DOI:10.12113/202202014

基于狄利克雷多项式过程模型与 K-means 结合的菌群分析

彭　显,贺建峰*
(昆明理工大学 信息工程及其自动化学院,昆明 650000)

摘　要:群体分型是一种有助于更好的理解人类身心健康等复杂生物学问题的有效方法,聚类是一种为了对样本分组来降低复杂性的定义肠型的方法,而传统 K-means 聚类算法的 K 值选取无法确定,本文在传统 K-means 聚类算法的基础上进行了改进,并公开数据集上进行了验证,实验表明改进算法能够解决 K 值选取无法确定的问题,且聚类结果的稳定性、准确性和聚类质量都得到显著提高。将改进后的模型运用于肠道菌群 OTUs 数据,发现不仅能够有效地区分 2-型糖尿病患者样本间的相似性,而且能鉴定出影响菌群结构异质性最大的 OTUs 菌,为临床解决 2-型糖尿病问题提供了一种新的思路。

关键词:K-means 算法;狄利克雷过程混合模型;菌群分析;群体分型;聚类
中图分类号:TP181　　**文献标志码**:A　　**文章编号**:1672-5565(2024)01-047-11

Flora analysis based on Dirichlet polynomial process model and K-means

PENG Xian, HE Jianfeng*
(*School of Information Engineering and Automation*, *Kunming University of Technology*, *Kunming* 650000, *China*)

Abstract: Population typing is an effective method to better understand complex biological problems such as human physical and mental health. Clustering is a method to define intestinal type in order to reduce complexity by grouping samples. However, the selection of K value of traditional K-means clustering algorithm cannot be determined. This paper improves the traditional K-means clustering algorithm and verifies it on the public dataset, The experimental results show that the improved algorithm can solve the problem of undetermined K value selection, and the stability, accuracy and quality of clustering results are significantly improved. Applying the improved model to the OTUs data of intestinal flora, it is found that it can not only effectively distinguish the similarities between samples of patients with type 2 diabetes, but also identify the OTUs bacteria that have the greatest impact on the heterogeneity of flora structure, providing a new perspective for clinical solutions to the problem of type 2 diabetes.
Keywords: K-means algorithm; Dirichlet process mixed model; Flora analysis; Population typing; Clustering

人体内定殖着数万亿有益于身心健康的微生物,通常不同的微生物群落定殖在不同的人体部位。研究表明不同人同一部位的微生物群落组成差异很大,同一个人的不同部位的微生物群落组成也显著不同。近年来,越来越多的研究表明人类的疾病与微生物的分布有关。例如,Hill-Burns 等[1]的研究成果中就找到了将帕金森氏症与肠道微生物群联系起来的证据,Oren 等[2]计算了 JS、BC、加权和非加权 UniFrac 距离的根平方和无平方,并选择了具有 PAM 算法中使用的预测强度和轮廓指数的集群数量。与 Koren 的方法不同,Hong 等[3]应用了具有欧几里得距离的 K 均值来识别两个集群,他们认为集群的数量在他们的研究中具有生物学意义。Holmes 等[4]Dirichlet 多项式混合物(DMM),用于微生物元基因组学数据的概率建模,混合成分将群落聚集到不同的"元社区",从而确定环境型或肠型,即组成

收稿日期:2022-02-24;**修回日期**:2022-03-29;**网络首发日期**:2023-03-17.
网络首发地址:https://kns.cnki.net/kcms/detail/23.1513.Q.20230316.1704.002.html
* **通信作者**:贺建峰,男,教授,博导,研究方向:数据挖掘,图像处理. E-mail:120112624@ qqcom.
引用格式:彭显,贺建峰.基于狄利克雷多项式过程模型与 K-means 结合的菌群分析[J].生物信息学,2024,22(1):47-57.
PENG Xian, HE Jianfeng.Flora analysis based on Dirichlet polynomial process model and K-means[J].Chinese Journal of Bioinformatics, 2024,22(1):47-57.

