Count Data Modeling using MCMC-Based Learing of finite EMSD Mixture Models

Xuanbo Su,[1] Nuha Zamzami,[2] Nizar Bouguila[1]
[1]Concordia Institute for Information Systems Engineering, Concordia University, Montreal, QC, Canada
[2]University of Jeddah, College of Computer Science and Engineering, Department of Computer Science and Artificial Intelligence
Jeddah, Saudi Arabia
s_xuanbo@encs.concordia.ca, nezamzami@uj.edu.sa, nizar.bouguila@concordia.ca

Abstract

Statistical approaches are widely used to analyze and model data. Among the successful statistical approaches, finite mixture models have received a lot attention with their flexibility and ease of use. There are already many finite mixture models to deal with this task, but Exponential Multinomial Scaled Dirichlet (EMSD) has recently show higher accuracy model compared to other state-of-the-art generative models for count data clustering. Thus, in this paper, we present Bayesian learning method for a finite mixture model of Exponential Multinomial Scaled Dirichlet (EMSD) distribution. We developed the estimation method based on a Markov Chain Monte Carlo with Metropolis-Hastings algorithm for learning this model parameters. This proposed method is verified by CON-19 information sentiment clustering and a comparison with other approaches with different models for count data

Introduction

With the advancement of technology, more and more complex data are generated, analyzing such valuable data and extraction of the latent pattern is a topic of interest in different areas of science and technology. One of the main attention grabbing approaches is clustering method, finite mixture models have been frequently used to cluster data into homogeneous groups and finite mixture models are flexibility and ease of use, the count data is widely used in many areas such as machine learning, computer vision and economic (Everitt 2005). This type of data has an obvious feature that it is positively skewed with the high frequency of zeros (Sturman 1999). Thus, using an effective model to analysis of this data is essential. Considering that count data, it always put considerable challenge for the researchers as the burstiness phenomenon (Bouguila and Ziou 2004), the new model Multinomial Scaled Dirichlet (MSD) is the composition of the scaled Dirichlet distribution and the multinomial in the same way that Generalized Dirichlet Distribution (MGD), Multinomial Beta-Liouville Distribution (MBL) are the compositions of the Dirichlet (Zamzami and Bouguila 2019). Elkan has shown the exponential approximation for Dirichlet Compound Multinomial (DCM) distributions (Elkan 2006), which is adjustable in high-dimensional spaces and flexibility. It has shown a better performance and faster computation. Similarly, exponential multinomial scaled Dirichlet (EMSD) has shown a superior performance in many challenging applications that involve high-dimensional count data (Zamzami and Bouguila 2019). In this paper, we will use Monte Carlo simulation technique of Gibbs sampling mixed with Metropolis-Hastings step for Bayesian analysis of complex statistical models. Bayesian estimation is based on learning from data using Bayes' theorem (Bouguila, Ziou, and Hammoud 2009), gaining its efficiency from the fact that it combines both the prior information and the information brought by the data to produce the posterior distribution (Bouguila and Ziou 2004), (Bolstad and Curran 2016). Other similar work have already gained the excellent results from the Bayesian approaches in case of mixture models(Bouguila, Wang, and Hamza 2010),(Bouguila, Ziou, and Hammoud 2009),(Stoneking 2014),(Amayri and Bouguila 2016). We validate the proposed algorithm with high dimensional real count data for coronavirus information sentiment clustering. The remainder of this paper is organized as follows. Section 2 introduces the Multinomial Scaled Dirichlet (MSD) Distribution. In section 3, we will present the introduction for Exponential Multinomial Scaled Dirichlet Distribution (EMSD). Then, we present proposed Bayesian learning algorithm using Gibbs sampling with a combination of the Metropolis-Hastings method in section 4. Section 5 shows the result from real count data and compares with other models and methods. Section 6 concludes this paper.

