Received: 15 February 2017 | Revised: 21 November 2017 | Accepted: 2 December 2017 | Published on: 10 January 2018

DOI: 10.1002/sam.11369

ORIGINAL ARTICLE

WILEY

Clustering over-dispersed data with mixed feature types

Lu Wang | **Dongxiao Zhu** | **Ming Dong**

Department of Computer Science, Wayne State University, Detroit, Michigan

Correspondence
Dongxiao Zhu, Department of Computer Science, Wayne State University, 5057 Woodward Avenue, Detroit, MI 48202.
Email: dzhu@wayne.edu

Funding Information
Division of Computing and Communication Foundations, 1451316. National Science Foundation, CCF-1451316, CNS-1637312.

Despite many data clustering methods are available, most of them uncover compactness or connectivity as the intrinsic structure of unlabeled data. Very few approaches explicitly consider the cluster size distribution, especially over-dispersed (high variance), which may represent yet another important aspect of structural information of unlabeled data. In this paper, we propose a novel joint mixture model framework to estimate cluster size distribution together with cluster compactness (density). Our framework is sufficiently flexible and general to capture a wide range of cluster size distributions from data with mixed feature types. Experiments on clustering synthetic and real-world data demonstrate a superior performance of our clustering approach in recovering the hidden structure of unlabeled data.

KEYWORDS

clustering, data mining, mixed feature-type data, over-dispersed cluster size distribution, unlabeled data

1 | INTRODUCTION

Recent years have seen a sharp increase in the volume of data with imbalanced and overlapping clusters. This type of data has been intensively studied in supervised learning [17,18] and briefly touched upon in semi-supervised learning [27,36] but barely in unsupervised learning. The latter represents a more challenging task of uncovering the intrinsic structures in unlabeled data. Intuitively, clusters not only differ in data compactness and connectivity, but also can differ dramatically in their sizes. Distribution of cluster sizes, particularly over-dispersed with high variance, may shed light on the hidden structure of unlabeled data.

Many existing data clustering approaches are effective in uncovering overlapping clusters from noisy unlabeled data; however, they focus more on detecting data compactness or connectivity whereas overlooking the cluster sizes as independent yet discriminative information on the structure of unlabeled data. In real-world scenario, over-dispersed cluster size distribution with high variance is ubiquitous. In addition, these data are often available with mixed feature types, such as continuous and categorical.

For instance, in medical science, the prevalence of breast cancer subtypes can be quite diverse and identification of these subtypes, regardless prevalence, are equally important [26]. In environmental science, the distribution of species abundances, represented by imbalanced clusters, describes key elements of biodiversity [20,42]. In social science, an actively studied topic is group imbalance, which measures the skewness of the group sizes [10,43]. In these applications, over-dispersion (high variance) in cluster size distribution represents a rich source of previously unattended information, which needs to be appropriately and explicitly modeled when developing an effective data clustering approach.

There are 2 main optimizing criterion for data clustering algorithms: data compactness used in K-means and mixture models [2,12] and data connectivity used in spectral clustering [33]. Both types of algorithms are effective in uncovering data clusters using structural information, either density or connectivity. However, cluster size distribution as yet another intrinsic data structure was often neglected.

The classical finite mixture model is one of the most commonly used probabilistic clustering methods to capture data compactness (density) [31]. Besides data compactness (density), the classical finite mixture model implicitly captures normally distributed cluster size by using multinomial distribution as we will show in Section 3.1, the latter asymptotically converges to normal distribution [11]. However, cluster size distribution in real-world data is frequently over-dispersed and heavily deviated from normal distribution. Using classical finite mixture model for capturing over-dispersed cluster size distribution may represent a significant misrepresentation of the intrinsic structure of the unlabeled data.

To accommodate over-dispersed cluster size distribution frequently seen in real-world data, we use Poisson distribution as an appropriate probabilistic model. Although several other probability models may also work with over-dispersed cluster size distribution to different extents, Poisson distribution is a more sensible choice. For example, Laplace distribution shares the similar core function with normal distribution in their probability density function [4], which limits its capability of accommodating over-dispersion. In addition, being a symmetrical distribution around the mean, it does not account for skewness in cluster size distribution.

In this paper, we propose a novel probabilistic framework, SizeDensity, to simultaneously model data compactness (density) and cluster size distribution. Specifically, we develop new joint mixture probability models and efficient algorithms to uncover the imbalanced and overlapping clusters from mixed feature-type data with over-dispersed size distributions. We focus on mathematical and algorithmic formulation, validation, and evaluation of integrating cluster size information into compactness-based clustering algorithms. The problem is sufficiently general and important but has barely been studied in the literature.

Our original contribution is to incorporate the cluster size distribution, especially over-dispersed, into the probabilistic model as an independent component from the class conditional density. By employing the class indicator as a latent variable, we derive and maximize a new likelihood function to simultaneously estimate cluster size distribution and class conditional density. The advantages of SizeDensity framework over the traditional DensityOnly framework described above are duly demonstrated in Figure 1.

The rest of the paper is organized as follows. In Section 2, we review the related works in data clustering. In Section 4, we describe our new SizeDensity clustering framework in details. In Section 5, we present experimental results using both synthetic data and real-world data and compare with the selected clustering methods. Finally, we conclude with discussion in Section 6.

2 | RELATED WORKS

Clustering is unsupervised partitions of instances into classes where instances within the same class are more similar to each other than those from different classes. Clustering has played a central role in many data-rich science domains, such as biological, social, physical, and medical sciences [22]. Existing clustering approaches have been largely focusing on discovering the cluster membership of each instance and class conditional density. When no well-defined clusters exist a priori, inferring the correct number of clusters is also critical for the purpose of knowledge discovery [12].

Many clustering methods are based on an evaluation of pairwise dissimilarities between the instances. Earlier methods that optimize data compactness, including hierarchical clustering, condensation-based clustering [46], K-means type of clustering [23] and self-organizing map [25], effectively partition the instances into different clusters according to the differences in their means. They run fast, and the results are easy to be visualized. However, a number of key limitations exist in the traditional data compactness-based approaches: (1) they only detect mutually exclusive clusters that an instance can only be partitioned into 1 cluster at a time [44], (2) they only consider mean of the instances but not variance/covariance of the instances within a cluster by assuming equal variance/covariance across the clusters, and (3) they neglect cluster size distribution, especially over-dispersed, as an intrinsic structure of real-world unlabeled data.

One type of probabilistic clustering methods are frequentist (likelihood-based) approaches [34], which use the cluster indicator matrix as a latent variable, and infer the expected values of this matrix by maximizing a likelihood function. In the latter one, the mean and variance of each cluster as well as their class weight are clearly defined and given. The Expectation-Maximization (EM)-type algorithm [1,45] is often used to iteratively maximize the likelihood function to estimate the class conditional parameters. It is rational when the number of clusters is known a priori and cluster size distribution is truly normal. Another type of probabilistic clustering methods is based on Bayesian Infinite Mixture Model [32,38], also known as Dirichlet process mixture model. Instead of getting a "right answer" of the density parameters in class conditional density, their distributions are learned by sampling from the posterior distributions. This type of approach works better if the number of clusters is unknown and a representation of the modeling uncertainty is desired.

In addition to compactness (density)-based clustering methods, there are other effective clustering approaches, such as connectivity-based spectral clustering [6,33,36,39], grid-based clustering [3], affinity propagation [14], subspace clustering [24], and so on. Likewise, these approaches are not designed for uncovering the over-dispersed cluster size distribution, which is frequently seen, and may shed light on the intrinsic structure of unlabeled real-world data.

To our knowledge, there are few data clustering approaches that consider over-dispersed cluster size distribution as an intrinsic structure of the data. In ref. [13], cluster sizes were estimated in conjunction with cluster shapes subject to the stringent and unrealistic orthogonality assumption in eigen-decomposition of the covariance matrix. Moreover, such decomposition algorithm is limited to continuous data with simple cluster shapes and needs multiple starts. In refs. [19,47], cluster sizes were assumed to be known and incorporated into the clustering models as an extrinsic constraint as opposed to an intrinsic structure. Thus, new probabilistic clustering methods for uncovering both class conditional density and over-dispersed cluster size distribution are urgently needed to effectively overcome all the 3 aforementioned limitations.

