Simultaneous count data feature selection and clustering using Multinomial Nested Dirichlet Mixture

Fares Alkhawaja, Manar Amayri, Nizar Bouguila
Concordia Institute for Information Systems Engineering (CIISE), Concordia University
1515 St.Catherine Street West, Montreal, Quebec H3G 2W1
f_a68198@live.concordia.ca, manar.amayri@concordia.ca, nizar.bouguila@concordia.ca

Abstract

The elevating effect of the curse of dimensionality in count data has made clustering a challenging task. This paper solves this by adopting the concept of feature saliency as a feature selection method in the context of using the Multinomial Nested Dirichlet Mixture (MNDM). The MNDM is a generalization of the Dirichlet Compound Mixture (DCM) that suffers from several limitations. The model learning is accomplished through the expectation-maximization method. The Minimum Message Length criterion is used to simultaneously determine the best number of components in the mixture with the updated selected features. At the price of convergence times, the results show better performance through different metrics, as the model aims to select the salient features and tune away the non-salient anomalistic features.

Introduction

The dramatic increase in data and the forecasted higher pace of its increment have made it more challenging to cluster the different populations and extract patterns accordingly. Consequently, with the abundance of highly dimensional challenging datasets, it has been observed that many features in high-dimensional representations tend to be less effective or nonsalient (Guyon and Elisseeff, 2003). Despite the clustering powerfulness in the context of machine learning, its performance can be drastically affected by redundant or outlying features (Bdiri et al., 2016; Bouguila, 2010; Bouguila and Ziou, 2012; Bouguila et al., 2012; Liu et al., 2011).

A very well-known example of high-dimensional data is count data. Count data is a representation of frequencies for the occurrence of each feature (Harris et al., 2014). It has gained scholars' attention over the past decades due to its wide appearance in many data collections, such as the words in a document, the visual keywords in images, and the taxa in microbial data. However, it usually appears in overly-sparsed vectors where many indices are zeros as a result of feature absences (Dhillon and Modha, 2001). This, in turn, hinders the clustering process, resulting in numerical divergence. The multinomial distribution has shown a competing performance over many other distributions (Bouguila, 2007, 2009; Bouguila and Ziou, 2004; Najar and Bouguila, 2022).

Burstiness is another challenge in count data clustering that is associated with the Multinomial assumption (Naïve Bayes assumption) (McCallum et al., 1998). The burstiness problem is linked to the correlation between the first and second appearance of a rare word[1] or visual word that is failed to be captured by the Multinomial distribution (Church and Gale, 1995). Therefore, clustering techniques of count data have explored prior distribution that characterizes the parameters of the Multinomial distribution, such as the Dirichlet distribution. The Dirichlet distribution has been an effective prior to the Multinomial distribution, resulting in a Multinomial Dirichlet distribution or the DCM (Mosimann, 1962). Dirichlet distribution has attracted a lot of attention due to its multinomial conjugacy and its powerful adaptation to small-sized datasets as part of statistical models (Bouguila and Ziou, 2005a,b, 2006; Oboh and Bouguila, 2017). However, the Dirichlet distribution suffers from its negative covariance matrix and its direct proportionality between its mean and variance (Alkhawaja and Bouguila, 2023).

The Multinomial Nested Dirichlet distribution (MNDD) is a generalization that was first introduced in (Null, 2008), which solves the earlier-mentioned limitations of the Dirichlet distribution. It is based on the Nested Dirichlet distribution (NDD) that generalized the Dirichlet distribution by adding a hierarchical structure. Indeed, the NDD is a special case of the generalized Dirichlet distribution (Connor and Mosimann, 1969), where its hierarchy is limited to two features per node (Minka, 1999). Figure 1 illustrates the hierarchical difference between the later distributions. In this paper, the Multinomial Nested Dirichlet Mixture (MNDM) is adopted in modeling visual data using Bag of Visual Words (BoVW) (Null, 2008).

Moreover, in endeavors to tackle the curse of dimensionality and the sparseness in count data, the concept of feature saliency for the high-dimensional observational vectors is adopted. The paradigm of feature saliencies is based on extracting the most effective unique features in all the components, where the number of components is determined by the Minimum Message Length criterion since the number

[1]The likelihood of a specific word to appear twice (if it appears once in a document) is higher than its first appearance.

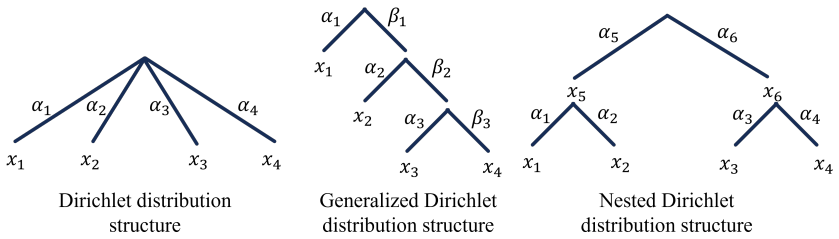

Figure 1: Tree structure for Dirichlet, generalized Dirichlet, and Nested Dirichlet distributions.

of components and the optimal selected features are interrelated (Hong et al., 2019).

