Generalizing Conditional Naive Bayes Model

Sahar Salmanzade Yazdi[1], Fatma Najar[2] and Nizar Bouguila[1]
[1]*Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC., Canada*
[2]*City University of New York (CUNY), John Jay College, New York, NY, U.S.A.*

Keywords: Conditional Naive Bayes Model (CNB), Latent Dirichlet Allocation (LDA), LD-CNB Model, LGD-CNB Model, LBL-CNB Model.

Abstract: Given the fact that the prevalence of big data continues to evolve, the importance of information retrieval techniques becomes increasingly crucial. Numerous models have been developed to uncover the latent structure within data, aiming to extract necessary information or categorize related patterns. However, data is not uniformly distributed, and a substantial portion often contains empty or missing values, leading to the challenge of "data sparsity". Traditional probabilistic models, while effective in revealing latent structures, lack mechanisms to address data sparsity. To overcome this challenge, we explored generalized forms of the Dirichlet distributions as priors to hierarchical Bayesian models namely the generalized Dirichlet distribution (LGD-CNB model) and the Beta-Liouville distribution (LBL-CNB model). Our study evaluates the performance of these models in two sets of experiments, employing Gaussian and Discrete distributions as examples of exponential family distributions. Results demonstrate that using GD distribution and BL distribution as priors enhances the model learning process and surpass the performance of the LD-CNB model in each case.

1 INTRODUCTION

In the realm of unsupervised learning, the structure of the data remains hidden from the observer which prompted the development of probabilistic mixture models. Indeed, a powerful approach aimed at figuring out this hidden structure, given that the data comprises a mixture of multiple underlying components (Li et al., 2016). Naive-Bayes (NB) models are a type of generative mixture models known for their simplicity, accuracy, and speed, making them widely used in tasks like product recommendations, medical diagnoses, software defect predictions, and cybersecurity. In addition, researchers have shown that these models tend to outperform other approaches, such as C4.5, PEBLS, and CN2 classifiers especially in cases with small datasets (Wickramasinghe and Kalutarage, 2021). In the era of big data, it is common to encounter issues like sparsity, missing values, and unobserved data. This is often due to the fact that users have limited knowledge about the vast number of available items. Hence, employing traditional NB models won't be advantageous when dealing with such large-scale datasets. To tackle the sparsity problem, a generalized form of the Naive Bayes model, referred to as the conditional Naive Bayes (CNB) model, was introduced (Taheri et al., 2010). This model calculates the likelihood of each class for a given feature vector by utilizing a subset of observed features, rather than incorporating all of them, thus addressing the sparsity problem. However, unlike the traditional Naive Bayes model, the CNB model does not consider the assumption of feature independence. To tackle this limitation, alternative models were suggested including multi-case model (Sahami et al., 1996), overlapping mixture model (Fu and Banerjee, 2008), aspect model (Hofmann, 2001), LDA (Blei et al., 2003), and LD-CNB model (Banerjee and Shan, 2007). Latent Dirichlet Allocation (LDA) is a probabilistic generative model of a corpus, where documents are represented as random mixtures over latent low-dimensional topic space. Assuming K latent topics, a document is generated by sampling a mixture of these topics, with each topic represented as a probability distribution over the words in the document, and then sampling words from that mixture. The key aspect of LDA is that despite the CNB model, it allows documents to be associated with two or more topics (Blei et al., 2003). The latent Dirichlet conditional Naive-Bayes (LD-CNB) model was presented as a more adaptable model since it

Yazdi, S., Najar, F. and Bouguila, N.
Generalizing Conditional Naive Bayes Model.
DOI: 10.5220/0012524000003690

In *Proceedings of the 26th International Conference on Enterprise Information Systems (ICEIS 2024) - Volume 1*, pages 163-171
ISBN: 978-989-758-692-7; ISSN: 2184-4992

utilizes exponential family distribution in variational approximation for model inference and learning. In the research conducted by Banerjee et al.(Banerjee and Shan, 2007), they applied Gaussian and Discrete distributions as specific examples of such exponential family distributions. Through a comparison between the LD-CNB and the CNB models, it has been demonstrated that the LD-CNB model consistently outperforms the CNB model in terms of having lower perplexity. However, using Dirichlet as a prior distribution in the model can lead to some constraints. To address the limitations associated with the constricting negative covariance structure of Dirichlet distribution, this paper introduces an approach where we suggest employing alternative distributions, specifically the generalized Dirichlet (GD) distribution and the Beta-Liouville (BL) distribution, as priors to define the mixing weights for the data point in the model.

The paper's organization is as follows: In section 2, we provide an overview of the LD-CNB model, its instantiations for exponential family distributions such as Gaussian and Discrete distributions, and the variational Expectation Maximization algorithm used for learning and inference. Section 3 covers a review of the properties of the generalized Dirichlet (GD) distribution and the Beta-Liouville (BL) distribution, our proposed approaches, and the updated model based on each of those prior distributions. Section 4 presents the experimental results obtained from the UCI benchmark repository (Frank, 2010) and Movielens recommendation system dataset (Harper and Konstan, 2015). Finally, in section 5, we offer our conclusions.