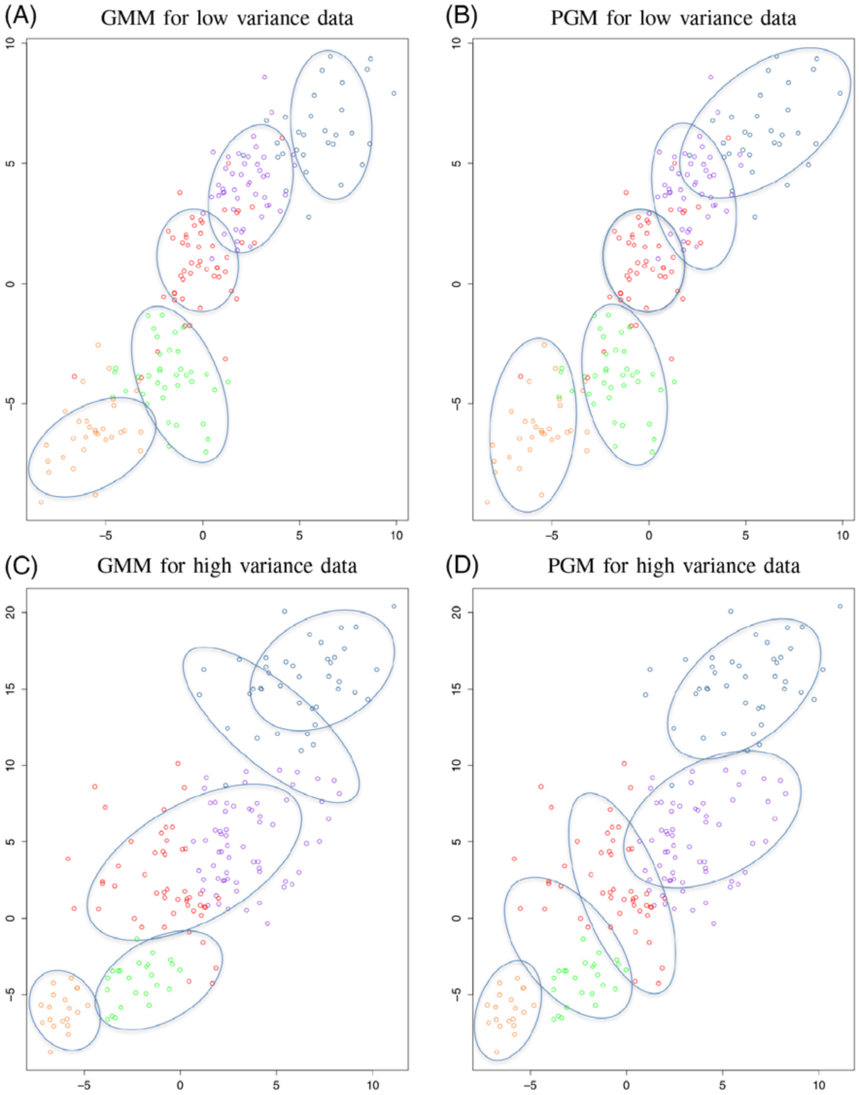


FIGURE 1 Top row represents the low variance data set, while bottom row represents the high variance data set. Both approaches perform well in the low variance data set. However, the traditional DensityOnly (Gaussian mixture model, GMM) approach (left column) does not accurately detect clusters with over-dispersed (high variance) cluster size distribution in the high variance data set whereas the proposed SizeDensity (Poisson Gaussian mixture model, PGM) approach (right column) does. The 5 true data clusters are shown in different colors and the corresponding ellipses are calculated from the density estimation of each cluster in both simulated data sets using DensityOnly and SizeDensity approaches, respectively. These 2 data sets are simulated using multivariate Gaussian distribution in R package MASS [40] and their cluster sizes are simulated using multinomial distribution

3 | PRELIMINARY: THE CLASSICAL FINITE MIXTURE MODELS

3.1 | The plain version of classical finite mixture model

The density for the classical mixture model with K components is defined as:

$$f(X) = \sum_{k=1}^{K} \pi_k f_k(X), \tag{1}$$

where X is a set of input data, π_k is the prior probability of a mixing proportion subject to $\sum_{k=1}^{K} \pi_k = 1$, and $f_k(y)$ is the density of a mixture component.

In classical finite mixture model, class conditional density $f_k(X)$ is often assumed to be multivariate normal [31]. A probability model is developed by introducing an indicator variable Z_{ik} for clustering n instances, where $Z_{ik} = 1$ represents the ith instance belonging to the kth cluster and 0 otherwise. Given a set of input data X_1, X_2, ..., X_n, the likelihood function of the indicator variable Z_{ik} is written as:

$$L(\Theta) = \prod_{i=1}^{n} f(X_i; \Theta)^{Z_{ik}}, k = 1, \ldots, K, \tag{2}$$

where Θ is a set of the mixture model-related parameters.

The EM algorithm, for estimating the parameter values that maximizes the above likelihood function, iterates the following computations until convergence:

E-step:

$$\tau_{ik}^{(l)} = \frac{\pi_k f_k(X_i))}{f(X_i))}, \tag{3}$$

where τ_{ik} is the expectation value of Z_{ik} and l is the iteration index.

M-step:

$$\pi_k^{(l+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}. \tag{4}$$

3.2 | The updated classical finite mixture model with independent component of modeling cluster size

In Section 3, Z_{ik} does not assume to follow any distribution so that the above model is DensityOnly for modeling cluster compactness (density). However, we here show the classical approach implicitly models the cluster size by assuming Z_{ik} follows a multinomial distribution. The likelihood function is written as:

$$L(\Theta) = \prod_{i=1}^{n} \left[\pi_k n! \prod_{k=1}^{K} \frac{\pi_k^{n_k}}{n_k!} f(X_i; \Theta) \right]^{Z_{ik}}, \tag{5}$$

where n_k is the number of instances in each cluster.

E-step:

$$\begin{aligned} \tau_{ik}^{(l)} &= E[Z_{ik} = 1 \mid \pi_k, \Theta] = p(Z_{ik} = 1 \mid \pi_k, \Theta) \\ &= \frac{p(Z_{ik} = 1)\pi_k f(X; \Theta)}{\sum_{k\prime=1}^{K} p(Z_{ik\prime} = 1)\pi_{k\prime} f(X; \Theta)}. \end{aligned} \tag{6}$$

M-step:

$$\pi_k^{(l+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik} = \frac{n_k}{n}. \tag{7}$$

The newly derived $\pi_k^{(l+1)}$ modeling Z_{ik} in Equation (7) is the same as the one without modeling Z_{ik} in Equation (4), due to multinomial distribution asymptotically converges to normal distribution [11]. Therefore, these 2 classical finite mixture models with or without modeling Z_{ik} are equivalent which perform well for normal cluster size distribution. However, the above 2 models do not perform well for data with over-dispersed cluster size distribution, because of strong deviation of over-dispersed cluster size distribution from normal distribution.