Multinomial Scaled Dirichlet Distribution

The scaled Dirichlet is a generalization of the Dirichlet distribution obtained after applying the perturbation and powering operations to a Dirichlet random composition. These operations define a vector-space structure in the simplex, and play the same role as sum and product by scalars in real space. We assume the dimension is D, the scaled Dirichlet with a set of parameters $\alpha = (\alpha_1 \cdots \alpha_D)$ which is the shape parameter, and $\beta = (\beta_1 \cdots \beta_D)$ which is the scale parameter. The scale Dirichlet distribution defined by(Aitchison 1982):

$$SD(\rho|\alpha, \beta) = \frac{\Gamma(\alpha) \prod_{d=1}^{D} \beta_d^{\alpha_d} \rho_d^{\alpha_d - 1}}{\prod_{d=1}^{D} \Gamma(\alpha_d)(\prod_{d=1}^{D} \beta_d \rho_d)^{a}} \quad (1)$$

where the $a = \sum_{d=1}^{D} \alpha_w$, and Γ is the Gamma function. Note that the scaled Dirichlet includes the Dirichlet as a special case when all elements of the vector β are equal to a common costant. Compared to the Dirichlet, the scaled Dirichlet has D extra parameters, which enhances the model flexibility (Hankin and others 2010). The good parameterization of scaled Dirichlet gives it the ability to better model variance and covariance. Moreover, unlike Dirichlet, the scaled Dirichlet takes into account relative positions between categories or multinomial cells. These properties make the scaled Dirichlet a more flexible choice as a prior to Multinomial.

The MSD model is composition of the Multinomial and scaled Dirichlet distribution, in this case, has two parameters, which are shape parameter α and scale parameter β, and we assume the $\mathbf{X_i} = [x_1 \cdots x_D]$. The scaled parameter controls how the density plot is spread out, The shape parameter is the form or shape of the scaled Dirichlet distribution (Zamzami and Bouguila 2019).

$$\begin{aligned} MSD(\mathbf{X_i}|\alpha,\beta) &= \int_{\rho} \mathcal{M}(\mathbf{X}|\rho)\mathcal{SD}(\rho|\theta)d\rho \\ &= \frac{n!}{\prod_{d=1}^{D} x_d!} \frac{\Gamma(a)}{\Gamma(a+n)\prod_{d=1}^{D} \beta_d^{x_d}} \prod_{d=1}^{D} \frac{\Gamma(\alpha_d + x_d)}{\Gamma(\alpha_d)} \end{aligned} \tag{2}$$

where D is the vocabulary size, Γ is the Gamma functionand $n = \sum_{d=1}^{D} x_d$

Exponential Multinomial Scaled Dirichlet Distribution

The exponential family of distribution have obvious benefits such as simplicity, effective optimization, it retains the essential information in dataset and reduces the computation time in high-dimension data. Thus, the MSD ditribution has been approximated as a member of the exponential family EMSD distribution (Zamzami and Bouguila 2019), using the following approximation for small α values .

$$\frac{\Gamma(\alpha_d + x_d)}{\Gamma(\alpha_d)} \simeq \Gamma(x_d)\alpha_d \tag{3}$$

Using the fact that if x is an integer, the $x! = x(x-1)!$, the authors have obtained the $\mathcal{EMSD}$ distribution:

$$\mathcal{EMSD}(\mathbf{X_i}|\lambda,\nu) = \frac{n!\Gamma(s)}{\prod_{d=1,x_d \geq 1}^{D} x_d \Gamma(s+n)} \prod_{d=1,x_d \geq 1}^{D} \frac{\lambda_d}{\nu_d^{x_d}} \tag{4}$$

where $s = \sum_{d=1}^{D} \lambda_d$.