2 LATENT DIRICHLET CONDITIONAL NAIVE BAYES

In this section, we will examine the LD-CNB model and discuss the constraints of both the LDA model and the Naive-Bayes model, which led to the development of LD-CNB. Additionally, we will delve into the details of the variational EM algorithm and the computational steps taken to accomplish the goals of model learning and inference.

The LD-CNB model was proposed in response to the limitations of NB models in handling sparsity within large-scale data sets. Because the observer has limited knowledge regarding the magnitude of the items, the likelihood of encountering missing or unobserved values rises. Although NB models demonstrated their accuracy and ability in processing small datasets, they are still not able to handle the sparsity in the case of big data. Furthermore, in the NB model, it is assumed that features come from a single mixture component, which imposes significant limitations on the modeling capabilities of the NB model.

In order to address the challenges associated with sparsity, the Conditional Naive-Bayes (CNB) model was introduced. This model conditions a Naive-Bayes model on only a subset of observed features. Let's assume that d represents the total number of features in the dataset, a subset of features is denoted as $f = \{f_1, ..., f_m\}$, where $m < d$. The conditional probability of the feature vector x is then computed as follows:

$$p(x|\pi, \Theta, f) = \sum_{z=1}^{K} p(z|\pi) \prod_{j=1}^{m} p_\psi(x_j|z, \Theta, f_j) \quad (1)$$

where π represents prior distribution over K components. The term ψ refers to the appropriate exponential family model for feature f_j and $p_\psi(x_j|z, \Theta, f_j)$ is the exponential family distribution for f_j. $z = (1, ..., K)$ and $\Theta = \{\theta_z\}$ are defined as the parameters for the exponential family distribution.

In the context of LDA, a 'data point' is presented as a sequence of tokens (feature), with each token generated from the same discrete distribution, since they are considered semantically identical (Griffiths and Steyvers, 2004). In some applications, instead of considering a feature as a token, each feature is associated with a measured value, which can be real or categorical. Besides that, various features within the feature set can carry distinct semantic meanings. Because the NB model assumes that features come from the same mixture component, they took a Dirichlet prior with parameter α for the mixing weight π to overcome the problem caused by that assumption. Therefore, the process of generating a sample x following the LD-CNB model can be outlined as follows:

1. Choose $\pi \sim \text{Dir}(\alpha)$
2. For each of the observed feature f_j (j=1, ..., m):
   (a) Choose $z_j \sim \text{Discrete}(\pi)$
   (b) Choose a feature value $x_j \sim p_\psi(x_j|z_j, \Theta, f_j)$

When taking the model parameters into account, the joint distribution of (π, z, x) can be expressed as:

$$p(\pi, z, x|\alpha, \Theta, f) = p(\pi|\alpha) \prod_{j=1}^{m} p(z_j|\pi) p_\psi(x_j|z_j, \Theta, f_j) \quad (2)$$

Given the feature set of the entire data set denoted as $F = \{f_1, ..., f_N\}$, the probability of the entire data set $X = \{x_1, ..., x_N\}$ can be calculated as follows:

$$p(X|\alpha,\Theta,f) = \prod_{i=1}^{N}\int_{\pi} p(\pi|\alpha)\left(\prod_{j=1}^{m_i}\sum_{z_{ij}=1}^{K} p(z_{ij}|\pi)p_{\psi}(x_{ij}|z_{ij},\Theta,f_{ij})\right)d\pi \tag{3}$$

It can be seen from Equation 3, that the model is dependent on the observed features and their potential values. Thus, when generating the value x_j for the feature f_j, it is necessary to select the suitable exponential family model (ψ). It's important to note that the choice of family distribution depends on the specific feature because each feature may have a different family distribution.

In the research conducted by Banerjee et al.(Banerjee and Shan, 2007), they utilized a univariate Gaussian distribution for real-valued features and a Discrete distribution for categorical features within each class. For the Gaussian distribution model (LD-CNB-Gaussian), the model parameters are denoted as $\Theta = \{(\mu_{(z,f_j)}, \sigma^2_{(z,f_j)})\}$, where $j = 1,...,d$, and $z = 1,...,K$ (d and K representing the total number of features and the number of latent classes in the dataset, respectively). Therefore, in equation 3, $p_{\psi}(x_{ij}|z_{ij},\Theta,f_{ij})$ can be updated as $p(x_j|\mu_{(z,f_j)},\sigma^2_{(z,f_j)})$. In the case of Discrete distribution (LD-CNB-Discrete model), each feature is allowed to be of a different type and a different number of possible values. Assuming K latent classes ($z = 1,...,K$), and d features with r_j ($j = 1,...,d$) possible values for each feature, the model parameters for latent class z and feature f_j are represented by discrete probability distribution over possible values $\Theta = \{p_{(z,f_j)}(r)\}$, where $r = (1,...,r_j)$.