相似的社区群落。Dong M 等[5]用 Dirichlet 多项式模型生成的三个模拟数据集和与帕金森病相关的真实肠道微生物群数据进行了实证研究,以调查三种预测分析策略的性能。Chen J 等[6]提出稀疏的 Dirichlet 回归可以更好地识别微生物组相关的协变量。Yang Y 等[7]提出了一种计算模型,即 k-Lognormal-Dirichlet-Multinomial 模型(kLDM)。Shi Y 等[8]通过采用贝叶斯建模方法,能够从数据中了解集群的数量。A Monleon-Getino 使用基于马尔可夫链蒙特卡罗模拟(MCMC)的贝叶斯方法来计算稀疏曲线期间 OTU 的最大预期数量。Irannia ZB 等[9]开发了一个马尔科夫随机场模型来预测未知微生物的分类群。Mao J 等[10]引入 Dirichlet 树多项式混合物(DTMM)作为微生物组研究中扩增子序列数据的生成模型。Niccolai E 等[11]找到了肠道微生物群的不健康状态通常是短链脂肪酸(SCFA)浓度的降低的原因。Harris K 等[12]使用分层的 Dirichlet 过程,为多站点 UNTB 开发了一种高效的贝叶斯拟合策略,Yu P 等[13]开发了一种计算对数可能性的新方法,以解决不产生长期运行时间的不稳定问题。

针对上述问题,本文提出了一种狄利克雷过程混合模型(DPMM)与 K-means 聚类相结合的方法,实验结果表明,改进后的算法选取的初始聚类中心点唯一,并且保证了聚类结果的稳定性,与传统 K-means 等聚类算法相比,聚类准确性和聚类质量都得到了提高。

1 K-means 算法

K-means 算法是有 MacQueen 于 1967 年总结多个领域学术成果后提出的,并提一次给出 K 均值的算法步骤,并用数学方法进行了证明。K 均值是机器学习中一种无监督的学习算法。K 均值的核心思想是给定一个数据集,它有 n 条数据记录,要把它分成 K 类。那么,最后所分的类,类间相似度最小,类内相似度最大。

从算法的核心思想可以看出,K 均值在迭代过程中,类内的惩罚函数不断变小,最终保证惩罚函数收敛。K 均值最大的缺点在于初始化过程中 K 的选择。鉴于 K 的值相较于数据集不会太大。往往采用枚举的方式进行测算。K 均值的惩罚函数往往选用欧式距离。具体步骤如下:

假设将 n 个 m 维数据 X 分成 K 类。

1)选择 K 个样本作为初始的聚类中心 O;

$$O = \{X_i, X_j, \cdots, X_k\}, i \neq j \neq k < K \tag{1}$$

2)针对每个样本计算其距离聚类中心距离 D,并计算代价函数 $L(X_n, O)$,并将距离最小的样本与聚类中心归为一类 C_m;

$$D = \{\|X_n - O_m\|\}, i \neq j \neq k \neq n < K \cap m = 1, \cdots, K \tag{2}$$

$$C_m = \{\{X_n \mid L(X_n, O) = min(L(X_n, O))\}, O_m\} \tag{3}$$

其中,min(*)为最小值。

3)对每类求算数平均作为新的聚类中心;

$$O_m = \frac{\sum C_m}{len(C_m)} \tag{4}$$

其中,len(*)为集合长度。

4)重复 2、3 步操作,知道代价函数达 L 到终止条件 L_0,并生成最终结果 Y。

对于 n 个 m 维数据分成 K 类的时间复杂度为 $O(tKnm)$,空间复杂度 $O(m(n+K))$,其中 t 为迭代次数。

构建拓扑见图 1。

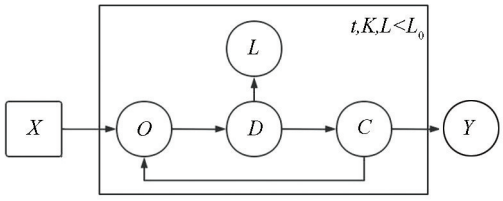

图 1 K 均值数学拓扑图

Fig.1 K-means mathematical topology

从 K-means 算法可以看出,K 值一旦给定将无法更改,算法会根据 K 值的不同,产生不同的惩罚函数 L,L 对 K 值敏感。而其迭代过程的次数是由终止条件决定。整套算法解释性良好,算法灵活度高。

鉴于 K 均值原理简单,实现容易,收敛速度快,聚类效果好,算法解释性强,需要主动调参参数少,不需要历史样本。

2 狄利克雷混合过程(DPMM)