TABLE 1 List of mathematical symbols

Notations	Comments
n	Total number of instances in a data set
n_k	Number of instances in each cluster
K	Number of clusters
k	Index of the clusters ∈ [1, ..., K]
$f_k(X)$	Density of the mixture component
X	Continuous features
X_i	*i*th instance's continuous feature' value
i	Index of the instances∈ [1, ..., n]
σ_k	Covariance matrix in cluster k
Y	Categorical features
Y_j	*j*th categorical feature
j	Index of the categorical feature
J	Number of categorical features
L_h	*h*th level in each categorical feature
h	Index of the levels in each categorical
H	Number of levels in each category
$n_{k_{Y_j L_h}}$	Number of instances in the *k*th cluster with the
	*h*th level in the *j*th categorical feature
$p^{n_{k_{Y_j L_h}}}$	$p^{n_{k_{Y_j L_h}}}$ probability that given extraction will be in the
	*k*th cluster with the *h*th level in *j*th categorical feature
D	Dummy variable
π_k	Mixing proportion
λ_k	Component parameter of Poisson model
Θ	A set of mixture model related parameters
Φ	PGMM-related parameters
Z_{ik}	Indicator that instance belongs to *k*th cluster
Z_i	Indicator of *i*th instance
F_i	Z_i's categorical distribution
τ_{ik}	Expectation value of Z_{ik}
l	Iteration index

4 | METHODOLOGY

4.1 | Poisson Gaussian multinomial mixture model for mixed feature-type data

Given a data set with continuous features X and categorical features Y with K clusters, and there are H different levels for each categorical feature. Using mathematical symbols listed in Table 1, the joint probability-based model for the mixed continuous and categorical data can be written as:

$$p(X, Y, Z|\pi, \Theta) = \sum_{k=1}^{K} \pi_k p(X|\Theta_X, \lambda_k) p(Y|\Theta_Y, \lambda_k). \tag{8}$$

As stated above, clusters not only differ in class conditional density but also in their sizes especially over-dispersed. Cluster size distribution, representing an intrinsic structure of unlabeled data, is often of practical interest together with cluster compactness or connectivity. Hence, we develop a novel probability model for clustering considering both class conditional density and cluster size distribution. The likelihood function for the new SizeDensity model with Poisson distribution representing the over-dispersed cluster size distribution in the mixed feature-type data is given as:

$$\begin{aligned} L_{\mathrm{PGMM}}(X, Y, Z|\pi_k, \Theta, \lambda_k) = \prod_{i=1}^{n} \prod_{k=1}^{K} \Bigg(\frac{\lambda_k^{n_k} e^{-\lambda_k}}{n_k!} \pi_k \\ \frac{1}{\sqrt{2}\pi\sigma_k} e^{-\frac{(X_1-\mu_k)(X_i-\mu_k)^{\intercal}}{2\sigma_k^2}} \prod_{j=1}^{J} \left[n_k! \left(\prod_{h=1}^{H} \frac{p_{k_{Y_j L_h}}^{n_{k Y_j L_h}}}{n_{k Y_j L_h}} \right) \right] \Bigg)^{Z_{ik}}, \end{aligned} \tag{9}$$

where $n_{kY_jL_h}$ is the number of instances in the kth cluster with the hth level in the jth categorical feature, $p^{n_{k_{Y_j L_h}}}$ is the probability that given extraction will be in the kth cluster with the hth level in the jth categorical feature and λ_k is a parameter representing cluster size.

Since we used Poisson distribution to capture the mean and variance of the over-dispersed cluster size distribution, we denote the SizeDensity model as Poisson Gaussian multinomial mixture model (PGMM). Note that the SizeDensity model is sufficiently flexible and can be extended to employ other discrete distributions for modeling a wide range of cluster size distributions in real-world data. Figure 2 presents the main idea of this work.

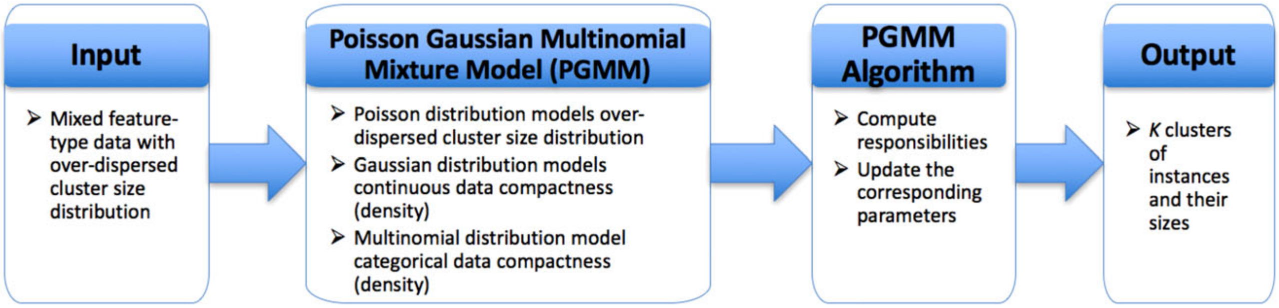

Input
➢ Mixed feature-type data with over-dispersed cluster size distribution
Poisson Gaussian Multinomial Mixture Model (PGMM)
➢ Poisson distribution models over-dispersed cluster size distribution
➢ Gaussian distribution models continuous data compactness (density)
➢ Multinomial distribution model categorical data compactness (density)
PGMM Algorithm
➢ Compute responsibilities
➢ Update the corresponding parameters
Output
➢ K clusters of instances and their sizes

FIGURE 2 A conceptual overview of the Poisson Gaussian multinomial mixture model (PGMM) for clustering mixed feature-type data with over-dispersed cluster size distribution

4.2 | The PGMM algorithm

Here, we develop an EM-type algorithm to maximize the complete data log-likelihood function $\log L_{\mathrm{PGMM}}$. The expected value of Z_{ik} is τ_{ik}, where Z_{ik} is a latent variable indicating whether the instance i belongs to kth cluster:

$$\begin{aligned}\tau_{ik}^{(l)} &= E[Z_{ik}=1 \mid \pi_k^{(l)}, \lambda_k^{(l)}, \Theta^{(l)}]\\ &= p(Z_{ik}=1 \mid \pi_k^{(l)}, \lambda_k^{(l)}, \Theta^{(l)})\\ &= \frac{p(N=n_k^{(l)})\pi_k^{(l)} p(x_X\Theta_x^{(l)}, \lambda_k^{(l)}) p(y_Y|\Theta_y^{(l)}, \lambda_k)^{(l)}}{\sum_{k\prime=1}^{K} p(N=n_{k\prime})\pi_{k\prime}^{(l)} p(x_X|\Theta_x^{(l)}, \lambda_{k\prime}^{(l)}) p(y_Y|\Theta_y^{(l)}, \lambda_{k\prime}^{(l)})}.\end{aligned} \tag{10}$$

Real-world data may also contain dummy variables, which are the binary variables with value either 1 or 0 indicating whether the specific instance is present or absent [15]. For example, gender is typically considered as a dummy variable in a medical data set. In our model, we use them to more accurately model the conditional distributions of other features rather than using them as features. Hence we arrive at the updated expected value of Z_{ik} for the data with dummy variables as:

$$\begin{aligned}\tau_{ik}^{(l)} &= E[Z_{ik}=1, D|\pi_k^{(l)}, \lambda_k^{(l)}, \Theta^{(l)}]\\ &= p(Z_{ik}=1, D=1|\pi_k^{(l)}, \lambda_k^{(l)}, \Theta^{(l)})\\ &= \frac{p(N=n_k)^{(l)}\pi_k^{(l)} p(x_X|\Theta_x^{(l)}, \lambda_k^{(l)}) p(y_Y|\Theta_y^{(l)}, \lambda_k^{(l)}) p(D=1)}{\sum_{k\prime=1}^{K} p(N=n_{k\prime}^{(l)})\pi_k^{(l)} p(x_X|\Theta_x^{(l)}, \lambda_{k\prime}^{(l)}) p(y_Y|\Theta_y^{(l)}, \lambda_{k\prime}^{(l)}) p(D{=}1)},\end{aligned} \tag{11}$$

where D represents a dummy variable. It equals to 1 when the specific instance is present and 0 otherwise. For example, in the dummy variable gender, 1 represents a male instance whereas 0 represents a female instance.

We calculate $Q(\Theta|\Theta^{[l]})$ of SizeDensity (PGMM) model as:

$$\begin{aligned}Q(\Theta|\Theta^{(l)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(l)} \Bigg\{ &\log\left[\frac{(\lambda_k^{(l)})^{n_k^{(l)}} e^{-\lambda_k^{(l)}}}{n_k^{(l)}!}\right]\\ &+ \log \pi_k^{(l)} + \log\left[\frac{1}{\sqrt{2\pi}\sigma_k^{(l)}} e^{-\frac{(X_i-\mu_k^{(l)})(X_i-\mu_k^{(l)})^{\intercal}}{2(\sigma_k^{(l)})^2}}\right]\\ &+ \log\left[\prod_{j=1}^{J} n_k^{(l)}! \left(\prod_{h=1}^{H} \frac{(p_{k_{Y_jL_h}}^{(l)})^{n_{k_{Y_jL_h}}^{(l)}}}{n_{k_{Y_jL_h}}^{(l)}!}\right)\right]\Bigg\},\end{aligned} \tag{12}$$

where $\log^{(l)}$ with parameter means the log value of the lth iteration.