2.1 Model Learning and Inference

2.1.1 Variational EM Algorithm

Consider y as the observed data generated through a set of latent variables x. Let Θ denotes the model parameter describing the dependencies between variables. Consequently, the likelihood of observing the data can be expressed as a function of Θ. The objective is to identify the optimal value for Θ that maximizes the likelihood, or equivalently, the logarithm of the likelihood, as illustrated in equation 4.

$$\log p(y|\Theta) = \log \int p(x,y|\Theta)\,dx \tag{4}$$

However, the computation of maximum log-likelihood is typically a complex task. As a solution, an arbitrary distribution for hidden variables, denoted as $q(x)$, is defined. The marginal likelihood can then be broken down with respect to $q(x)$ as outlined below:

$$\begin{aligned}\log p(y|\Theta) &= \log \int q(x)\,\frac{p(x,y|\Theta)}{q(x)}\,dx \\ &\quad - \log \int q(x)\,\frac{p(x|y,\Theta)}{q(x)}\,dx \\ &= \mathcal{L}(q(x)|\Theta) + \mathcal{KL}(q(x)||p(x|y,\Theta))\end{aligned} \tag{5}$$

The term $\mathcal{L}(q(x)|\Theta)$ is referred to as the evidence lower bound (ELBO), serving as a lower bound for $\log p(y|\Theta)$ due to the non-negativity of $\mathcal{KL}(q(x)||p(x|y,\Theta))$ (Li and Ma, 2023). To achieve the maximum log-likelihood, we can either minimize $\mathcal{KL}(q(x)||p(x|y,\Theta))$ or maximize the evidence lower bound (ELBO), denoted as $\mathcal{L}(q(x)|\Theta)$. Consequently, rather than directly maximizing the log-likelihood, the focus is on maximizing the ELBO (Li and Ma, 2023). This approach leads to the development of a variational EM algorithm, which iteratively optimizes the lower bound of the log-likelihood.

$$q(\pi,z|\gamma,\phi,f) = q(\pi|\gamma)\prod_{j=1}^{m} q(z_j|\phi_j) \tag{6}$$

$q(\pi,z|\gamma,\phi,f)$ is introduced as a variational distribution over the latent variables conditioned on free parameters γ and ϕ, where γ is a Dirichlet parameter, and $\phi = (\phi_1,...,\phi_m)$ is a vector of multinomial parameters.

Based on the information above, the associated ELBO can be computed as follows:

$$\begin{aligned}\mathcal{L}(\gamma,\phi;\alpha,\Theta) &= \mathbb{E}_q[\log p(\pi|\alpha)] + \mathbb{E}_q[\log p(z|\pi)] \\ &\quad + \mathbb{E}_q[\log p(x|z,\Theta)] + \mathcal{H}(q(\pi)) \\ &\quad + \mathcal{H}(q(z))\end{aligned} \tag{7}$$

The variational EM-step is derived by setting the partial derivatives, with respect to each variational and model parameter, to zero. The ELBO can be optimized iteratively by employing the following set of update equations:

$$\phi_{(z_j,f_j)} \propto \exp\left(\Psi(\gamma_{z_j}) - \Psi(\sum_{z_{j'}=1}^{K}\gamma_{z_{j'}})\right) p_{\psi}(x_j|z_j,\Theta,f_j) \tag{8}$$

$$\gamma_{z_j} = \alpha_{z_j} + \sum_{j=1}^{m}\phi_{(z_j,f_j)} \tag{9}$$

As previously shown, the respective distributions for LD-CNB-Gaussian and LD-CNB-Discrete will be substituted with $p_{\psi}(x_j|z_j,\Theta,f_j)$ in (8). The updated corresponding parameter (Θ) for each model is then calculated as follows:

- LD-CNB-Gaussain

$$\mu_{(z_j,f_j)} = \frac{\sum_{i=1}^{N} \phi_{i(z_j,f_j)} x_{ij}}{\sum_{i=1}^{N} \phi_{i(z_j,f_j)}} \tag{10}$$

$$\sigma^2_{(z_j,f_j)} = \frac{\sum_{i=1}^{N} \phi_{i(z_j,f_j)} (x_{ij} - \mu_{(z_j,f_j)})^2}{\sum_{i=1}^{N} \phi_{i(z_j,f_j)}} \tag{11}$$

- LD-CNB-Discrete

$$p_{(z_j,f_j)}(r) \propto \sum_{i=1}^{N} \phi_{i(z_j,f_j)} x_{ij} \mathbb{1}(r|i,f_j) + \varepsilon \tag{12}$$

In equation 12, the term $\mathbb{1}(r|i,f_j)$ refers to the indicator matrix of observed value r for feature f_j in observation x_i.

However, using Dirichlet as a prior presents some restrictions, especially when modeling correlated topics. First, all data features are bound to share a common variance, and their sum must be equal to one. Consequently, we cannot introduce individual variance information for each component of the random vector. In addition, when using a Dirichlet distribution, we have only one degree of freedom to convey our confidence in the prior knowledge. All the entries in the Dirichlet prior are always negatively correlated which means if the probability of one component increases, the probabilities of the other components must either decrease or remain the same to ensure they still sum up to one (Caballero et al., 2012). These limitations motivated us to employ a generalized form of Dirichlet distribution, namely generalized Dirichlet distribution, and Beta-Liouville distribution as potential priors for the multinomial distribution.

