working Session on Learning. Springer-Verlag, Berlin, Germany, pp 138–150

Connor RJ, Mosimann JE (1969) Concepts of independence for proportions with a generalization of the Dirichlet distribution. J Am Stat Assoc 64:194–206

Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 29:103–130

Dougherty J, Kohavi R, Sahami M (1995) Supervised and unsupervised discretization of continuous features. In: Proceedings of the 12th international conference on machine learning. Morgan Kaufmann, San Francisco, CA, pp 194–202

Fang KT, Kotz S, Ng KW (1990) Symmetric multivariate and related distributions. Chapman and Hall, New York

Hsu CN, Huang HJ, Wong TT (2003) Implications of the Dirichlet assumption for discretization of continuous attributes in naïve Bayesian classifiers. Mach Learn 53:235–263

Ishwaran H, James LF (2001) Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc 96:161–173

Kohavi R, Sahami M (1996) Error-based and entropy-based discretization of continuous features. In: Proceedings of the second international conference on knowledge discovery and data mining, Portland, OR, pp 114–119

Langley P, Iba W, Thompson K (1992) An analysis of Bayesian classifiers. AI Research Branch, NASA Ames Research Center, Moffett Field, CA 94035, USA

Lochner RH (1975) A generalized Dirichlet distribution in Bayesian life testing. J Roy Stat Soc Series B 37:103–113

Mitchell TM (1997) Machine learning. McGraw-Hill

Wilks SS (1962) Mathematical Statistics. John Wiley, New York

Wong TT (1998) Generalized Dirichlet distribution in Bayesian analysis. Appl Math Comput 97:165–181

Wong TT (2005) A Bayesian approach employing generalized Dirichlet priors in predicting microchip yields. J Chin Inst Ind Eng 22:210–217

Wong TT (2007) Perfect aggregation of Bayesian analysis on compositional data. Stat Papers 48:265–282