In the M-step, we find the parameter values that maximize the $Q(\Theta|\Theta^{[l]})$. π_k is updated by summarizing the expected counts of instances as:

$$\pi_k^{(l+1)} = \sum_{i=1}^{n} \frac{\tau_{ik}^{(l)}}{n}. \tag{13}$$

The Gaussian component parameters are updated as:

$$\mu_k^{(l+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(l)} x_{Xi}}{\pi_k^{(l)}}, \tag{14}$$

$$\begin{aligned}\Sigma_k &= \frac{\sum_{i=1}^{n} \pi_{ik}(X_i-\mu_k)(X_i-\mu_k)^{\intercal}}{\pi_k}\\ &= \frac{\sum_{i=1}^{n} \pi_{ik} X_i X_i^{\intercal}}{\pi_k} - \mu_k \mu_k^{\intercal},\end{aligned} \tag{15}$$

where μ_k and Σ_k are vector of means and covariance matrix for continuous feature, respectively.

The multinomial component parameters are updated as:

$$p_{k_{Y_jL_h}}^{(l+1)} = \frac{n_{k_{Y_jL_h}}^{(l)}}{n_k^{(l)}}, \tag{16}$$

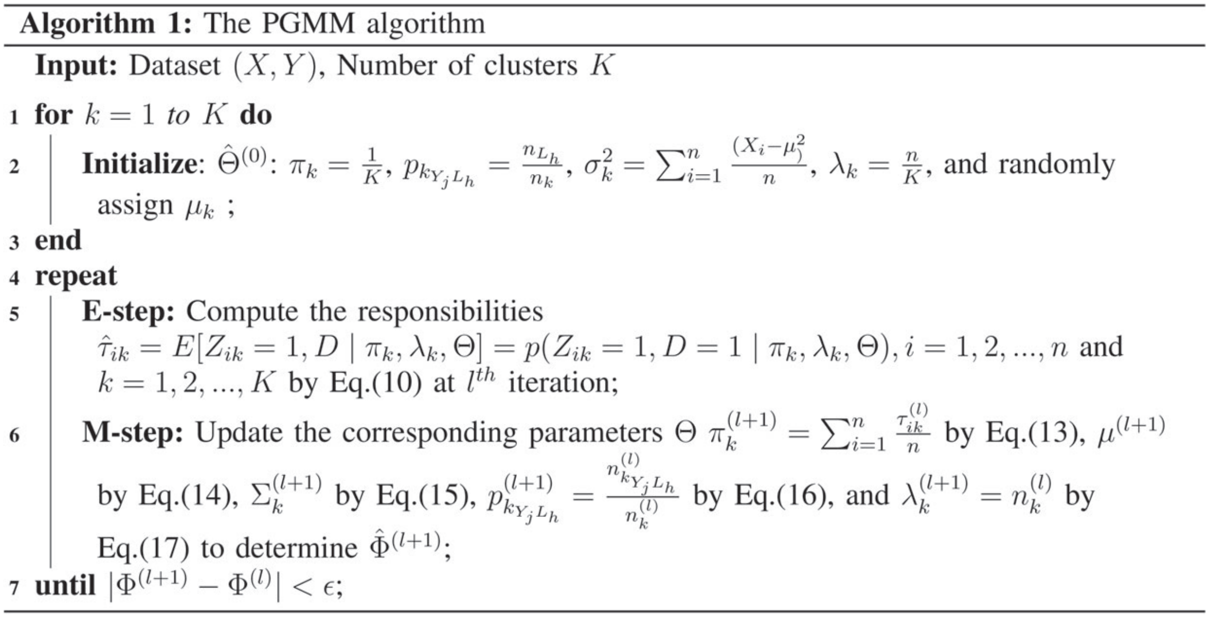
and the Poisson component parameter λ_k is estimated by calculating the first derivative of $Q(\Theta|\Theta^{[l]})$ as:

$$\lambda_k^{(l+1)} = n_k^{(l)}. \tag{17}$$

For the initialization of the parameters, each cluster is given an equal size, and each class is given an equal weight at the beginning. That is, $\lambda_k = \frac{n}{K}$ for cluster size and $\pi_k = \frac{1}{K}$ for the mixing proportion. We also set $p_{k_{Y_jL_h}} = \frac{n_{L_h}}{n_k}$ for cluster density of the categorical features. For cluster density of the continuous features, we set $\sigma_k^2 = \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{n}$ and randomly assign the values of μ_k. The complete PGMM method is given in Algorithm 1.

After the PGMM algorithm converges, we assign each instance to a cluster with the highest probability among all clusters according to the indicator matrix Z_{ik}, calculated as follows:

$$p(Z_{ik}=1|X, Y, \widehat{\Theta}) = \frac{\widehat{\pi}_k p(X_i, Y_j|\widehat{\lambda}_k, \widehat{\mu}_k \widehat{\Sigma}_k)\widehat{p}_{k_{Y_jL_h}}}{\sum_{k\prime=1}^{K} \widehat{\pi}_{k\prime} p(X_i, Y_j|\widehat{\lambda_{k\prime}}, \widehat{\mu}_{k\prime} \widehat{\Sigma}_{k\prime})\widehat{p}_{k\prime_{Y_jL_h}}}.$$



```
Algorithm 1: The PGMM algorithm
  Input: Dataset $(X, Y)$, Number of clusters $K$
for $k = 1$ to $K$ do
    Initialize: $\hat{\Theta}^{(0)}$: $\pi_k = \frac{1}{K}$, $p_{k_{Y_j L_h}} = \frac{n_{L_h}}{n_k}$, $\sigma_k^2 = \sum_{i=1}^{n} \frac{(X_i-\mu^2_{)}}{n}$, $\lambda_k = \frac{n}{K}$, and randomly
      assign $\mu_k$ ;
end
repeat
    E-step: Compute the responsibilities
      $\hat{\tau}_{ik} = E[Z_{ik} = 1, D \mid \pi_k, \lambda_k, \Theta] = p(Z_{ik} = 1, D = 1 \mid \pi_k, \lambda_k, \Theta), i = 1, 2, ..., n$ and
      $k = 1, 2, ..., K$ by Eq.(10) at $l^{th}$ iteration;
    M-step: Update the corresponding parameters $\Theta$ $\pi_k^{(l+1)} = \sum_{i=1}^{n} \frac{\tau_{ik}^{(l)}}{n}$ by Eq.(13), $\mu^{(l+1)}$
      by Eq.(14), $\Sigma_k^{(l+1)}$ by Eq.(15), $p_{k_{Y_j L_h}}^{(l+1)} = \frac{n_{k_{Y_j L_h}}^{(l)}}{n_k^{(l)}}$ by Eq.(16), and $\lambda_k^{(l+1)} = n_k^{(l)}$ by
      Eq.(17) to determine $\hat{\Phi}^{(l+1)}$;
until $|\Phi^{(l+1)} - \Phi^{(l)}| < \epsilon$;
```

4.3 | On convergence of the PGMM algorithm

In this section, we theoretically study the convergence of the PGMM algorithm. To this end we claim that the proofs given below can be applied to any algorithms falling into our SizeDensity framework. Formally, we need to show that the log-likelihood of PGMM denoted as $L_{\text{PGMM}}(\Phi)$ converges monotonically to an unique log-likelihood value $L_{\text{PGMM}}(\Phi)^*$, where Φ represents the set of PGMM related parameters. Our mathematical proof of PGMM convergence is written as below:

Proposition 1. For a convex function $f(x)$, $E[f(X)] \geq f(E[X])$ provided that the expectations exist and are finite. For a strictly convex function, $E[f(X)] = f(E[X])$ if only if $p(x = E[X]) = 1$. For concave function $f(x)$, $E[f(X)] \leq f(E[X])$ provided that the expectations exist and are finite (Jensen's inequality).