Metropolis-Within-Gibbs sampling Estimation Algorithm

A challenging problem when deploying a finite mixture model is learning the model's parameters. The approaches used for parameters estimation could be deterministic or Bayesian (Najar, Zamzami, and Bouguila 2019). Bayesian learning technique shows its superiority over the likelihood-based method in this work. For learning the model's parameters, we apply K-means to obtain the K clusters and initialize our parameters by applying method of moments (MOM) (Manouchehri and Bouguila 2018). Then, the new parameters are estimated by using Gibbs sampling within Metropolis-Hastings algorithm(Bouguila, Ziou, and Hammoud 2009). The entire set of documents is $\mathcal{X} = \{\mathbf{X_1} \cdots \mathbf{X_N}\}$, where each document is described by a D-dimensional $\mathbf{X_i}$, the likehood corresponding to a mixture of M distribution is :

$$P(\mathcal{X}|\lambda,\nu) = \prod_{i=1}^{N} P(\mathbf{X_i}|\lambda,\nu) = \prod_{i=1}^{N}(\sum_{j=1}^{M} \pi_j P(\mathbf{X_i}|\lambda_j,\nu_j)) \tag{5}$$

We proposed a M-dimentional membership indicator $\vec{Z_i} = (Z_{i1} \cdots Z_{iM})$ to each observation, where the $Z_{ij} = 1$ if $\mathbf{X_i}$ belongs to the the component j and zero, otherwise. Thus, for $\mathcal{X}$ we have $\mathcal{Z} = \{Z_1 \cdots Z_N\}$, and the posterior distribution of all parameters over the data set $\mathcal{X}$ is defined by:

$$P(\lambda,\nu|\mathcal{X},Z) \propto P(\lambda,\nu,\pi) \prod_{Z_{ij}} P(\mathbf{X_i}|\lambda_j,\nu_j,\pi_j) \tag{6}$$

where $P(\lambda,\nu,\pi)$ is the prior distribution of parameters, and $\prod_{Z_{ij}} P(\mathbf{X_i}|\lambda_j,\nu_j,\pi_j)$ is the likehood of the data given the model's parameters expressed also as:

$$P(\mathcal{X}|\mathcal{Z},\lambda,\nu) = \prod_{i=1}^{N}\prod_{j=1}^{M}(\pi_j P(\mathbf{X_i}|\lambda_j,\nu_j))^{\mathbf{Z_{ij}}} \tag{7}$$

the mixing weight is $\vec{\Pi} = (\pi_1 \cdots \pi_M)$, it should sums to one and all its values should be postive, thus, our natural choice is the Dirichlet distribution which is defined as:

$$P(\vec{\Pi}|\eta) = \frac{\Gamma(\sum_{j=1}^{M} \eta_j)}{\prod_{j=1}^{M} \Gamma(\eta_j)} \prod_{j=1}^{M} \pi^{\eta_j - 1} \tag{8}$$

where $\eta = (\eta_1 \cdots \eta_M)$ is the Dirichlet distribution's parameter vector. We also have:

$$P(\mathbf{Z}|\vec{\Pi}) = \prod_{i=1}^{N} P(Z_{ij}|\vec{\Pi}) = \prod_{i=1}^{N}\prod_{j=1}^{M} \pi_j^{Z_{ij}} = \prod_{j=1} \pi_j^{\delta_j} \tag{9}$$

where $\delta_j = \sum_{i=1}^{N} Z_{ij}$.

Thus, we can gain the posterior of mixing weight:

$$\begin{aligned} &P(\vec{\Pi}|\mathbf{Z}) \propto P(\vec{\Pi}|\eta)P(\mathbf{Z}|\vec{\Pi}) \\ &= \frac{\Gamma(\sum_{j=1}^{M} \eta_j)}{\prod_{j=1}^{M} \Gamma(\eta_j)} \prod_{j=1}^{M} \pi^{\eta_j + \delta_j - 1} \propto \mathbf{M}(\eta_1 + \delta_1 \cdots \eta_M + \delta_M) \end{aligned} \tag{10}$$