狄利克雷混合过程分为:嵌套 DPMM、关联 DPMM、层次 DPMM(图 2)。

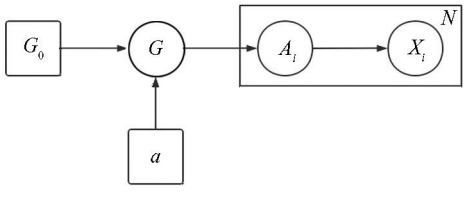

图 2 DPMM 数学拓扑

Fig.2 DPMM mathematical topology

从图中可以看出,根据参数 α 在 G_0 中选出 G 的过程是 DPMM 过程,在 A_i 的条件下不断得试探 G 到 x_i ,选取过程可以是无限多次。

在共享样本中进行采样,如果该过程符合马尔可夫蒙特卡洛采样法就可以构造为分层 DPMM(图 3)。

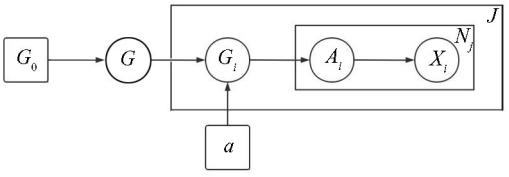

图 3　分层 DPMM 数学拓扑

Fig.3　Hierarchical DPMM mathematical topology

从图 3 中可以看出在 G_0 中选出 G_j 的过程是分层 Dirichlet 过程,首先在 G_0 中选出 G 的但是 G 的不能满足 α 的条件,又在 G 中选出 G_j 满足 α 的条件,在 A_i 的条件下不断得试探 G_j 到 x_i ,重复选出 G_j 满足 α 的条件继续在 A_i 的条件下不断得试探 G_j 到 x_i ,知道满足条件 J。

既可以通过构造采样函数可以进行 DPMM 的无限循环过程,也可以通过无线叠加样本进行 DPMM 的无限循环过程。鉴于实验所用 OTUs 是有限的,采用抽样函数进行 Dirichlet 过程,考虑到实现过程中不可能真正无限运行,以样本遍历最大次数为结束无线循环过程的条件。

狄利克雷混合过程下的 OTUs 数据分类模型前描述,描述结果见图 4,图 5。

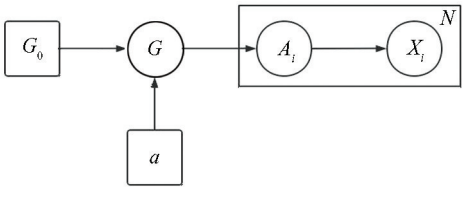

图 4　DPMM OTUs 数据聚类拓扑

Fig.4　DPMM OTUs data clustering topology

从图 4 可以看出,此时 G_0 变成 OTUs 数据集, α 是随机采样规则, A_i 为聚类过程, x_i 为聚类结果,N 为聚类寻优过程。

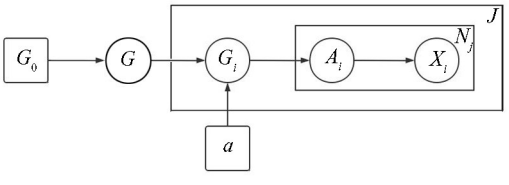

图 5　分层 DPMMOTUs 数据聚类拓扑

Fig.5　Hierarchical dpmmotus data clustering topology

从图 5 可以看出,此时 G_0 变成 OTUs 数据集, G 部分 OTUs 数据集, α 是随机采样规则, A_i 为聚类过程, x_i 为聚类结果,N 为聚类寻优过程,J 是 K 均值的 K 选取过程。

3　K-means 的 Dirichlet 过程

引入 Dirichlet 过程可以充分解决 K 均值选择 K 值难的问题,此处选用分层 DPMM。外层通过对 K 值的变化构建不同的 K-means 算法,内层通过对数据集采样 Dirichlet 过程实现。可以看出分层 DPMM 天生适用于聚类算法的调优。其拓扑结构见图 6。

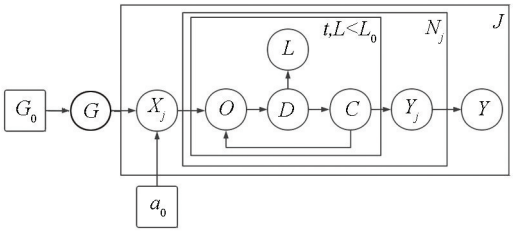

图 6　PDMM 改进的 K 均值数学拓扑图

Fig.6　Improved K-means mathematical topology of PDMM

构建算法:

1)采用抽样算法从数据 X 中抽出样本 X_{nk} , X_{nk} 与 X 列属性相同;

2)给 K-means 设置最大 K 值;