Therefore, in this paper, the main contributions are:

1. Perform clustering using the MNDM for count data.
2. Simultaneous feature selection based on the feature saliency concept.
3. Using MML to determine the number of components simultaneously.

The rest of the paper is divided into the Related Works in Section 2, followed by the proposed method in Section 3. Section 4 shows the derivations of the MNDM and its learning. Section 5 shows the obtained results. Finally, Section 6 concludes this paper.

Related works

Due to their importance in the era of rapid information expansion, dimensionality reduction methods have been extensively explored and introduced in many fields, including sensory-robotics (including localization tasks) (Panteleris and Argyros, 2022), computer vision (Wang et al., 2022), and medical fields (Remeseiro and Bolon-Canedo, 2019). Feature extraction and feature selection are two main approaches that help reduce the number of features and obtain explainable and effective features. Feature extraction methods perform dimensionality reduction faster at the cost of information loss and less interpretability in the retrieved features (Solorio-Fernández et al., 2022). The examples of feature extraction methods are many, such as principal component analysis (Zhang et al., 2017), linear discriminant analysis (Xanthopoulos et al., 2013), partial least square (Marquetti et al., 2016), t-distributed stochastic neighbor embedding (Sharma et al., 2021), and latent Dirichlet allocation in topics models (Blei et al., 2003).

Feature selection methods aim at approving selected features within a dataset that can be interpretable. The unsupervised feature selection methods have three main categories: filter/ranking, wrapper, and wrapper-embedded/hybrid techniques (Hong et al., 2019). Filter methods refine the features based on their property through distance, entropy, and dependency metrics, resulting in multiple ranks for the features (Solorio-Fernández et al., 2022). This, in turn, makes it independent of the components of the learning model, resulting in a faster evaluation. However, this comes at the cost of having redundant features or equally contributing features if they have the same rank. Filter methods were first introduced in (Dash et al., 1997), where the authors used the entropy measure to select features sequentially in a reversed manner through the so-called sequential backward selection. Therefore, a lower entropy results in more well-defined components. Other techniques include the textual TF-IDF measure (Bouillot et al., 2013), mutual information (Kraskov et al., 2004), and laplacian score (He et al., 2005).

Unlike filter methods, wrapper-embedded is based on simultaneous unsupervised clustering and feature selection. Therefore, the independence assumption is dropped, leading to a better understanding of the salient features of the current clusters (Hong et al., 2019). The wrapper approaches are a sub-division of wrapper-embedded approaches, where the clustering and feature selection are accomplished in two steps. Therefore, wrapper-embedded techniques are sometimes referred to as hybrid techniques. The hybrid techniques combine the two benefits of filter and wrapper techniques through faster convergence and higher accuracy. Since wrapper techniques are computationally demanding, the convergence time is higher than the filter ones (Boutemedjet et al., 2009). Examples of wrapper techniques include scatter class separability (Dy and Brodley, 2004), and backward elimination (Law et al., 2002), where the authors estranged features based on their class conditional dependence iteratively. Hybrid methods have recently gained higher interest, as they are adopted in the context of mixture models as in (Boutemedjet et al., 2009; Law et al., 2004; Zamzami and Bouguila, 2022), where the number of components is estimated through model selection criteria.

Indeed, model selection is pivotal for the selected features, and the overall model performance, as it optimizes the model towards the best number of clusters to represent the data Mashrgy et al. (2014). Model selection criteria include stochastic and deterministic methods. Deterministic methods have shown cost efficiency and relatively good accuracy levels. Among multiple deterministic methods such as Akaike Information Criterion (AIC) (Sakamoto et al., 1986), Bayesian Information Criterion (BIC) (Weakliem, 1999), and Minimum Description Length (MDL) (Barron et al., 1998), Minimum Message Length (MML) has shown a remarkable performance in determining the number of clusters, comparatively. Therefore, in this work, simultaneous unsupervised clustering is implemented along with feature selection using the MNDM and the concept of the salient features, respectively.