We assume $\vec{Z_i}^{(t)}$generate from Multinomial distribution $\mathcal{M}(1, P(1|\vec{X_i}^{(t-1)}) \cdots P(M|\vec{X_i}^{(t-1)}))$,where $P(j|\vec{X_i})$ is the posterior distribution defined by:

$$P(j|\vec{X_i}) = \frac{\pi_j P(\vec{X_i}|\lambda_j,\nu_j)}{\sum_{j=1}^{M} \pi_j P(\vec{X_i}|\lambda_j,\nu_j)} \tag{11}$$

The EMSD is a member of the expontial family of distributations, if s-parameters density belongs to the exponential family, it can be written:

$$P(\vec{x}|\vec{\theta}) = H(\vec{x}) exp(\sum_{l=1}^{s} G_l(\vec{\theta}) T_l(\vec{x}) + \Phi(\vec{\theta})) \tag{12}$$

where $T_l(\vec{x})$ is a vector of suffcient statistic, $G_l(\vec{\theta})$ is the vector for natural parameters, $H(\vec{x})$ is the underlying measure and $\Phi(\vec{\theta})$ is called log normalizer which ensures that the distribution in tegrates to one. Then, by letting:

$$\begin{aligned}
H(\vec{x}) &= n!(\prod_{x_d \geq 1} x_d^{-1}) \\
G_l(\vec{\theta}) &= [\log(\lambda_{jd}) - \log(\nu_{jd})] \\
T_l(\vec{x}) &= \begin{bmatrix} \sum_{d=1}^{D} I(x_d \geq 1) \\ \sum_{d=1}^{D} x_d \end{bmatrix} \\
\log(\Gamma(s+n_i)) &= \log(\Gamma(s)) + \sum_{t=1}^{n-1} \log(s+t) \\
\Phi_l(\vec{\theta}) &= \log(\frac{\Gamma(s)}{\Gamma(s+n_i)}) = -\sum_{t=1}^{n-1} \log(s+t)
\end{aligned} \tag{13}$$

Thus, the EMSD can be rewritten as:

$$\begin{aligned}
\mathcal{EMSD}(\mathbf{X_i}|\lambda_j, \nu_j) &= (\prod_{x_d \geq 1} {x_d}^{-1}) n! \frac{\Gamma(s)}{\Gamma(s+n)} \\
&= \left\{ \exp(\sum_{d=1}^{D} I(x_d \geq 1)(\log(\lambda_{jd}) - x_w \log(\nu_{jd})) \right\}
\end{aligned} \tag{14}$$

where the $\vec{\theta} = (\lambda_{jd}, \nu_{jd})$.
In this case, a conjugate prior for $\vec{\theta}$ is given by:

$$\begin{aligned}
P(\theta_j) &\propto \exp(\sum_{d=1}^{D} \rho_l G_l(\theta_j) + k\Phi(\theta_j)) \\
&\propto \exp\left[\sum_{d=1}^{D} (\rho_1 \log(\lambda_{jd}) - \right. \\
&\qquad \left. \rho_2 \log(\nu_{jd})) - k \sum_{t=1}^{n-1} \log(s+t) \right]
\end{aligned} \tag{15}$$

where (ρ_1, ρ_2, k) are the prior's hyperparameters. Thus, we can determine the posterior distribution as follows:

$$\begin{aligned}
P(\theta_j|\mathcal{X}, \mathcal{Z}) &\propto P(\theta_j) P(\mathbf{X_i}|\theta_j) \\
&\propto \exp\left\{ \left[\sum_{d=1}^{D} (\log(\lambda_{jd})(\rho_1 \right.\right. \\
&\quad + \sum_{i, z_{ij}=1} I(x_{id} \geq 1)) \\
&\quad \left.\left. - \log(\nu_{jd})(\rho_2 + \sum_{i=1, z_{ij}=1} x_{id}) \right] \right\} \\
&\quad - k \sum_{t=1}^{n-1} \log(s+t) \\
&\quad + \sum_{i, z_{ij}=1} \left\{ \left[log(n!) - \sum_{d, x_{id} \geq 1} log(x_{id}) \right.\right. \\
&\quad \left.\left. - \sum_{t=1}^{n_i - 1} \log(s+t) \right] \right\}
\end{aligned} \tag{16}$$