3 PROPOSED APPROACHES

In this section, we provide a concise overview of the generalized Dirichlet distribution and the Beta-Liouville distribution. We then proceed to adjust the equations in accordance with these new priors.

3.1 Latent Generalized Dirichlet Conditional Naive Bayes

To overcome the limitations associated with the Dirichlet distribution, (Bouguila and ElGuebaly, 2008; Bouguila and Ghimire, 2010), Connor and Mosimann introduced the concept of neutrality and developed the generalized Dirichlet distribution (Connor and Mosimann, 1969) which is conjugate to the multinomial distribution Najar and Bouguila (2022a).

In this context, a random vector $\overrightarrow{X}$ is considered completely neutral when, for all values of j ($j < K$), the vector $(x_1, x_2, ..., x_j)$ is independent of the vector $(x_{j+1}, x_{j+2}, ..., x_K)/(1 - \sum_j (x_1, x_2, ..., x_j))$, which means that a neutral vector does not impact the proportional division of the remaining interval among the rest of the variables. By assuming a univariate beta distribution with parameters α and β for each component of $(x_1, x_2, ..., x_{K-1})$, the probability density function for the generalized Dirichlet distribution is derived as follows:

$$GD(\overrightarrow{X}|\alpha_1, ..., \alpha_K, \beta_1, ..., \beta_K) = \prod_{i=1}^{K} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} x_i^{(\alpha_i - 1)} (1 - \sum_{j=1}^{i} x_j)^{\gamma_i} \tag{13}$$

where

$$\text{for } i = 1, 2, ..., K-1, \ \gamma_i = \beta_i - (\alpha_{i+1} + \beta_{i+1})$$
$$\text{and } \gamma_K = \beta_K - 1$$

Note that $\alpha_i, \beta_i > 0$. For $i = 1, 2, ..., K$, $x_i \geq 0$ and $\sum_{i=1}^{K} x_i \leq 1$ (Epaillard and Bouguila, 2019). Assuming $\beta_{i-1} = \alpha_i + \beta_i$ the generalized Dirichlet distribution is reduced to the Dirichlet distribution, which indicates Dirichlet distribution as a special case of the generalized Dirichlet distribution (Bouguila, 2008). The mean, the variance, and the covariance in the case of the generalized Dirichlet distribution, for $i = 1, ..., K-1$ are as follows:

$$E(X_i) = \frac{\alpha_i}{\alpha_i + \beta_i} \prod_{j=1}^{i-1} \frac{\beta_j + 1}{\alpha_j + \beta_j} \tag{14}$$

$$Var(X_i) = E(X_i) \times \left(\frac{\alpha_i + 1}{\alpha_i + \beta_i + 1} \prod_{j=1}^{i-1} \frac{\beta_j + 1}{\alpha_j + \beta_j + 1} - E(X_i) \right) \tag{15}$$

$$COV(X_i, X_d) = E(X_d) \times \left(\frac{\alpha_i}{\alpha_i + \beta_i + 1} \prod_{j=1}^{i-1} \frac{\beta_j + 1}{\alpha_j + \beta_j + 1} - E(X_i) \right) \tag{16}$$

Unlike Dirichlet distribution, the GD distribution has a more general covariance structure, and variables with the same means are not obligated to have the same covariance. Moreover, for GD distribution covariance between two variables is not negative (Najar and Bouguila, 2021). This flexibility and properties of the GD distribution make it desirable prior to the topic modeling and finding the hidden structure of the data (Koochemeshkian et al., 2020).

3.2 Model Learning and Inference

Variational EM Algorithm. In the proposed approach, we consider a GD prior with parameters α, β for the mixing weights of the data points of the model ($\pi \sim GD(\alpha, \beta)$), and Θ as the model parameter. Therefore, the joint distribution of (π, z, x) is calculated as:

$$p(\pi, z, x|\alpha, \beta, \Theta, f) = p(\pi|\alpha, \beta) \prod_{j=1}^{m} p(z_j|\pi) p_\psi(x_j|z_j, \Theta, f_j) \tag{17}$$

Following that, the variational distribution for updated model parameters is defined as:

$$q(\pi, z|\gamma, \lambda, \phi, f) = q(\pi|\gamma, \lambda) \prod_{j=1}^{m} q(z_j|\phi_j) \tag{18}$$

where γ and λ are the parameters for the generalized Dirichlet distribution, and $\phi = (\phi_1, ..., \phi_m)$ denotes a vector of parameters for the multinomial distribution.

Further, in order to determine the maximum likelihood of the data, we seek to maximize the associated lower bound (ELBO), computed as follows:

$$\begin{aligned}\mathcal{L}(\gamma, \lambda, \phi; \alpha, \beta, \Theta) =& \mathbb{E}_q[\log p(\pi|\alpha, \beta)] + \mathbb{E}_q[\log p(z|\pi)] \\ &+ \mathbb{E}_q[\log p(x|z, \Theta)] + \mathcal{H}(q(\pi)) \\ &+ \mathcal{H}(q(z))\end{aligned} \tag{19}$$

It has been demonstrated that the GD distribution belongs to the exponential family, so its expected value is calculated by taking the derivative of its cumulant function (Appendix A). By setting the partial derivatives to zero with regard to each parameter and subsequently deriving the revised equations for variational and model parameters (equations 20-22), we can find the maximum value for the ELBO.