The proposed method

The MNDD offers multiple advantages thanks to its conjugacy and its generalized form over the Multinomial Dirichlet distribution. Indeed, the MNDD, as mentioned earlier, is based on the NDD, which is a generalization over the Dirichlet distribution, in which multiple nestings/nodes are introduced. These nestings can be at different levels to build the so-called Dirichlet tree. The NDD was first introduced by (Dennis III, 1991), and was further revisited by (Minka, 1999), and (Null, 2008). The two forms are similar, with a difference in the probability density function (PDF) representation, as was further investigated by (Alkhawaja and Bouguila, 2023). In this paper, we are using the form introduced by (Null, 2008), due to the obtained advantages introduced in (Alkhawaja and Bouguila, 2023). Therefore, for an (i^{th}) observation, a $D+K$ dimensional vector $\vec{X}_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{i(D+K)})$ follows a Multinomial distribution with the parameters vector $\vec{P} = (P_1, \ldots, P_j, \ldots, P_{D+K})$, with the following NDD PDF:

$$P(\vec{P} \mid \vec{\alpha}) = \frac{\prod_{j=1}^{D} {P_j}^{\alpha_j - 1} \prod_{k=1}^{K} {P_{D+k}}^{\alpha_{D+k} - \bar{A}_k}}{\prod_{k=0}^{K} B(A_k)} \tag{1}$$

where $\vec{\alpha} = (\alpha_1, \ldots, \alpha_j, \ldots, \alpha_{D+K})$ is the parameters vector for the Dirichlet distribution, ($\bar{A}_k$) is the sum of the (A_k) vector underneath each (k) nesting. The authors in (Null, 2009) showcased that a nesting tree can represent any tree structure using only two nestings at each level $K = D - 2$. Where, K and D are the total nesting and nested variables, respectively. Therefore, the NDD can represent the GDD and any other tree structure using two parameters at each nesting (k). These parameters are $\zeta_k = (\alpha_k, \beta_k)$. Therefore, following to the derivations in (Null, 2008), the PDF of the MNDD, for a (k^{th}) nesting, is written as:

$$\begin{aligned} P(\vec{X}_{ik} \mid \zeta_k) = {} & \frac{\Gamma(X_{ik_1} + X_{ik_2} + 1)}{\Gamma(X_{ik_1} + 1)\Gamma(X_{ik_2} + 1)} \times \\ & \frac{\Gamma(\alpha_k + X_{ik_1})\Gamma(\beta_k + X_{ik_2})\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)\Gamma(\alpha_k + \beta_k + X_{ik_1} + X_{ik_2})} \end{aligned} \tag{2}$$

where $\vec{X}_{ik} = (X_{ik_1}, X_{ik_2})$. Note that this is scalable to encapsulate more than two variables, replacing the Beta distribution with an independent Dirichlet distribution for the targetted (k^{th}) node. Due to the existence of the Gamma (Γ) function in equation (2), the calculations become intractable, leading to approximations that might cause biasing to the estimator (Null, 2008). Despite the abundance of methods that were proposed to solve this issue, the authors in (Null, 2008), followed an efficient re-parameterization technique introduced by (Paul et al., 2005), and generalized the form by replacing the variables $\zeta_k = (\alpha_k, \beta_k)$ by $\Omega_k = (\pi_k, \phi_k)$, which are the mean and the dispersion factor (dispersion among the features), respectively. Then, by using the integer factorial function generalization (Bhargava, 2000), $(X_k - 1)! = \Gamma(X_k)$, the PDF of the MNDD is written as:

$$\begin{aligned} P(\vec{X}_k \mid \Omega_k) = {} & \frac{(X_{k_1} + X_{k_2})!}{(X_{k_1})!(X_{k_2})!} \times \\ & \frac{\prod_{r=0}^{X_{k_1}-1}(r\phi_k + \pi_k) \prod_{r=0}^{X_{k_2}-1}(r\phi_k + 1 - \pi_k)}{\prod_{r=0}^{Y_k-1} 1 + r\phi_k} \end{aligned} \tag{3}$$

where $Y_k = X_{k_1} + X_{k_2}$, and

$$\pi_k = \frac{\alpha_k}{\alpha_k + \beta_k}, \phi_k = \frac{1}{\alpha_k + \beta_k} \tag{4}$$

Note that each nesting (k) is independent of the other at the same level, enabling their PDF to be written independently.

Implmenting the mixture model

The implementation of the model is divided into three parts. In the first part, the model is built with the augmentation of salient features constraint, followed by the estimation of the parameters in the second part. Finally, determining the number of components is illustrated in the third part.

Feature saliency augmentation

As the mixture model is a collection of components, the MNDM model generalizes equation (3) by combining the prior probability vector $\vec{P} = (P(1), \ldots, P(m), \ldots, P(M))$ for M components and N observations, yielding:

$$P(\mathcal{X} \mid \theta) = \prod_{i=1}^{N} \sum_{m=1}^{M} P(\vec{X}_i \mid \vec{\Omega}_m) P(m) \tag{5}$$

where $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$ and $\theta = \langle \Omega = \{\vec{\Omega}_1, \ldots, \vec{\Omega}_m = (\Omega_{m1}, \ldots, \Omega_{mk}, \ldots, \Omega_{mK}), \ldots, \vec{\Omega}_M\}, \vec{P} \rangle$.