Considering (Kleiter 1992), once the sample $\mathcal{X}$ is konwn, it can be used to get the prior hyperparameters (Bensmail et al. 1997). The hyperparameters fixed at: $k = 1$, $\rho_1 = 1$, $\rho_2 = \sum_{z_i=1} x_{id}$. the accepted ratio is whether the sample at iteration t should be accepted or refused for the next iteration t+1, the ratio defined as follows:

$$r = \frac{\mathcal{P}(\tilde{\theta_j}|\mathcal{X}, \mathcal{Z}) \mathbf{q}(\theta_j^{(t-1)}|\tilde{\Theta_j})}{\mathcal{P}(\theta_j|\mathcal{X}, \mathcal{Z}) \mathbf{q}(\tilde{\theta_j}|, \theta_j^{(t-1)})}$$

where the $\mathbf{q}$ is proposal distribution. Because the parameter $\lambda \in (0,1)$, we consider Gamma distribution for the λ with $\sigma_0 = 0.03$ as scale parameter and Inverse Gamma distribution for the ν with $\varphi_0 = 1$ as scale parameter, thus, the Metropolis-Hastings algorithm as fllows:
1) Generate $\tilde{\lambda_j} \sim \mathcal{G}(\lambda_j^{(t-1)}, \sigma_0), \tilde{\nu_j} \sim invG(\nu_j^{(t-1)}, \varphi_0)$ and $U \sim \mathcal{U}[0,1]$
2) Computation the acceptance ratio:

$$r = \frac{\mathcal{P}(\tilde{\theta_j}|\mathcal{X}, \mathcal{Z}) \mathcal{G}(\lambda_j^{(t-1)}|\tilde{\lambda_j}, \sigma_0) invG(\nu_j^{(t-1)}|\tilde{\nu_j}, \varphi_0)}{\mathcal{P}(\theta_j|\mathcal{X}, \mathcal{Z}) \mathcal{G}(\tilde{\lambda_j}|, \lambda_j^{(t-1)} \sigma_0) invG(\tilde{\nu_j}|\nu_j^{(t-1)}, \varphi_0)}$$

3) If $r < \mathcal{U}$ then $\lambda_j^t = \tilde{\lambda_j}, \nu^t = \tilde{\nu_j}$ else $\lambda_j^t = \lambda_j^{(t-1)}, \nu_j^t = \nu_j^{(t-1)}$ Therefore, the Gibbs sampling algorithm as follows:

Algorithm 1 Gibss sampling within Metropolis-Hastings

```
Initialization
a. Apply K-means to obtain K clusters
b. Apply method of moments on each component j to get inintial $\vec{\lambda}$
c. set initial $\vec{\nu}$ as 1
repeat
    Generate $\vec{Z_i^{(t)}} \sim \mathcal{M}(1, P(1|\vec{X_i}^{(t-1)}) \cdots P(M|\vec{X_i}^{(t-1)}))$
    Generate $\prod_j^{(t)}$ from $P(\prod|\mathbf{Z})$
    Generate $\vec{\theta_j^{(t)}}$ from Eq.14 using Gibbs sampling algorithm
until Convergence of parameters
```

CON-19 Information Sentiment Clustering

In 2020, a new virus swept the world, it brings many disasters to the world. In addition to the epidemic, it also has become important to keep abreast of people's attitudes towards the virus from the comments related to the CON-19 on the internet. We use 3798 information from Twitter[1], the data in these sources have been classified into five labels including: Neutral, Positive, Negative, Extremely Negative, Extremely Positive.