Given data with mixed feature types, we write the log-likelihood function of PGMM as:

$$L_{\text{PGMM}}(\Phi) = \sum_i \sum_k \log p(X, Y|\Phi) = \sum_i \sum_k \log \sum_{Z_i} p(X, Y, Z|\Phi), \tag{18}$$

where Z_i is the class indicator of ith instance, and k is the class index. Assuming Z_i follows a categorical distribution denoted as F_i, the log-likelihood of PGMM can be shown as:

$$L_{\text{PGMM}}(\Phi) = \sum_i \sum_k \log \sum_{Z_i} F_i(Z_i) \frac{p(X_i, Y_i, Z_i|\Phi)}{F_i(Z_i)}, \tag{19}$$

where X_i and Y_i are the continuous and categorical feature values of ith instance, respectively.

Proposition 2. Assume the continuous and categorical data distributions are from canonical exponential families, in their natural parameterization, $L_{\text{PGMM}}(\Phi)$ is a concave function [7].

Proposition 3. If random variable X follows categorical distribution denoted as $g(X)$ and its probability mass function can be donated as f_X. Then the expected value of $g(X)$ is $E[g(X)] = \sum_x g(x) f_X(x)$ [9].

Lemma 1. For each data instance, $L_{\text{PGMM}}(\Phi) \geq \sum_i \sum_k \sum_{Z_i} F_i(Z_i) \log \frac{p(X_i,Y_i,Z_i|\Phi)}{F_i(Z_i)}$.

Proof. According to Proposition 3, the expectation value of $\frac{p(X_i,Y_i,Z_i|\Phi)}{F_i(Z_i)}$ is $\sum_{Z_i} F_i(Z_i) \log \frac{p(X_i,Y_i,Z_i|\Phi)}{F_i(Z_i)}$. Combining Propositions 1, 2, and 3, we can get:

$$f\left\{E_{Z_i \sim F_i}\left[\frac{p(X_i, Y_i, Z_i|\Phi)}{F_i(Z_i)}\right]\right\} \geq E_{Z_i \sim F_i}\left\{f\left[\frac{p(X_i, Y_i, Z_i|\Phi)}{F_i(Z_i)}\right]\right\}, \tag{20}$$

so that Lemma 1 is proved. □

Theorem 1. Given $L_{\text{PGMM}}(\Phi)$ is a concave function for continuous and categorical data distributions from canonic exponential families, $L_{\text{PGMM}}(\Phi)^{(l+1)} \geq L_{\text{PGMM}}(\Phi)^{(l)}$.

Proof. In E-Step, for each instance i, compute $F_i(Z_i) = p(Z_i|X_i, Y_i|\Phi)$. Then, the log-likelihood of PGMM is written as:

$$L_{\text{PGMM}}(\Phi^{(l)}) = \sum_i \sum_k \sum_{Z_i} F_i^{(l)}(Z_i) \log \frac{p(X_i, Y_i, Z_i|\Phi^{(l)})}{F_i^{(l)}(Z_i)}. \tag{21}$$

In M-Step, compute

$$\Phi =* \arg \max_{\Phi} \sum_{i=1} \sum_{k=1} \sum_{Z_i} F_i(Z_i) \log \frac{p(X_i, Y_i, Z_i|\Phi)}{F_i(Z_i)}. \tag{22}$$

And the log-likelihood of PGMM is rewritten as:

$$\begin{aligned} L_{\mathrm{PGMM}}(\Phi^{(l+1)}) &\geq \sum_{i} \sum_{k} \sum_{Z_i} F_i^{(l)}(Z_i) \log \frac{p(X_i, Y_i, Z_i|\Phi^{(l+1)})}{F_i^{(l)}(Z_i)} \\ &\geq \sum_{i} \sum_{k} \sum_{Z_i} F_i^{(l)}(Z_i) \log \frac{p(X_i, Y_i, Z_i|\Phi^{(l)})}{F_i^{(l)}(Z_i)} \\ &= L_{\mathrm{PGMM}}(\Phi^{(l)}), \end{aligned} \tag{23}$$

so that $L_{\mathrm{PGMM}}(\Phi)^{(l+1)} \geq L_{\mathrm{PGMM}}(\Phi)^{(l)}$ is proved. □

Thus, $L_{\mathrm{PGMM}}(\Phi)$ is nondecreasing over iterations. At $(l+1)$ iterations, $|L_{\mathrm{PGMM}}(\Phi)^{(l+1)} - L_{\mathrm{PGMM}}(\Phi)^{(l)}| \leq \epsilon$, where ϵ is an arbitrarily small number greater than 0. So, the convergence of PGMM is proved.

5 | EXPERIMENTS AND RESULTS

To evaluate the performance of our proposed PGMM (SizeDensity) algorithm, we comprehensively compared our algorithm with other 7 popular algorithms using a total of 9 data sets, including 4 synthetic data sets, 4 Heart Disease data sets, and 1 Walmart Recruiting data set.

5.1 | Experimental setup

In our experiments, all the algorithms were implemented in R language. Due to the heuristic nature of the clustering algorithms, we ran each algorithm multiple times using different parameter values attempting to report their best performance. We ran all algorithms 10 times on the 4 synthetic and 4 heart disease data sets and we ran all the algorithms 3 times on Walmart Recruiting data set due to extremely large data volume.

The 7 methods selected for comparison can be summarized into the following 3 categories:

- **Model-based clustering**: We selected 3 model-based clustering algorithms, that is, DensityOnly (Gaussian Multinomial Mixture model, GMM), SizeOnly (Poisson Mixture Model, PMM), and Dirichlet Process Mixture Model (DPMM). R package *mixtools* [5] was used to implement GMM and PMM. Please refer to Appendixes 1 and 2 for details of GMM and PMM models, respectively. DPMM was implemented in *PReMiuM* package [29]. In each run, the concentration parameter of Dirichlet process was set to be a nonnegative random number in the range of 0.001 to 1. In addition, the number of iterations in the burn-in period of the Markov Chain Monte Carlo (MCMC) sampling as well as the number of sweeps after the burn-in period were set in the range of 10 to 1000.
- **Distance-based clustering**: We select 4 clustering methods based on different distance metrics, which are density peak clustering (DPC)and 3 hierarchical clustering methods. DPC algorithm was implemented in R package [37]. The hierarchical clustering algorithms were implemented using *cluster* package [30] with Gower's distance [16]. In our experiments, we explored all the 3 commonly used group similarity measurements, that is, group average, single link, and complete link, corresponding to Hierarchical Clustering Average (HC-A), Hierarchical Clustering Single (HC-S), and Hierarchical Clustering Complete (HC-C) algorithms.
- **Spectral clustering**: In our experiments, due to the mixed feature types, we generated Gower's similarity matrix using *CluMix* package [35] and then plugged it into *SNFtool* [41] to perform spectral clustering. In each run, we tried different weights to calculate Gower's distance between objects in R package *CluMix*.

5.2 | Synthetic data

We generated 4 synthetic data sets using R with various cluster size distributions to evaluate the performance of our clustering methods and compare with selected other clustering methods. The 4 synthetic data sets are generated based on the parameters learned from lung cancer outcomes study from Institute for Digital Research and Education of UCLA [8].

Each synthetic data set contains a single label (target) representing the 5 stages of lung cancer, 7 continuous features (age, length of stay in hospital after surgery, white blood count, red blood count, body mass index, interleukin 6, and C-reactive protein) and 4 categorical features (married, family hx, smoking hx and gender). In our PGMM model, we treated gender as a dummy variable instead of a categorical feature.

In ref. [28], the authors examined means and variances of continuous features, for example, body mass index and interleukin 6 for patients in various cancer stages. With these means and variances, Gaussian distribution is used to simulate continuous features. Binomial and multinomial distributions are used for binary features and multicategorical features, respectively. Cluster sizes are simulated using multinomial distribution.