By assuming conditional independence among the features, based on each m^{th} component, equation (5) becomes:

$$P(\mathcal{X} \mid \theta) = \prod_{i=1}^{N} \sum_{m=1}^{M} P(m) \prod_{k=1}^{K} P(\vec{X}_{ik} \mid \Omega_{mk}) \tag{6}$$

The feature saliency concept is based on imposing a binary feature relevance parameter. This parameter is selected to be correlated with the current component as in (Zamzami and Bouguila, 2022). Therefore, the binary feature relevance vector is $\vec{\lambda}_m = (\lambda_{m1}, \ldots, \lambda_{mk}, \ldots, \lambda_{mK})$ indicates the pertinence of each k^{th} feature nesting to the m^{th} component. Therefore, by setting $\theta' = \langle \theta, \lambda = \{\vec{\lambda}_1, \ldots, \vec{\lambda}_m \ldots, \vec{\lambda}_M\}, \Omega' = \{\Omega'_1, \ldots, \Omega'_k = (\pi'_k, \phi'_k), \ldots, \Omega'_K)\} \rangle$, an irrelevant/nonsalient feature distribution $Q(\vec{X}_{ik} \mid \Omega'_k)$ is added, as shown below:

$$\begin{aligned} P(\mathcal{X} \mid \theta') = {} & \prod_{i=1}^{N} \sum_{m=1}^{M} P(m) \prod_{k=1}^{K} (P(\vec{X}_{ik} \mid \Omega_{mk}))^{\lambda_{mk}} \\ & \times (Q(\vec{X}_{ik} \mid \Omega'_k))^{(1-\lambda_{mk})} \end{aligned} \tag{7}$$

Note that, the relevance vector determines the relevance of a certain nesting, not a certain feature, as the two

features at each nesting are negatively correlated in the case of beta-binomial nestings. Afterward, following the binary feature relevance parameter, a probability relevance parameter is introduced $\rho = \{\vec{\rho}_1, \ldots, \vec{\rho}_m = (\rho_{m1}, \ldots, \rho_{mk}, \ldots, \rho_{mK}), \ldots, \vec{\rho}_M\}$, which determines the saliency of a certain feature and it is used as a measure to keep or drop the feature. Therefore, equation 7 is reformulated to the generative model equation (8) below:

$$P(\mathcal{X} \mid \Theta) = \prod_{i=1}^{N} \sum_{m=1}^{M} P(m) \prod_{k=1}^{K} [(\rho_{mk}) P(\vec{X}_{ik} \mid \Omega_{mk})]^{\lambda_{mk}} \times [(1-\rho_{mk}) Q(\vec{X}_{ik} \mid \Omega'_k)]^{(1-\lambda_{mk})} \tag{8}$$

where $\Theta = \langle \theta', \rho \rangle$. From the equation above, two main benefits can be observed:

1. The number of degrees of freedom is doubled, which gives it more flexibility and helps avoid over-fitting.
2. The salient features are dependent on the components through the generation of the components and the features consequently.

Parameters estimation

With the high amount of parameters, Table 1 shows all the parameters that are required to be estimated, where the first three are the latent parameters.

The parameters are estimated through the Expectation-Maximization (EM) method (Dempster et al., 1977), which has shown wide usage due to its effectiveness in estimating the parameters. The EM method comprises two main steps: the E-step and the M-step. Using the log-likelihood equation written below, the E-step and M-step are iteratively expecting the posterior of the latent parameters and maximizing the parameters, as in (Boutemedjet et al., 2009).

$$\mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} z_{im} \log \Big[P(m) \times \prod_{k=1}^{K} (\rho_{mk}) P(\vec{X}_{ik} \mid \Omega_{mk}))^{\lambda_{mk}} \times ((1-\rho_{mk}) Q(\vec{X}_{ik} \mid \Omega'_k))^{(1-\lambda_{mk})} \Big] \tag{9}$$

As the value of $\lambda_{mk} \in \{0, 1\}$, it is intuitive to replace it with a posterior probability conditioned on the m^{th} component. Therefore, following to u_{imk} and v_{imk} in Table 1, and by using the logarithmic properties, the log-likelihood can be rewritten as:

$$\begin{aligned} \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} z_{im} \Big[& \log[P(m)] \\ & + \sum_{k=1}^{K} u_{imk} \log[\rho_{mk}] \\ & + u_{imk} \log[P(\vec{X}_{ik} \mid \Omega_{mk})] \\ & + v_{imk} \log[(1-\rho_{mk})] \\ & + v_{imk} \log[Q(\vec{X}_{ik} \mid \Omega'_k)] \Big] \end{aligned} \tag{10}$$