$$\phi_{(z_j, f_j)} \propto p_\psi(x_j|z_j, \Theta, f_j) \times \exp\left(\Psi(\gamma_{z_j}) - \Psi(\lambda_{z_j}) - (\sum_{i=1}^{z_j} \Psi(\gamma_i + \lambda_i) - \Psi(\lambda_i))\right) \tag{20}$$

$$\gamma_{z_j} = \alpha_{z_j} + \sum_{j=1}^{m} \phi_{(z_j, f_j)} \tag{21}$$

$$\lambda_{z_j} = \beta_{z_j} + \sum_{j=1}^{m} \phi_{(z_j, f_j)} \tag{22}$$

Given that the model parameter Θ form is independent of the prior distribution, the update equations for exponential family parameters remain unchanged, as presented in (10-12).

3.3 Latent Beta-Liouville Conditional Naive Bayes

In this study, we will incorporate another distribution known as the Beta-Liouville (BL) distribution as a prior in our model. Research has demonstrated that the BL distribution offers a viable alternative to the Dirichlet and GD distributions for statistically representing proportional data. The BL distribution also serves as a conjugate prior for the multinomial distribution and, similar to the GD distribution, it has a more general covariance structure (Luo et al., 2023). The probability density function for a random vector $\overrightarrow{X}$ following a BL distribution with positive parameter vector $\overrightarrow{\Phi} = (\alpha_1, ..., \alpha_K, \alpha, \beta)$ is expressed as:

$$\begin{aligned}BL(\overrightarrow{X}|\overrightarrow{\Phi}) =& \\ &\Gamma(\sum_{k=1}^{K} \alpha_k) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{k=1}^{K} \frac{X_k^{(\alpha_k - 1)}}{\Gamma(\alpha_k)} \\ &\times \left(\sum_{k=1}^{K} X_k\right)^{\alpha - \sum_{k=1}^{K} \alpha_k} \left(1 - \sum_{k=1}^{K} X_k\right)^{\beta - 1}\end{aligned} \tag{23}$$

The Beta-Liouville distribution transforms into the Dirichlet distribution when the generator density follows a beta distribution with parameters $\sum_{i=1}^{K-1} \alpha_i$ and α_K, as explained in (Fan and Bouguila, 2015; Bouguila, 2010). The mean, the variance, and the covariance of the Beta-Liouville distribution are calculated as follows:

$$E(X_i) = \frac{\alpha}{\alpha + \beta} \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k} \tag{24}$$

$$\begin{aligned}Var(X_i) =& E(X_i) \left(\frac{\alpha + 1}{\alpha + \beta + 1} \frac{\alpha_k + 1}{\sum_{k=1}^{K} \alpha_k + 1}\right) \\ &- E(X_i)^2 \left(\frac{\alpha_k^2}{(\sum_{k=1}^{K} \alpha_k)^2}\right)\end{aligned} \tag{25}$$

$$\begin{aligned}COV(X_i, X_d) =& \frac{\alpha_i \alpha_d}{\sum_{i=1}^{K} \alpha_i} \Big(- \frac{\alpha^2}{(\alpha+\beta)^2 \sum_{k=1}^{K} \alpha_k} \\ &+ \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)(\sum_{k=1}^{K} \alpha_k + 1)} \Big)\end{aligned} \tag{26}$$

By considering a Beta-Liouville (BL) prior with parameter vector $\overrightarrow{\Phi}$ from which the mixing weight π is generated, we compute the joint distribution of (π, z, x), the associated variational distribution, and the lower bound, respectively, as outlined below:

$$p(\pi,z,x|\overrightarrow{\Phi},\Theta,f) = p(\pi|\overrightarrow{\Phi})\prod_{j=1}^{m} p(z_j|\pi)p_\psi(x_j|z_j,\Theta,f_j) \tag{27}$$

$$q(\pi,z|\overrightarrow{\Phi},\phi,f) = q(\pi|\overrightarrow{\Omega})\prod_{j=1}^{m} q(z_j|\phi_j) \tag{28}$$

$$\begin{aligned}\mathcal{L}(\overrightarrow{\Omega},\phi;\overrightarrow{\Phi},\Theta) = {} & \mathbb{E}_q[\log p(\pi|\overrightarrow{\Phi})] + \mathbb{E}_q[\log p(z|\pi)] \\ & + \mathbb{E}_q[\log p(x|z,\Theta)] + \mathcal{H}(q(\pi)) \\ & + \mathcal{H}(q(z))\end{aligned} \tag{29}$$

where, $\overrightarrow{\Omega} = (\gamma_1,...,\gamma_K,\gamma,\lambda)$ is the Beta-Liouville parameter vector and $\phi = (\phi_1,...,\phi_m)$ are the multinomial parameters. The BL distribution is also a member of the exponential family (Appendix B), thus the expected value will be obtained by computing the derivative of its cumulant function (Bakhtiari and Bouguila, 2016). Consequently, the corresponding variational parameters are updated as follows:

$$\phi_{(z_j,f_j)} \propto p_\psi(x_j|z_j,\Theta,f_j)\times \exp\left(\Psi(\gamma_{z_j}) - \Psi(\sum_{z_j=1}^{K}\gamma_{z_j}) + \Psi(\lambda) - \Psi(\gamma+\lambda)\right) \tag{30}$$

$$\gamma_{z_j} = \alpha_{z_j} + \sum_{j=1}^{m}\phi_{(z_j,f_j)} \tag{31}$$

4 EXPERIMENTAL RESULTS

To evaluate the performance of our LGD-CNB and LBL-CNB models and to compare them with LD-CNB model for each experiment, we selected different sets of data. This assessment examines how three different priors affect the Gaussian and Discrete models.