We compared the performance of the 8 clustering methods in terms of adjusted Rand index (ARI) [21] (please see the Appendix 3 for the calculation of ARI). To highlight the unique advantage of PGMM in recovering the over-dispersed (high variance) cluster size distribution, we designed a set of case-control experiments. Specifically, we used a true cluster size distribution with over-dispersion (high variance) to generate the data sets 1 and 2, and a true cluster size distribution with low variance to generate the data sets 3 and 4.

In Table 2, we report ARI values to assess the performance of each clustering method on each data set. PGMM (SizeDensity) performs the best in the data sets 1 and 2 featured with over-dispersion (high variance) in cluster sizes whereas DPC (DensityOnly) and GMM (DensityOnly) performs the best in the data sets 3 and 4 featured with low variance in cluster sizes, respectively. Note some methods (eg, DPC) do not require the number of clusters as an input whereas others

TABLE 2 The adjusted Rand index (ARI) values of 8 selected clustering methods: PGMM (SizeDensity), GMM (DensityOnly), PMM (SizeOnly), Dirichlet process mixture model (DPMM), density peak clustering (DPC), hierarchical clustering average (HC-A), hierarchical clustering single (HC-S), hierarchical clustering complete (HC-C), and spectral clustering using 4 simulated data sets. (the best performance results are in bold face)

	PGMM	GMM	PMM	DPMM	DPC	HC-A	HC-S	HC-C	Spectral clustering
1	**0.781**	0.724	0.413	0.637	0.721	0.753	0.717	0.732	0.713
2	**0.767**	0.553	0.661	0.695	0.758	0.749	0.682	0.721	0.704
3	0.673	0.772	0.440	0.587	**0.781**	0.718	0.691	0.705	0.711
4	0.600	**0.754**	0.500	0.744	0.629	0.706	0.727	0.673	0.696

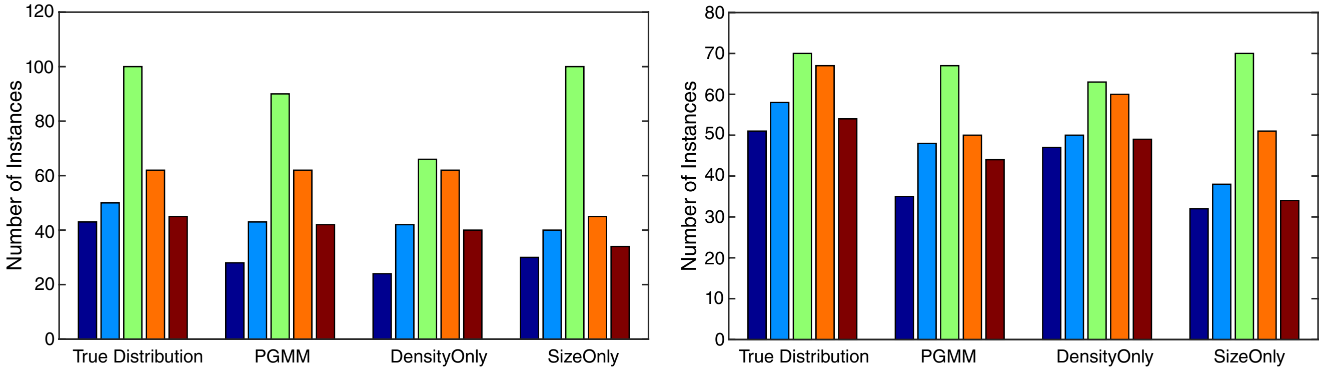


FIGURE 3 A set of case-control synthetic data sets experiments to demonstrate performance of the proposed PGMM method in uncovering the true cluster size distribution. The true cluster size distribution of synthetic data set 2 (upper panel) is with over-dispersion (high variance), while the true cluster size distribution of synthetic data set 4 (lower panel) is with low variance. Histograms in each panel represent, from left to right, the true distribution of cluster size, PGMM (SizeDensity) approach, GMM (DensityOnly) approach and PMM (SizeOnly) approach

do, we gave the true number of clusters to all the compared methods to maximize their performances.

In Figure 3, we gain visual impression on how well PGMM uncovers the true cluster size distribution. Specifically, the upper panel depicts the real-world scenario that cluster sizes are over-dispersed. It is evident that PGMM effectively recovers the true distribution, so does the SizeOnly model. In comparison, the DensityOnly model tends to undermine the cluster size information by normalizing the sizes of all clusters, leading to a more uniform cluster size distribution. In the lower panel, on the other hand, DensityOnly model that neglects cluster size information works as well as PGMM due to the low variance in cluster size distribution. The SizeOnly model performs poorly since there is little information (variance) on the cluster size distribution to capture. Our case-control experiments clearly illustrate the unique capability of PGMM in uncovering data clusters of diverse sizes in addition to data compactness.

5.3 | Real-world data

We used 4 real-world Heart Disease data sets and 1 Walmart Recruiting data set to evaluate the performance of our clustering algorithm. The 4 Heart Disease data sets are collected by Cleveland, Hungary, Switzerland, and the VA Long Beach and we downloaded them from UCI Machine Learning Repository. Each Heart Disease data set contains a single target, which represents the categories of heart disease labeled by the doctors, 6 continuous features (age, blood pressure, serum cholesterol, blood sugar, maximum heart rate, and height at rest) and 8 categorical features (gender, chest pain type, electrocardiographic results, exercise, slope of peak exercise ST segment, major vessels, thal, and angiographic disease results). We treated gender as a dummy variable as described in the Section 4.2 in consistence with what we did for synthetic data.

The Walmart Recruiting data set was downloaded from Kaggle competition website under the name "Walmart Recruiting: Trip Type Classification". It was originally analyzed using market basket analysis to classify shopping trips in order to improve their segmentation process. In this data set, there are 647 054 customers/instances and 6 features including 2 identifying features (VisitNumber and UPSNumber), 1 continuous feature (ScanCount), and 3 categorical features (Weekday, DepartmentDescription, and FinelineNumber), along with the target (TripType). There are 38 TripTypes that we use as the true clusters for evaluating the performance of our clustering algorithm.

In the Walmart Recruiting data set, there are only 3 out of 38 clusters that contain more than 5% of total instances. In the remaining 35 clusters, 33 clusters contain less than 4% of total instances, 28 clusters contain less than 3% of total instances, 23 clusters contain less than 2% of total instances and 19 clusters contain less than 1% of total instances. As a result, we filtered out the very small clusters with very few instances, that is, those with less than 1% of the instances. We used the remaining 19 clusters for further experimental comparison and evaluation.

TABLE 3 The ARI values of the 8 selected clustering methods using 5 real-world data sets. D1-D4 are the 4 heart disease data sets and WR is the Walmart recruiting data set. (N/a entries are due to intractable memory requirement of the corresponding algorithms. the best performance results are in bold face)

	PGMM	GMM	PMM	DPMM	DPC	HC-A	HC-S	HC-C	Spectral clustering
D1	**0.760**	0.514	0.241	0.715	0.717	0.749	0.671	0.679	0.712
D2	**0.751**	0.483	0.540	0.603	0.651	0.685	0.677	0.741	0.673
D3	**0.737**	0.564	0.480	0.629	0.697	0.733	0.716	0.734	0.669
D4	**0.729**	0.664	0.552	0.597	0.704	0.723	0.698	0.712	0.708
WR	**0.396**	0.216	0.143	N/A	N/A	N/A	N/A	N/A	N/A