3)选择 K 个样本作为初始的聚类中心 O_n ;

$$O_n = \{X_{ni}, X_{nj}, \cdots, X_{nk}\}, i \neq j \neq k < K \tag{5}$$

4)针对样本 X_{nk} 计算其距离聚类中心距离 D_n ,并计算代价函数 $L(X_{nk}, O_n)$,并将距离最小的样本与聚类中心归为一类 C_{nm} ;

$$D_n = \{\|X_{nk} - O_n\|\}, i \neq j \neq k < K \tag{6}$$

$$C_{nm} = \{\{X_n \mid L(X_{nk}, O_n) = min(L(X_{nk}, O_n))\}, O_n\} \tag{7}$$

5)对每类求算数平均作为新的聚类中心;

$$O_{nm} = \frac{\sum C_{nm}}{len(C_{nm})} \tag{8}$$

6)重复 2、3 步操作,知道代价函数达 L 到终止条件 L_0 。

改进算法保留了 K 均值实现容易,收敛速度快,距离效果好,算法解释性强的同时,还解决了 K 值选择难的问题。

对于 n 个 m 维数据分成 K 类的时间复杂度为 $O(tKnmNJ)$,空间复杂度 $O(m(n+K)NJ)$ 。

4　实验结果与分析

K-means 算法是一种基于距离的聚类方法,它

以 k 为参数,把包含 n 个对象的数据集划分为 k 个簇,使得簇内具有较高的相似度,而簇与簇之间的相似度较低。其基本思想是:根据预先设定的的聚类中心数随机选取 K 个样本点作为初始聚类中心,把剩余的对象分配给离它最近的簇中;然后将各簇所有对象的平均值作为新的聚类中心,并把每个对象重新分配到最近的簇中,不断迭代该过程,直到聚类质量不再发生变化,目标函数最小化为止。实验结果见表 1,表 2。

4.1 改进的 K-means 算法

1)算法首先计算每个样本点的局部密度和相对距离,将每个样本点的局部密度和相对距离作积,并将其降序排列,然后根据输入的聚类数自动选取前 K 个决策值较大的样本点作为初始聚类中心点。

2)将选取的 K 个聚类中心应用到 K-means 算法中,并利用各簇的中位数代替原来的均值进行后续聚类中心的更迭。

从表 1 中的两个有效性指标来看,改进后的K-means算法的聚类准确性明显优于K-means这是因为K-means算法随机选取初始聚类中心严重影响聚类的准确性和稳定性。在 R15 数据集上,改进的K-means算法的准确率和标准化互信息高达 99%以上,聚类效果与真实情况将近吻合。

表 1 实验结果对比

Table 1 Comparison of experimental results

数据集	评价指标	K-means	改进的 K-means
Aggregation	ACC	0.769 0	**0.860 9**
	NMI	0.838 3	**0.855 7**
R15	ACC	0.811 7	**0.996 7**
	NMI	0.923 3	**0.994 2**
R5	ACC	0.641 5	0.995 5
	NMI	0.794 0	0.984 8
Iris	ACC	0.823 7	**0.900 0**
	NMI	0.758 2	**0.766 1**
Wine	ACC	0.578 7	**0.719 1**
	NMI	0.414 0	**0.434 0**
Seeds	ACC	0.891 2	**0.895 2**
	NMI	0.685 8	**0.710 1**
Glass	ACC	0.490 7	**0.514 0**
	NMI	0.286 2	0.359 3

表 2 在 Iris 数据集上的有效性结果对比

Table 2 Comparison of effectiveness results on Iris dataset

序号	K-means			改进后的 K-means		
	初始中心	真实类别	迭代次数	初始中心	真实类别	迭代次数
1	(144,73,119)	(3,2,3)	9	(8,100,113)	(1,2,3)	5
2	(22,63,136)	(1,2,3)	3	(8,100,113)	(1,2,3)	5
3	(119,143,98)	(3,3,3)	8	(8,100,113)	(1,2,3)	5
4	(59,98,26)	(2,2,1)	6	(8,100,113)	(1,2,3)	5
5	(106,5,41)	(3,1,1)	13	(8,100,113)	(1,2,3)	5
6	(7,15,122)	(1,1,3)	11	(8,100,113)	(1,2,3)	5
7	(105,48,141)	(3,1,3)	4	(8,100,113)	(1,2,3)	5
8	(6,66,57)	(1,2,2)	9	(8,100,113)	(1,2,3)	5
9	(74,67,96)	(2,2,2)	5	(8,100,113)	(1,2,3)	5
10	(102,98,25)	(3,2,1)	7	(8,100,113)	(1,2,3)	5
平均	—	—	8	—	—	5