4.1 Gaussian Models

As mentioned earlier, Gaussian models are suitable for features with real values. Table 1 displays the calculated perplexities for LD-CNB-Gaussian, LGD-CNB-Gaussian, and LBL-CNB-Gaussian models across five different datasets. These datasets are chosen from the UCI benchmark repository, in which all features are available for every instance. The model was trained using 70% of the data, and the remaining 30% was utilized for testing. Perplexity values are then computed on the testing set using equation 32, with the same number of selected features for all instances in the dataset. The perplexity values after 10 iterations are presented in Table 1. According to equation 32, lower perplexity indicates a higher log-likelihood probability, suggesting a better fit for the model.

$$Perplexity(X) = \exp\Big\{-\frac{\sum_{i=1}^{N}\log p(x_i)}{\sum_{i=1}^{N} m_i}\Big\} \tag{32}$$

Table 1: Perplexity of LD-CNB, LGD-CNB, and LBL-CNB Gaussian models.

	LD	LGD	LBL
Wine	0.9936	0.9804	0.93262
Balance	0.9966	0.9810	0.9953
HeartFailure	0.9837	0.8792	0.7987
WDBC	0.9967	0.9943	0.9925
Yeast	0.9974	0.9963	0.9959

Results indicate that LGD-CNB-Gaussian and LBL-CNB-Gaussian models perform better than LD-CNB-Gaussian, showing that the generalized structure of GD distribution and BL distribution makes them more suitable as prior distributions. Table 2 displays the outcomes of assessing the models on the WDBC dataset (Wolberg and Street, 1995). These results represent the averages obtained from 20 runs with distinct randomly assigned initial values. According to the table, both the LGD-CNB model and LBL-CNB model outperform the LD-CNB model, showcasing higher accuracy in those instances.

Table 2: Accuracy, precision, recall, and f-score in percent for LD-CNB, LGD-CNB, LBL-CNB using the WDBC dataset.

	Accuracy	Precision	Recall	F-score
LD	0.63	0.55	0.075	0.132
LGD	0.85	0.75	0.90	0.828
LBL	0.82	0.92	0.55	0.688

4.2 Discrete Models

To assess Discrete models, we utilized the 100K MovieLens dataset from the Grouplens Research Project (Harper and Konstan, 2015). This dataset comprises 100,000 ratings (1-5) provided by 943 users for 1682 movies. Users with fewer than 20 ratings or incomplete demographic information were excluded. Due to users not rating all movies, there is sparsity in this dataset. Similar to the prior experi-

ment, we conducted the experiment on our three models and computed the perplexity for each using Equation 33.

$$Perplexity(X) = \exp\left\{-\frac{\sum_{i=1}^{N}\log p(x_i)}{N}\right\} \quad (33)$$

The disparity in the number of rated movies among users serves as evidence for the dataset's sparsity, a notable distinction between the Discrete model and the Gaussian model. Moreover, there are no constraints on the covariance of data points, implying that a user rating fewer movies does not necessitate another user to rate more. As illustrated in Figure 1, despite the sparsity, the perplexity of the LGD-CNB-Discrete model is lower than that of the LBL-Discrete model, which, in turn, is lower than the LD-CNB-Discrete model. This finding suggests that similar to real-valued features, the more general covariance structure of the GD and BL priors allows them to better describe proportional data, leading to their superior performance over the Dirichlet prior to categorical features. Additionally, the presence of two vector parameters in the GD distribution enhances its flexibility with sparse data, enabling it to assign low values effectively (Najar and Bouguila, 2022b). In this experiment, we compute similarly the accuracy, precision, recall, and F-score for this model using the testing set. Table 3 illustrates that the suggested approaches have led to an enhancement in the overall performance.

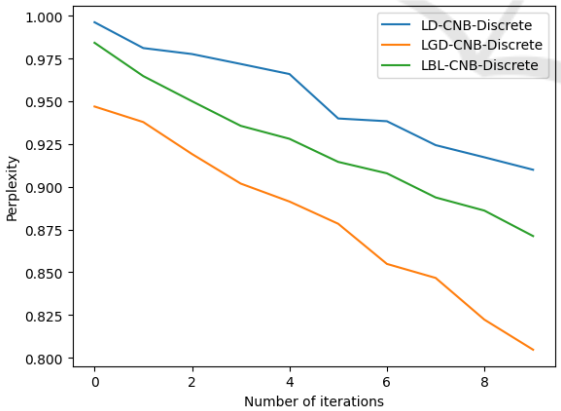

Figure 1: Perplexity for LD-CNB-Discrete, LGD-CNB-Discrete, and LBL-CNB-Discrete.