E-step:

$$z_{im} = \frac{P(\vec{X}_i \mid \Theta_m)}{\sum_{m=1}^{M} P(\vec{X}_i \mid \Theta_m)} \tag{11}$$

$$u_{imk} = \frac{\rho_{mk} P(X_{ik} \mid \Omega_{mk})}{P(\vec{X}_i \mid \Theta_m)} \tag{12}$$

$$v_{imk} = z_{im} - u_{imk} \tag{13}$$

$P(m)$ follows the two following verification rules:

$$\sum_{m=1}^{M} P(m) = 1, 0 < P(m) < 1,\ m = (1, \ldots, M) \tag{14}$$

Therefore, **M-step**:

$$P(m) = \frac{\sum_{i=1}^{N} z_{im}}{N} \tag{15}$$

$$\rho_{mk} = \frac{\sum_{i=1}^{N} u_{imk}}{N} \tag{16}$$

$$\pi_{mk} = \frac{\sum_{i=1}^{N} u_{imk} x_{ik1}}{\sum_{i=1}^{N} u_{imk}} \tag{17}$$

$$\phi_{mk} = \frac{\sum_{i=1}^{N} u_{imk} (x_{ik1} - \pi_{mk})^2}{\sum_{i=1}^{N} u_{imk}} \tag{18}$$

$$\pi'_k = \frac{\sum_{i=1}^{N} \sum_{m=1}^{M} v_{imk} x_{ik1}}{\sum_{i=1}^{N} \sum_{m=1}^{M} v_{imk}} \tag{19}$$

$$\phi'_k = \frac{\sum_{i=1}^{N} \sum_{m=1}^{M} v_{imk} (x_{ik1} - \pi_{mk})^2}{\sum_{i=1}^{N} \sum_{m=1}^{M} v_{imk}} \tag{20}$$

It is noteworthy that, the value of ρ_{mk} is dependent on the number of components, therefore, it is eventually averaged throughout all the determined number of components through the MML, as will be shown next.

Model selection

The minimum message length is written as

$$\begin{aligned} MML(\mathcal{X}, \Theta) = & -\log h(\Theta) - \log P(\mathcal{X} \mid \Theta) \\ & - \frac{1}{2} \log |F(\Theta)| + \frac{N_p}{2} (1 + \log(\kappa_{N_p})) \end{aligned} \tag{21}$$

where $N_p = M(1 + 2K)$ and the value of κ_{N_p} is chosen based on the value N_p through a look-up table (Lyu et al., 2022). Moreover, h is the prior distribution, and $F(\Theta)$ is the Fisher information matrix (FIM), which can be approximated as the negative expected value of the Hessian matrix, as shown below:

$$F(\Theta) = -E[(2k)^2 \log P(\mathcal{X} \mid \Theta)] \tag{22}$$

As mentioned in (Law et al., 2004), the FIM is intractable analytically. Therefore, an approximation is used through a block-diagonal matrix of size $(M + 2K(1 + MR + S))$, where R and S are the sums of salient $(M \times K)$ and non-salient feature (K) distribution parameters, respectively. Therefore, to find the MML, the following equation is used:

$$MML(\mathcal{X}, \Theta) = \arg\min_{\theta} \Big[\frac{M + 2K}{2} \log N - \log P(\mathcal{X} \mid \Theta) + \frac{R}{2} \sum_{m=1}^{M} \sum_{k=1}^{K} \log N P(m) \rho_{mk} + \frac{S}{2} \sum_{m=1}^{M} \sum_{k=1}^{K} \log N (1 - \rho_{mk}) \Big] \tag{23}$$

Table 1: The complete set of Parameters.

Parameter	Size	Description
z_{im}	$N \times M$	$P(z_{im} = 1 \mid \vec{X}_i)$
u_{imk}	$N \times M \times K$	$P(z_{im} = 1, \lambda_{mk} = 1 \mid \vec{X}_i)$
v_{imk}	$N \times M \times K$	$P(z_{im} = 1, \lambda_{mk} = 0 \mid \vec{X}_i)$
$P(m)$	M	prior probability for m^{th} component
ρ_{mk}	$M \times K$	feature relevance probability
π_{mk}	$M \times K$	MNDD salient features mean
ϕ_{mk}	$M \times K$	MNDD salient features dispersion
π'_k	K	MNDD non-salient features mean
ϕ'_k	K	MNDD non-salient features dispersion

Experimental results

The novel model was assessed by three visual datasets, namely Natural Scenes (**NS**[2]), KTH-TIPS (**KT**[3]), and **MMI**[4], that are represented using BoVW, in addition to the UCI handwritten digits dataset (**UCI HW**[5]). The latter dataset has 64-dimensional observation vectors that are originally clustered from (32×32) bitmap pixels to a 64-dimensional (8×8) vector for each digit by calculating the sum of black pixels in a (4×4) window. Table 2 shows the details of all datasets used and the associated tree structure. Despite the visual nature of the selected datasets, the introduced model is applicable using Bag of Words (BoW) instead of BoVW.