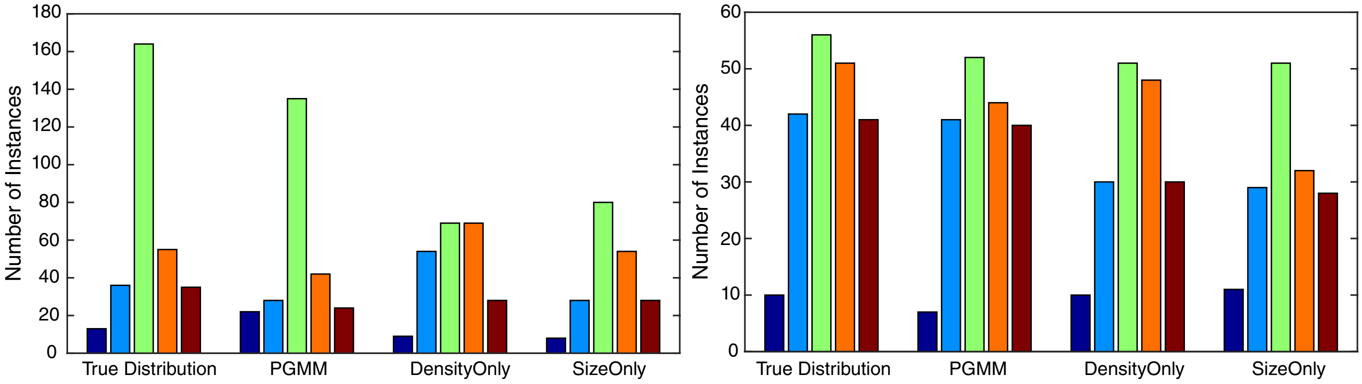


FIGURE 4 Comparison of clustering approaches for mixed continuous and categorical data using heart disease data sets D1 and D3 as first and second panel, respectively. Histogram in each panel represents, from left to right, the true distribution of cluster size, PGMM (SizeDensity) approach, GMM (DensityOnly) approach and PMM (SizeOnly) approach

In Table 3, we report ARI to compare the performance of each of the 8 clustering methods on each of the 5 real-world data sets. We run these methods using a server with the configuration,that is, 4 × 2.6 GHz CPU's and 256 GB of memory. DPMM, DPC, HC-A, HC-S, HC-C, and Spectral Clustering do not scale to this big data set with 647 054 instances due to intractable memory requirement of loading a 647 054 × 647 054 dissimilarity/similarity matrix.

We, once again, observed the superior performance of PGMM to the selected clustering methods implemented in R packages. Similarly, to show the advantage of PGMM in detecting the clusters with highly over-dispersed size distribution, Figure 4 compares cluster size distributions uncovered by the selected methods using the 4 Heart Disease data sets with mixed continuous and categorical features. The upper panel represents a more common structure of real-world data which cluster size distribution is highly over-dispersed where the true cluster sizes are very diverse, PGMM successfully uncovers this important true structure from the real-world data but not the others. The lower panel, on the contrary, represents a less common structure of real-world data where the cluster sizes are more uniform. PGMM again performs the best but others also yield decent results. Our real-world data analysis further demonstrates the superior performance of PGMM to the group of widely accepted clustering methods in clustering real-world data.

6 | CONCLUSION

In this paper, we presented PGMM, a novel probability model to capture both compactness and size distribution of the data clusters for over-dispersed cluster size distribution. We also derived, validated, and evaluated a new EM-type algorithm to estimate the parameters of the model. The over-dispersed cluster size distribution is often the case for real-world data as we have demonstrated using the Walmart's and the Heart Disease data. These real-world data sets are adequate in emulating the complex structure of the real-world data with continuous and categorical features observed in a large number of instances.

Our SizeDensity framework is sufficiently flexible that it accommodates mixed categorical and continuous feature types and over-dispersed cluster size distribution. In addition, our model also employs a dummy variable to represent presence or absence of a specific instance. Along this line, our model can naturally handle missing values by employing additional dummy variables to represent the absence of data points. The cluster size distribution in many real-world data can be very complex, and in certain cases Poisson distribution may not be the best for modeling the cluster size distribution. To this end, our model provides a general framework that can incorporate any distributions for count data, such as negative binomial, Hermite, and exponential distributions.

Our SizeDensity framework is also sufficiently general that it can be extended to SizeConnectivity framework, including

Nonnegative Matrix Factorization (NMF)-based clustering, kernel K-means, and spectral clustering, to detect imbalanced and overlapping clusters by incorporating the cluster size distribution, for example, for NMF-based clustering methods, we will propose a new clustering objective function that minimizes the squared error of matrix factorization while encouraging a diversity of cluster sizes. The latter is achieved by minimizing entropy of the cluster size distribution to encourage the diversity of cluster sizes at the same time of minimizing the errors. Further, our SizeDensity framework can be generalized to NumberSizeDensity or NumberSizeConnectivity frameworks, in which the variation in the number of clusters can be integrated with the variation in the cluster sizes using Bayesian infinite mixture models.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under grants CNS-1637312 and CCF-1451316.

ORCID

Lu Wang http://orcid.org/0000-0003-4016-4096
Dongxiao Zhu http://orcid.org/0000-0002-0225-7817