Table 3: Accuracy, precision, recall and f-score in percent for LD-CNB, LGD-CNB, LBL-CNB Discrete models.

	Accuracy	Precision	Recall	F-score
LD	0.83	0.62	0.052	0.095
LGD	0.87	0.75	0.51	0.607
LBL	0.86	0.74	0.39	0.51

5 CONCLUSION

In this paper, we have presented the incorporation of the GD and BL distributions as priors in the CNB model to address sparsity in large-scale datasets. Utilizing the Conditional Naive Bayes (CNB) model, we conditioned the model on observed feature subsets, enhancing sparsity management. The traditional approach assumed a Dirichlet distribution as a prior, LD-CNB acknowledges that feature values are generated from an exponential family distribution, varying depending on the considered feature. We have outlined the advantages of employing GD and BL distributions over the Dirichlet distribution. Our investigation into the Gaussian and Discrete distributions as exponential families for LGD-CNB and LBL-CNB models revealed that the more generalized covariance structure of GD and BL distributions makes them desirable as prior distributions for uncovering latent structures in sparse data, especially when feature vectors follow a discrete distribution.

REFERENCES

Bakhtiari, A. S. and Bouguila, N. (2016). A latent beta-liouville allocation model. *Expert Systems with Applications*, 45:260–272.

Banerjee, A. and Shan, H. (2007). Latent dirichlet conditional naive-bayes models. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 421–426. IEEE.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bouguila, N. (2008). Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):462–474.

Bouguila, N. (2010). Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, 22(2):186–198.

Bouguila, N. (2011). Hybrid generative/discriminative approaches for proportional data modeling and classification. *IEEE Transactions on Knowledge and Data Engineering*, 24(12):2184–2202.

Bouguila, N. and ElGuebaly, W. (2008). On discrete data clustering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 503–510. Springer.

Bouguila, N. and Ghimire, M. N. (2010). Discrete visual features modeling via leave-one-out likelihood estimation and applications. *Journal of Visual Communication and Image Representation*, 21(7):613–626.

Caballero, K. L., Barajas, J., and Akella, R. (2012). The generalized dirichlet distribution in enhanced topic

detection. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 773–782.

Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206.

Epaillard, E. and Bouguila, N. (2019). Data-free metrics for dirichlet and generalized dirichlet mixture-based hmms–a practical study. *Pattern Recognition*, 85:207–219.

Fan, W. and Bouguila, N. (2015). Expectation propagation learning of a dirichlet process mixture of beta-liouville distributions for proportional data clustering. *Engineering Applications of Artificial Intelligence*, 43:1–14.

Frank, A. (2010). Uci machine learning repository. *http://archive. ics. uci. edu/ml*.

Fu, Q. and Banerjee, A. (2008). Multiplicative mixture models for overlapping clustering. In *2008 Eighth IEEE International Conference on Data Mining*, pages 791–796. IEEE.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42:177–196.

Koochemeshkian, P., Zamzami, N., and Bouguila, N. (2020). Flexible distribution-based regression models for count data: Application to medical diagnosis. *Cybern. Syst.*, 51(4):442–466.

Li, T. and Ma, J. (2023). Dirichlet process mixture of gaussian process functional regressions and its variational em algorithm. *Pattern Recognition*, 134:109129.

Li, X., Ling, C. X., and Wang, H. (2016). The convergence behavior of naive bayes on large sparse datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(1):1–24.

Luo, Z., Amayri, M., Fan, W., and Bouguila, N. (2023). Cross-collection latent beta-liouville allocation model training with privacy protection and applications. *Applied Intelligence*, pages 1–25.

Najar, F. and Bouguila, N. (2021). Smoothed generalized dirichlet: A novel count-data model for detecting emotional states. *IEEE Transactions on Artificial Intelligence*, 3(5):685–698.

Najar, F. and Bouguila, N. (2022a). Emotion recognition: A smoothed dirichlet multinomial solution. *Engineering Applications of Artificial Intelligence*, 107:104542.

Najar, F. and Bouguila, N. (2022b). Sparse generalized dirichlet prior based bayesian multinomial estimation. In *Advanced Data Mining and Applications: 17th International Conference, ADMA 2021, Sydney, NSW, Australia, February 2–4, 2022, Proceedings, Part II*, pages 177–191. Springer.

Sahami, M., Hearst, M., and Saund, E. (1996). Applying the multiple cause mixture model to text categorization. In *ICML*, volume 96, pages 435–443.

Taheri, S., Mammadov, M., and Bagirov, A. M. (2010). Improving naive bayes classifier using conditional probabilities.

Wickramasinghe, I. and Kalutarage, H. (2021). Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, 25(3):2277–2293.

Wolberg, William, M. O. S. N. and Street, W. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B.

Zamzami, N. and Bouguila, N. (2019a). Model selection and application to high-dimensional count data clustering - via finite EDCM mixture models. *Appl. Intell.*, 49(4):1467–1488.