表 2 给出了改进后的 K-means 算法与 K-means 在 Iris 数据集上的详细情况。两种种算法共运行 10 次,并给出了每次运行时所选取的初始中心点(用数据集中对应的编号表示)、初始聚类中心所对应的实际类别、迭代次数、运行时间。

从表 2 中可以看出,传统的 K-means 算法每次选取的初始聚类中心都是随机的,导致每次聚类结果不一致;并且在很多情况下,选取的初始聚类中心可能位于同一个簇中,这样使得最初的聚类中心过于邻近,导致算法迭代次数增加。改进后的 K-means 算法每次运行所选取的初始聚类中心是唯一的,且每个聚类中心点与真实类别相对应,算法具有很好的稳定性。

4.2 实验数据集与评价标准

实验数据集来源于云南第一人民医院内分泌科在 2015 年到 2017 年对 2 型糖尿病患者统计。其中 N 为学校正常对照 20 例,PN 为医院正常对照 19 例,P 为 2 型糖尿病合并胃肠自主神经病变 30 例,D 为 2 型糖尿病 76 例。

本文引入了许多信息准则。通过加大模型复杂

度的惩罚函数来避免拟合。假设聚类结果有参数 k 个,观察数 n 个,测试值为 y,中心点值为 $\tilde{y}$ 。

1) AIC(赤池信息准则)

AIC 由日本赤池弘次在 1974 年提出的,是衡量模型拟合优良性的一个标准。它的定义是建立在熵的概念上,为模型复杂度和拟合数据优良性的评价提供了标准。具体函数如下:

$$AIC = 2k + nln\left(\frac{\sum (y - \tilde{y})^2}{n}\right) \tag{9}$$

2) BIC(贝叶斯信息准则)

BIC 惩罚考虑了样本数量,对样本数量大,防止模型精度过高而造成模型复杂度过高。BIC 于 1978 年由 Schwarz 提出,具体公式如下:

$$BIC = kln(n) + nln\left(\frac{\sum (y - \tilde{y})^2}{n}\right) \tag{10}$$

相较于与 RMSE(均方根误差)只关注残差,AIC、BIC 综合衡量了模型的复杂度与拟合程度。调整 R^2 ($Adj - R^2$)虽然突破了 R^2 解释样本与拟合模型相关系数的局限,但是当参数不断变多时:

$$Adj - R^2 = 1 - \frac{RSS}{TSS} \tag{11}$$

其中 $\frac{RSS}{TSS}$ 下降,但 $\frac{N-1}{N-k-1}$ 提升,对的参数变化带来的波动不明显。

本文选择 AIC 与 BIC 作为实验评估,并作为实验过程中的惩罚函数。

为了反映出模型过拟合的特性,给定 5 个二维高斯分布,每个分布随机采样 500 个点,$\begin{bmatrix} Z_{500\times2} \cdot \begin{bmatrix} 1.3 & -0.3 \\ -0.7 & 0.4 \end{bmatrix} \\ Z_{500\times2} + [-6 \quad 3] \end{bmatrix}$,分别采用传统聚类方法与 DPMM 改进的聚类算法进行聚类,当最大类别设置成 10,进行过拟合。拟合分布效果见图 7。

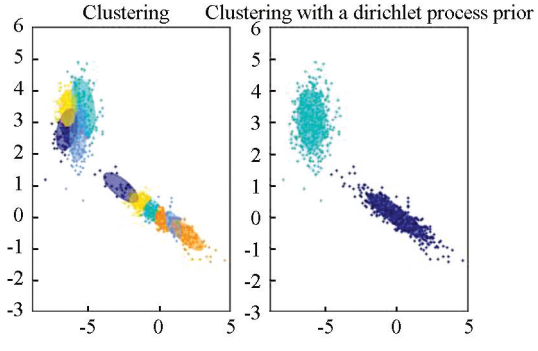

图 7　过拟合实验

Fig.7　Over fitting experiment

从图 7 可以看出传统聚类方法为最求高精度将原始数据分为 10 个类,但是通过 DPMM 改进的聚类算法将数据分为 2 类。主观上看:散点主要集中在两个区,没有必要拟合成 10 类。此类过拟合是典型的参数太多,模型复杂度过高造成的。进一步评估 DPMM 改进算法的 AIC 与 BIC 见图 8。