We use bag-of-words and the preprocess for this data in our experiments is removing all the stops words and rare words (words less than 15 occurrences), then we transform each text into a vector of counts containing the number of occurrences for each given word in a text document. Given the estimated EMSD parameters, the clustering is performed by applying the Bayes's rule. We evaluate our proposed Bayesian inference methods with different multinomial-based methods based on accuracy, precision, recall, and F-measure. The compared models are: mixture of multinomial model (MM), the Dirichlet compound multinomial mixture model (DCM),Multinomial Scaled Dirichlet mixture model (MSD), the mixture of EDCM models and EMSD with expectation maximization (EM) algorithm. The obtained results are shown in table 1. According to the evaluation metrics, mixture of multinomial model (MM) has the worst results due to the "naïve Bayes assumption", the DCM and MSD have better results compared to MM. Noting EMSD and EDCM models outperform their corresponding models (DCM and MSD). Our proposal model achieves clearly superior results.

Table 1: RESULT

Models	Accuracy(%)	Precision	Recall(%)	F-measure(%)
MM	21.3	20.3	20.0	21.3
DCM	47.3	48.42	48.23	48.32
MSD	48.66	43.70	64.61	52.14
EDCM-EM	71.88	72.80	73.09	72.95
EMSD-EM	76.72	77.56	77.64	77.60
EMSD-Bayesian	**83.21**	**82.43**	**83.11**	**83.32**

[1] https://www.kaggle.com/datatattle/covid-19-nlp-text-classification

Conclusion

In this paper, we introduced a new Bayesian method for learning the parameters of a mixture model based on the exponential family approximation to MSD. We presented an Markov Chain Monte Carlo method to evaluate the posterior distribution and bayes estimator by Gibbs sampling. We demonstrated the result in real test data of CON-19 sentiment clustering and compare with other models and methods. The promising results are explained by taking advantage of the posterior information which is not considered in the other methods. We can conclude the EMSD model with the proposed inference method offers a promising clustering approach and can be used in other real life applications with sparse high-dimensional data.

References

Aitchison, J. 1982. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* 44(2):139–160.

Amayri, O., and Bouguila, N. 2016. A bayesian analysis of spherical pattern based on finite langevin mixture. *Applied Soft Computing* 38:373–383.

Bensmail, H.; Celeux, G.; Raftery, A. E.; and Robert, C. P. 1997. Inference in model-based cluster analysis. *Statistics and Computing* 7(1):1–10.

Bolstad, W. M., and Curran, J. M. 2016. *Introduction to Bayesian statistics*. John Wiley & Sons.

Bouguila, N., and Ziou, D. 2004. Improving content based image retrieval systems using finite multinomial dirichlet mixture. In *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, 23–32. IEEE.

Bouguila, N.; Wang, J. H.; and Hamza, A. B. 2010. Software modules categorization through likelihood and bayesian analysis of finite dirichlet mixtures. *Journal of Applied Statistics* 37(2):235–252.

Bouguila, N.; Ziou, D.; and Hammoud, R. I. 2009. On bayesian analysis of a finite generalized dirichlet mixture via a metropolis-within-gibbs sampling. *Pattern Analysis and Applications* 12(2):151–166.

Elkan, C. 2006. Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning*, 289–296.

Everitt, B. S. 2005. Latent class analysis. *Encyclopedia of Statistics in Behavioral Science*.

Hankin, R. K., et al. 2010. A generalization of the dirichlet distribution. *Journal of Statistical Software* 33(11):1–18.

Kleiter, G. D. 1992. Bayesian diagnosis in expert systems. *Artificial Intelligence* 54(1-2):1–32.

Manouchehri, N., and Bouguila, N. 2018. Learning of finite two-dimensional beta mixture models. In *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*.

Najar, F.; Zamzami, N.; and Bouguila, N. 2019. Fake news detection using bayesian inference. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, 389–394. IEEE.

Stoneking, C. J. 2014. Bayesian inference of gaussian mixture models with noninformative priors. *arXiv preprint arXiv:1405.4895*.

Sturman, M. C. 1999. Multiple approaches to analyzing count data in studies of individual differences: The propensity for type i errors, illustrated with the case of absenteeism prediction. *Educational and Psychological Measurement*.

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