REFERENCES

1. L. Acharya et al., *Gsgs: A computational approach to reconstruct signaling pathway structures from gene sets*, IEEE Trans. Comput. Biol. Bioinform. 9(2) (2012), 438–450.
2. L. R. Acharya and D. Zhu, *Estimating an optimal correlation structure from replicated molecular profiling data using finite mixture models*, International Conference on Machine Learning and Applications, 2009. ICMLA'09, IEEE, 2009, pp. 119–124.
3. R. Agrawal et al., *Automatic subspace clustering of high dimensional data for data mining applications*, ACM 27(2) (1998).
4. D. F. Andrews and C. L. Mallows, *Scale mixtures of normal distributions*, J. R. Stat. Soc. Ser. B. Methodol. (1974), 99–102.
5. T. Benaglia et al., *Mixtools: An r package for analyzing finite mixture models*, J. Stat. Softw. 32(6) (2009), 1–29.
6. A. R. Benson, D. F. Gleich, and J. Leskovec, *Tensor spectral clustering for partitioning higher-order network structures*, Proceedings of the 2015 SIAM International Conference on Data Mining, 2015, pp. 118–126.
7. P. J. Bickel and K. A. Doksum, *Mathematical statistics: Basic ideas and selected topics*, Vol I, 117, CRC Press, 2015.
8. J. Bruin. Newtest: Command to compute new test @ONLINE (Online), 2011, Feb., available at http://www.ats.ucla.edu/stat/stata/ado/analysis/
9. G. Casella and R. L. Berger, *Statistical inference*, Vol 2, Duxbury, Pacific Grove, CA, 2002.
10. K.-Y. Chiang et al., *Prediction and clustering in signed networks: A local to global perspective*, J. Mach. Learn. Res. 15(1) (2014), 1177–1213.
11. W. H. DuMouchel, *On the asymptotic normality of the maximum-likelihood estimate when sampling from a stable distribution*, Ann. Stat. 1 (1973), 948–957.
12. M. A. Figueiredo and A. K. Jain, *Unsupervised learning of finite mixture models*, IEEE Trans. Pattern Anal. Mach. Intell. 24(3) (2002), 381–396.
13. C. Fraley and A. E. Raftery, *Model-based clustering, discriminant analysis, and density estimation*, J. Am. Stat. Assoc. 97(458) (2002), 611–631.
14. B. J. Frey and D. Dueck, *Clustering by passing messages between data points*, Science 315(5814) (2007), 972–976.
15. S. Garavaglia and A. Sharma, *A smart guide to dummy variables: Four applications and a macro*, Proceedings of the Northeast SAS Users Group Conference, 1998.
16. J. C. Gower, *A general coefficient of similarity and some of its properties*, Biometrics 27 (1971), 857–871.
17. H. He, E. Garcia, et al., *Learning from imbalanced data*, IEEE Trans. Knowl. Data Eng. 21(9) (2009), 1263–1284.
18. X. Hong, S. Chen, and C. J. Harris, *A kernel-based two-class classifier for imbalanced data sets*, IEEE Trans. Neural Netw. 18(1) (2007), 28–41.
19. F. Höppner and F. Klawonn, *Clustering with size constraints*, in *Computational Intelligence Paradigms*, Springer, Berlin and Heidelberg, Germany, 2008, 167–180.
20. S. P. Hubbell, *The unified neutral theory of biodiversity and biogeography (MPB-32)*, Vol 32, Princeton Univ. Press, Princeton, NJ, 2001.
21. L. Hubert and P. Arabie, *Comparing partitions*, J. Classif. 2(1) (1985), 193–218.
22. A. K. Jain, *Data clustering: 50 years beyond k-means*, Pattern Recogn. Lett. 31(8) (2010), 651–666.
23. A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice-Hall, Inc., 1988.
24. K. Kailing, H.-P. Kriegel, and P. Kröger, *Density-connected subspace clustering for high-dimensional data*, Proceedings of SDM, **4**, SIAM, 2004.
25. T. Kohonen et al., *Self organization of a massive document collection*, IEEE Trans. Neural Netw. 11(3) (2000), 574–585.
26. M. L. Kwan et al., *Epidemiology of breast cancer subtypes in two prospective cohort studies of breast cancer survivors*, Breast Cancer Res. 11(3) (2009), R31.
27. T. Li, C. Ding, and M. I. Jordan, *Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization*, Seventh IEEE International Conference on Data Mining, 2007. ICDM 2007*,* IEEE, 2007, pp. 577–582.
28. L. Liu et al., *Fatigue and sleep quality are associated with changes in inflammatory markers in breast cancer patients undergoing chemotherapy*, Brain Behav. Immun. 26(5) (2012), 706–713.
29. S. Liverani, D. I. Hastie, L. Azizi, M. Papathomas, and S. Richardson, *Premium: An r package for profile regression mixture models using dirichlet processes*, arXiv preprint arXiv:1303.2836, 2013.
30. M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, cluster: Cluster Analysis Basics and Extensions, 2016, r package version 2.0.5 — For new features, see the 'Changelog' file (in the package source).
31. G. McLachlan and D. Peel, *Finite mixture models*, John Wiley & Sons, 2004.
32. R. M. Neal, *Markov chain sampling methods for dirichlet process mixture models*, J. Comput. Graph. Stat. 9(2) (2000), 249–265.
33. A. Y. Ng et al., *On spectral clustering: Analysis and an algorithm*, Adv. Neural Inf. Proces. Syst. 2 (2002), 849–856.
34. W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip, *Efficient clustering of uncertain data*, Sixth International Conference on Data Mining, 2006. ICDM'06, IEEE, 2006, pp. 436–445.
35. E. Paradis, J. Claude, and K. Strimmer, *APE: Analyses of phylogenetics and evolution in R language*, Bioinformatics 20 (2004), 289–290.
36. J. Qian and V. Saligrama, *Spectral clustering with unbalanced data*, arXiv preprint arXiv:1302.5134, 2013.
37. A. Rodriguez and A. Laio, *Clustering by fast search and find of density peaks*, Science 344(6191) (2014), 1492–1496.
38. H. Shan and A. Banerjee, *Bayesian co-clustering*, Eighth IEEE International Conference on Data Mining, 2008. ICDM'08. IEEE, 2008, pp. 530–539.
39. J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell. 22(8) (2000), 888–905.
40. W. N. Venables and B. D. Ripley, *Modern applied statistics with S*, 4th ed., Springer, New York, 2002. ISBN: 0&hyphen;387&hyphen;95457&hyphen; 0 [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4.
41. B. Wang et al., *Similarity network fusion for aggregating data types on a genomic scale*, Nat. Methods 11(3) (2014), 333–337.
42. L. Wang et al., *Poisson-markov mixture model and parallel algorithm for binning massive and heterogeneous dna sequencing reads*, in *International Symposium on Bioinformatics Research and Applications*, Springer, Cham, Switzerland, 2016, 15–26.
43. L. Wang, D. Zhu, Y. Li, and M. Dong, *Modeling over-dispersion for network data clustering*, 16th International Conference on Machine Learning and Applications (ICMLA), IEEE, 2017.
44. J. J. Whang, I. S. Dhillon, and D. F. Gleich, *Non-exhaustive, overlapping k-means*, Proceedings of the 2015 SIAM International Conference on Data Mining, 2015, pp. 936–944.
45. C. J. Wu, *On the convergence properties of the em algorithm*, Ann. Stat. 11 (1983), 95–103.

46. R. Xu, D. Wunsch, et al., *Survey of clustering algorithms*, IEEE Trans. Neural Netw. 16(3) (2005), 645–678.
47. S. Zhu, D. Wang, and T. Li, *Data clustering with size constraints*, Knowl.-Based Syst. 23(8) (2010), 883–889.

How to cite this article: Wang L, Zhu D, Dong M. Clustering over-dispersed data with mixed feature types. *Stat Anal Data Min: The ASA Data Sci Journal*, 2018;11:55–65. https://doi.org/10.1002/sam.11369.

APPENDIX A

A1 | DENSITYONLY: GMM

Most of the compactness-based clustering methods consider only the mean and the variance/covariance of each cluster, which were captured by class conditional density. If we use Gaussian and multinomial distributions to model continuous and categorical data density in each cluster, the likelihood function of the DensityOnly model (Gaussian Multinomial Mixture model, GMM) can be shown as:

$$L_{\mathrm{PGMM}}(X,Y,Z|\pi,\Theta)=\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\pi_k\frac{1}{\sqrt{2}\pi\sigma_k}\right.$$
$$\left.e^{-\frac{(X_1-\mu_k)(X_i-\mu_k)^{\mathrm{T}}}{2\sigma_k^2}}\prod_{j=1}^{J}\left[n_k!\left(\prod_{h=1}^{H}\frac{p_{k_{Y_jL_h}}^{n_{kY_jL_h}}}{n_{kY_jL_h}!}\right)\right]\right)^{Z_{ik}}. \tag{A1}$$

APPENDIX B

B1 | SIZEONLY: PMM

Besides the class conditional density, cluster size distribution is also an important structural information to discriminate clusters from clusters. The size of *k*th cluster can be modeled using a discrete distribution Poisson distribution, so that we can discriminate clusters simply by their sizes. The likelihood function of the Size Only model (Poisson Mixture Model, PMM) can be shown as:

$$L_{\mathrm{PGMM}}(X,Y,Z|\pi_k,\lambda_k)=\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\pi_k\frac{\lambda_k^{n_k}e^{-\lambda_k}}{n_k!}\right)^{Z_{ik}}. \tag{B1}$$

APPENDIX C

C1 | ADJUSTED RAND INDEX

Consider a pair of instances to be positive if they are from the same cluster, otherwise it is negative. Here is the formula for calculating ARI of the clustering algorithm. Assume T is the number of true clusters as shown in Table C1, we also use t_1, t_2, … , t_r to represent truth clusters and use k_1, k_2, … , k_K to represent the calculated clusters. $n_{i.}$ represents the number of instances that belongs to cluster t_i, n_j represents the number of instances assigned to cluster k_j, and n_{ij} represents the number of instances belongs to cluster t_i and are assigned to cluster k_j. Hence $n_{i.}=\sum_{j=1}^{K}n_{ij}$, $n_{.j}=\sum_{j=1}^{T}n_{ij}$, and $n=\sum_{i=1}^{T}n_{i.}=\sum_{j=1}^{K}n_{.j}$.

ARI [21] is the corrected-for-chance version of the Rand index. It is used to access the overall performance of the clustering algorithm in the Section 5. It can be calculated as follows:

$$ARI=\frac{\sum_{i,j}\binom{n_{ij}}{2}-\left[\sum_i\binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{n_{i.}}{2}+\sum_j\binom{n_{.j}}{2}\right]-\left[\sum_i\binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}\right]/\binom{n}{2}}. \tag{C1}$$

TABLE C1 A confusion matrix of symbols for defining ARI

Truths\clusters	k_1	k_2	…	k_K	Total
t_1	n_{11}	n_{12}	…	n_{1K}	$n_{1.}$
t_2	n_{21}	n_{22}	…	n_{2K}	$n_{2.}$
.	.	.		.	.
.	.	.		.	.
t_T	n_{T1}	n_{T2}	…	n_{TK}	$n_{T.}$
Total	$n_{.1}$	$n_{.2}$	…	$n_{.K}$	$n_{..}=n$