Table 3 shows the obtained results of the enhanced MNDM with feature saliences (E-MNDM) in comparison with other models, namely, the Gaussian Mixture Model (GMM), Multinomial Mixture (MM), DCM, Multinomial Generalized Dirichlet Mixture (MGDM), and MNDM using the PDF in equation 3. The results are compared through three factors: precision, recall, and convergence time. It is noteworthy that, the relatively small selected size of the dataset observations testifies to the powerfulness of the listed statistical models, which outperform deep learning models as shown in (Alkhawaja and Bouguila, 2023). As can be inferred from the table, despite the superiority of the MNDM over most of the models, the E-MNDM is able to achieve results better than the MNDM by 5% in terms of precision and recall, approximately. This is due to the consideration of the relevant features (salient features) per component. The number of components has a lower and an upper bound for each dataset with a range of $(-2/+2)$ of the actual number of components. Therefore, the approximated number of components shown in the last column of Table 3 (Approx. M) shows the best MML (minimum) value, where each result is obtained through equation (23).

It is noteworthy that, the tree structure is crucial for the overall performance. This criticality can be inferred in the **UCI HW** dataset, where the MGDM is able to achieve better results than the MNDM and E-MNDM. Therefore, this creates an opportunity for potential enhancement to the model through the NDD tree structure optimization.

Moreover, the use of salient features naturally optimizes the determination of the number of clusters, as the features have a more solid connection to their associated component. Therefore, a simultaneous enhancement is achieved for the feature selection and the determination of the number of components, enhancing overall accuracy.

Conclusion

This work introduces a framework to exploit the advantages of the NDD and the MNDD by utilizing the concept of salient features and simultaneous determination of the number of components. As the NDD offers a hierarchical structure that enables more generalization to the covariance matrix, it can model the data at higher accuracies than the Dirichlet and the generalized Dirichlet distributions. Therefore, the augmentation of feature saliency assists the component determination criterion (MML) and creates a relationship between the features and the associated components. This framework was applied to four datasets, and its excellence was shown through different metrics in comparison with other competing models.

The use of BoW or BoVW is enough to validate the performance of the introduced model. However, despite their effectiveness, they lack semantic representation of the words and the visual keywords. On the contrary, feature representation methods such as Word embeddings or Word2Vec are capable of embedding word semantics, which could be another advancement to the obtained performance.

The promising results of this work introduce potential improvements that address the estimation method and the tree structure. According to (Boutemedjet et al., 2009), the minorization-maximization framework could lead to better results for feature selection. Moreover, as concluded from the results section, the tree structure is critical for the data representation, as it is essential for the feature relevance judgment (Null, 2008).

References

Fares Alkhawaja and Nizar Bouguila. Unsupervised nested dirichlet finite mixture model for clustering. *Applied In-*

[2]https://www.kaggle.com/datasets/zaiyankhan/15scene-dataset

[3]https://www.csc.kth.se/cvap/databases/kth-tips/download.html

[4]https://mmifacedb.eu/

[5]https://archive.ics.uci.edu/dataset/80/optical+ recognition+of+handwritten+digits

Table 2: The properties of the selected image datasets and their trees' configuration

	KT	NS	MMI	UCI HW
#Classes (M)	10	10	6	10
Resolution	200x200	250x250	750x576	32x32
Total Observations	810	2500	1140	5622
Training (N)	610	2000	700	3822
Testing	200	500	440	1800
Dimensions (D)	32	64	128	64
Nestings (K)	30	62	126	62
Nested Variables	2	2	2	2
Depth	5	6	7	6
Total Dim. (D+K)	62	126	254	126