Zamzami, N. and Bouguila, N. (2019b). A novel scaled dirichlet-based statistical framework for count data modeling: Unsupervised learning and exponential approximation. *Pattern Recognit.*, 95:36–47.

Zamzami, N. and Bouguila, N. (2022). Sparse count data clustering using an exponential approximation to generalized dirichlet multinomial distributions. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1):89–102.

APPENDIX

A Exponential Form of the Generalized Dirichlet Distribution

The exponential family of distributions is a group of parametric probability distributions with specific mathematical characteristics, making them easily manageable from both statistical and mathematical perspectives. This family encompasses various distributions like normal, exponential, log-normal, gamma, chi-squared, beta, Dirichlet, Bernoulli, and more. Given a measure η, an exponential family of probability distributions is identified as distributions whose density (in relation to η) follows a general form:

$$p(x|\eta) = h(x)\exp\left(\eta^T T(x) - A(\eta)\right) \qquad (34)$$

where, $h(x)$ is referred to as the base measure, $T(x)$ is the sufficient statistic. η is known as natural parameter, and $A(\eta)$ is defined as the cumulant function.

It has been shown that the generalized Dirichlet distribution is a member of the exponential family distributions (Zamzami and Bouguila, 2019a,b, 2022), as evidenced by its representation in the aforementioned form, as illustrated below:

$$GD(\overrightarrow{X}|\alpha_1,...,\alpha_K,\beta_1,...,\beta_K) = \exp\Big(\sum_{k=1}^{K}\Big(\log(\Gamma(\alpha_k+\beta_k)) - \log(\Gamma(\alpha_k)) - \log(\Gamma(\beta_k))\Big) + \alpha_1\log(X_1) + \sum_{k=2}^{K}\alpha_k\Big(\log(X_k) - \log(1-\sum_{t=1}^{k-1}X_t)\Big) + \beta_1\log(1-X_1) + \sum_{k=2}^{K}\beta_k\Big(\log(1-\sum_{t=1}^{k}X_t) - \log(1-\sum_{t=1}^{k-1}X_t)\Big) - \sum_{k=1}^{K}\log(X_k) - \log(1-\sum_{k=1}^{K}X_k)\Big) \tag{35}$$

Based on that we can calculate the base measure, the sufficient statistic, and the cumulant function as (Bouguila, 2011):

$$h(\overrightarrow{X}) = -\sum_{k=1}^{K}\log(X_k) - \log(1-\sum_{t=1}^{K}X_t) \tag{36}$$

$$T(\overrightarrow{X}) = \Big(\log(X_1), \log(X_2)-\log(1-X_1), \log(X_3)-\log(1-X_1-X_2),...,\log(1-X_1), \log(1-X_1-X_2)-\log(1-X_1),..., \log(1-\sum_{t=1}^{K}X_t)-\log(1-\sum_{t=1}^{K-1}X_t)\Big) \tag{37}$$

$$A(\eta) = \Big(\sum_{k=1}^{K}\log(\Gamma(\alpha_k)) + \log(\Gamma(\beta_k)) - \log(\Gamma(\alpha_k+\beta_k))\Big) \tag{38}$$

given $\eta = (\alpha_1,...,\alpha_K,\beta_1,...,\beta_K)$.

B Exponential Form of the Beta-Liouville Distribution

Besides the generalized Dirichlet distribution, the Beta-Liouville distribution can also be expressed in the framework of exponential family distributions as it is shown below:

$$BL(\overrightarrow{X}|\alpha_1,...,\alpha_K,\alpha,\beta) = \exp\Big(\log(\Gamma(\sum_{k=1}^{K}\alpha_k)) - \log(\Gamma(\alpha+\beta)) - \log(\Gamma(\alpha)) - \log(\Gamma(\beta)) - \sum_{k=1}^{K}\log(\Gamma(\alpha_k)) + \sum_{k=1}^{K}\alpha_k\Big(\log(X_K) - \log(\sum_{k=1}^{K}X_k)\Big) + \alpha\log(\sum_{k=1}^{K}X_k) + \beta\log(1-\sum_{k=1}^{K}X_k) - \sum_{k=1}^{K}\log(X_k) - \log(1-\sum_{k=1}^{K}X_k)\Big) \tag{39}$$

In this scenario, the determination of the base measure, sufficient statistic, and cumulant function is carried out as follows:

$$h(\overrightarrow{X}) = -\sum_{k=1}^{K}\log(X_k) - \log(1-\sum_{t=1}^{K}X_t) \tag{40}$$

$$T(\overrightarrow{X}) = \Big(\log(X_1) - \log(\sum_{k=1}^{K}X_k), \log(X_2) - \log(\sum_{k=1}^{K}X_k),..., \log(X_K) - \log(\sum_{k=1}^{K}X_k)\Big) \tag{41}$$

$$A(\eta) = \log\Big(\Gamma(\sum_{k=1}^{K}\alpha_k)\Big) + \log(\Gamma(\alpha+\beta)) - \log(\Gamma(\alpha)) - \log(\Gamma(\beta)) - \sum_{k=1}^{K}\log(\Gamma(\alpha_k)) \tag{42}$$

given $\eta = (\alpha_1,...,\alpha_K,\alpha,\beta)$.