Table 3: Obtained performances by the Clustering models for different datasets

Dataset	Model	Precision	Recall	Time(s)	Approx. M
KT	GMM	58.9%	49.7%	132.70	9
	MM	59.1%	53.4%	122.27	10
	DCM	75.1%	76.7%	156.80	11
	MGDM	76.9%	75.8%	243.10	10
	MNDM	76.8%	78.0%	395.83	10
	E-MNDM	**82.2%**	**81.5%**	731.13	10
NS	GMM	69.5%	64.4%	184.40	10
	MM	74.7%	76.4%	144.00	12
	DCM	77.4%	76.4%	207.33	10
	MGDM	82.3%	81.9%	405.39	10
	MNDM	86.8%	87.0%	437.87	10
	E-MNDM	**93.6%**	**91.9%**	961.25	10
MMI	GMM	61.1%	55.5%	20.70	7
	MM	67.7%	60.1%	18.38	5
	DCM	70.1%	65.6%	35.86	6
	MGDM	71.1%	67.4%	59.75	6
	MNDM	74.9%	76.6%	63.86	6
	E-MNDM	**79.0%**	**78.8%**	114.46	6
UCI HW	GMM	78.5%	77.6%	202.98	10
	MM	81.2%	80.7%	178.46	10
	DCM	87.8%	88.0%	316.14	10
	MGDM	95.4%	95.3%	612.84	10
	MNDM	91.3%	91.4%	662.93	10
	E-MNDM	**92.8%**	**92.3%**	1231.95	10

telligence, pages 1–27, 08 2023. doi: 10.1007/s10489-023-04888-8.

Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE transactions on information theory*, 44(6): 2743–2760, 1998.

Taoufik Bdiri, Nizar Bouguila, and Djemel Ziou. Variational bayesian inference for infinite generalized inverted dirichlet mixtures with feature selection and its application to clustering. *Appl. Intell.*, 44(3):507–525, 2016.

Manjul Bhargava. The factorial function and generalizations. *The American Mathematical Monthly*, 107(9):783–799, 2000.

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

Nizar Bouguila. Spatial color image databases summarization. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 1, pages I–953–I–956, 2007.

Nizar Bouguila. A model-based approach for discrete data clustering and feature weighting using map and stochastic complexity. *IEEE Transactions on Knowledge and Data Engineering*, 21(12):1649–1664, 2009.

Nizar Bouguila. On multivariate binary data clustering and feature weighting. *Comput. Stat. Data Anal.*, 54(1):120–134, 2010.

Nizar Bouguila and Djemel Ziou. Mml-based approach for finite dirichlet mixture estimation and selection. In Petra Perner and Atsushi Imiya, editors, *Machine Learning and Data Mining in Pattern Recognition, 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005, Proceedings*, volume 3587 of *Lecture Notes in Computer Science*, pages 42–51. Springer, 2005a.

Nizar Bouguila and Djemel Ziou. On fitting finite dirichlet mixture using ECM and MML. In Peng Wang, Maneesha Singh, Chidanand Apté, and Petra Perner, editors, *Pattern Recognition and Data Mining, Third International Conference on Advances in Pattern Recognition, ICAPR 2005, Bath, UK, August 22-25, 2005, Proceedings, Part I*, volume 3686 of *Lecture Notes in Computer Science*, pages 172–182. Springer, 2005b.

Nizar Bouguila and Djemel Ziou. Online clustering via finite mixtures of dirichlet and minimum message length. *Eng. Appl. Artif. Intell.*, 19(4):371–379, 2006.

Nizar Bouguila and Djemel Ziou. A countably infinite mixture model for clustering and feature selection. *Knowl. Inf. Syst.*, 33(2):351–370, 2012.

Nizar Bouguila and Djerriel Ziou. 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.*, pages 23–32. IEEE, 2004.

Nizar Bouguila, Khaled Almakadmeh, and Sabri Boutemedjet. A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection. *Expert Syst. Appl.*, 39(7):6641–6656, 2012.

Flavien Bouillot, Phan Nhat Hai, Nicolas Béchet, Sandra Bringay, Dino Ienco, Stan Matwin, Pascal Poncelet, Mathieu Roche, and Maguelonne Teisseire. How to extract relevant knowledge from tweets? In *Information Search, Integration and Personalization: International Workshop, ISIP 2012, Sapporo, Japan, October 11-13, 2012. Revised Selected Papers*, pages 111–120. Springer, 2013.

Sabri Boutemedjet, Nizar Bouguila, and Djemel Ziou. A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (8):1429–1443, 2009. doi: 10.1109/TPAMI.2008.155.

Kenneth W Church and William A Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.

Robert J. Connor and James E. Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969. doi: 10.1080/01621459.1969.10500963.

Manoranjan Dash, Hua Liu, and Jun Yao. Dimensionality reduction of unsupervised data. In *Proceedings ninth ieee international conference on tools with artificial intelligence*, pages 532–539. IEEE, 1997.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Samuel Y Dennis III. On the hyper-dirichlet type 1 and hyper-liouville distributions. *Communications in Statistics-Theory and Methods*, 20(12):4069–4081, 1991.

Inderjit S Dhillon and Dharmendra S Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42:143–175, 2001.

Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *Journal of machine learning research*, 5(Aug):845–889, 2004.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Tammy Harris, Joseph M Hilbe, and James W Hardin. Modeling count data with generalized distributions. *The Stata Journal*, 14(3):562–579, 2014.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. *Advances in neural information processing systems*, 18, 2005.

Xin Hong, Hailin Li, Paul Miller, Jianjiang Zhou, Ling Li, Danny Crookes, Yonggang Lu, Xuelong Li, and Huiyu Zhou. Component-based feature saliency for clustering.

IEEE transactions on knowledge and data engineering, 33(3):882–896, 2019.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Martin Law, Anil Jain, and Mário Figueiredo. Feature selection in mixture-based clustering. *Advances in neural information processing systems*, 15, 2002.

M.H.C. Law, M.A.T. Figueiredo, and A.K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, 2004. doi: 10.1109/TPAMI.2004.71.

Huawen Liu, Xindong Wu, and Shichao Zhang. Feature selection using hierarchical feature clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 979–984, 2011.

Shanxiang Lyu, Zheng Wang, Cong Ling, and Hao Chen. Better lattice quantizers constructed from complex integers. *IEEE Transactions on Communications*, 70(12): 7932–7940, 2022.

Izabele Marquetti, Jade Varaschim Link, André Luis Guimarães Lemes, Maria Brígida dos Santos Scholz, Patrícia Valderrama, and Evandro Bona. Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. *Computers and Electronics in Agriculture*, 121:313–319, 2016. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2015.12.018.

Mohamed Al Mashrgy, Taoufik Bdiri, and Nizar Bouguila. Robust simultaneous positive data clustering and unsupervised feature selection using generalized inverted dirichlet mixture models. *Knowl. Based Syst.*, 59:182–195, 2014.

Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI, 1998.

Tom Minka. The dirichlet-tree distribution. July 1999.

James E Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82, 1962.

Fatma Najar and Nizar Bouguila. Exact fisher information of generalized dirichlet multinomial distribution for count data modeling. *Information Sciences*, 586:688–703, 2022.

Brad Null. The nested dirichlet distribution: properties and applications. 11 2008.

Brad Null. Modeling baseball player ability with a nested dirichlet distribution. *Journal of Quantitative Analysis in Sports*, 5:5–5, 01 2009.

Bromensele Samuel Oboh and Nizar Bouguila. Unsupervised learning of finite mixtures using scaled dirichlet distribution and its application to software modules categorization. In *2017 IEEE international conference on industrial technology (ICIT)*, pages 1085–1090. IEEE, 2017.

Paschalis Panteleris and Antonis Argyros. Pe-former: Pose estimation transformer. In Mounîm El Yacoubi, Eric Granger, Pong Chi Yuen, Umapada Pal, and Nicole Vincent, editors, *Pattern Recognition and Artificial Intelligence*, pages 3–14, Cham, 2022. Springer International Publishing. ISBN 978-3-031-09282-4.

Sudhir R. Paul, Uditha Balasooriya, and Tathagata Banerjee. Fisher information matrix of the dirichlet-multinomial distribution. *Biometrical Journal*, 47(2):230–236, 2005. doi: https://doi.org/10.1002/bimj.200410103.

Beatriz Remeseiro and Veronica Bolon-Canedo. A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, 112:103375, 2019. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2019.103375.

Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81(10.5555):26853, 1986.

Nonita Sharma, Monika Mangla, Sachi Nandan Mohanty, and Suneeta Satpaty. A stochastic neighbor embedding approach for cancer prediction. In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 599–603, 2021. doi: 10.1109/ESCI50559.2021.9396902.

Saúl Solorio-Fernández, J. Ariel Carrasco-Ochoa, and José Francisco Martínez-Trinidad. A survey on feature selection methods for mixed data. *Artif. Intell. Rev.*, 55(4):2821–2846, apr 2022. ISSN 0269-2821. doi: 10.1007/s10462-021-10072-6.

Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:1285–1297, 2022. doi: 10.1109/TIP.2022.3140606.

David L. Weakliem. A critique of the bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3):359–397, 1999.

Petros Xanthopoulos, Panos M Pardalos, Theodore B Trafalis, Petros Xanthopoulos, Panos M Pardalos, and Theodore B Trafalis. Linear discriminant analysis. *Robust data mining*, pages 27–33, 2013.

Nuha Zamzami and Nizar Bouguila. A novel minorization–maximization framework for simultaneous feature selection and clustering of high-dimensional count data. *Pattern Analysis and Applications*, 26, 07 2022. doi: 10.1007/s10044-022-01094-z.

Rui Zhang, Feiping Nie, and Xuelong Li. Auto-weighted two-dimensional principal component analysis with robust outliers. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6065–6069, 2017. doi: 10.1109/ICASSP.2017.7953321.