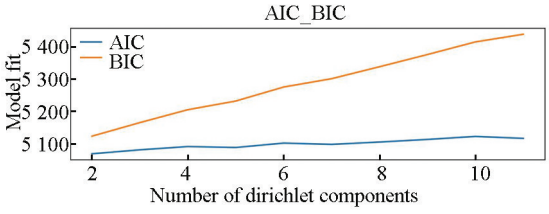

图 8　模型评价

Fig.8　Model evaluation

分类越低 AIC 与 BIC 越好,因此 DPMM 改进的聚类算法,会随着参数的增多拟合度更。可以看出,BIC 的抬升幅度远大于 AIC。两种评估准则第二项相同,造成差异来源于第一项。每一个 Dirichlet 组成值都是对 K 均值进行一次聚合。传统的 K 均值方法无法实现 K 值由 2 ~ 10 的光滑过渡,但是 DPMM 改进的 K-means 算法可以很好的实现。K 值由一到极大的某值计算。

4.3　II 型糖尿病合并胃肠自主神经病变与 II 型糖尿病菌群差异

在对 OTUs 数据进行聚类分析之前,还需要分析一下样本的种类。并且对每类样本进行主成分分析。本实验所用的 OTUs 数据包含生物种类包括囊括海量细菌与古生菌。16s RNA 获得了近 2 000 种 OTU 序列,包含 2 界,OTUs 识别率 100%;28 门,OTUs 识别率 99.39%;55 纲,OTUs 识别率 97.83%;108 目,OTUs 识别率 95.38%;185 科,OTUs 识别率 85.41%;408 科,OTUs 识别率 62.47%;284 种,OTUs 识别率 16.76%,见表 3。

分 P、PN、N、D4 类分析 OTU 分布。以 85%识别率为主要分析目标,鉴于界的种类只有 2 种,没有聚类的必要,后续分析将集中在门纲目的级别进行分析与实验。实验以 0.001 为识别精度。

II 型糖尿病合并胃肠自主神经病变记为 P 类,与 OTUs 关联有效数据有 268 个,包含生物分类见表 4,表 5。

表 3 生物种类

Table 3 Biological species

界	门	纲	目	科	属	种
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Escherichia_coli
Archaea	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia-Shigella	Bacteroides_plebeius
	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Romboutsia	Bacteroides_coprocola
	Fusobacteria	Negativicutes	Selenomonadales	Prevotellaceae	Prevotella_9	Bacteroides_stercoris
	Actinobacteria	Betaproteobacteria	Burkholderiales	Ruminococcaceae	Faecalibacterium	Bacteroides_uniformis
	WD272	unidentified_ Acidobacteria	Actinomycetales	Lactobacillaceae	Dialister	Sutterella_ wadsworthensis
	Chlamydiae	KD4-96	Neisseriales	boneC3G7	Prevotella_2	Clostridium_ bifermentans
	Planctomycetes	Methanobacteria	Rickettsiales	Rhodospirillaceae	Collinsella	Oscillibacter_sp._KLE_1745
	TM6	MB-A2-108	Victivallales	Succinivibrionaceae	Alloprevotella	Bacteroides_eggerthii
	Deinococcus-Thermus	Nitrospira	Micrococcales	Carnobacteriaceae	Anaerostipes	Clostridiales_bacterium_ canine_oral_taxon_085
		Holophagae	unidentified_ Saccharibacteria	Family_XIII	Dorea	Lactobacillus_ruminis
		…	…	…	…	…

表 4 II 型糖尿病合并胃肠自主神经病变生物种类

Table 4 Biological species of type II diabetes mellitus with gastrointestinal autonomic neuropathy

门	纲	目
Bacteroidetes	Bacteroidia	Bacteroidales
Proteobacteria	Gammaproteobacteria	Enterobacteriales
Firmicutes	Negativicutes	Selenomonadales
Actinobacteria	Clostridia	Clostridiales
Fusobacteria	unidentified_Actinobacteria	Bifidobacteriales
Verrucomicrobia	Fusobacteriia	Fusobacteriales
Tenericutes	Betaproteobacteria	Burkholderiales
Cyanobacteria	Bacilli	Lactobacillales
Synergistetes	unidentified_Firmicutes	Pasteurellales
	Verrucomicrobiae	Verrucomicrobiales
	Erysipelotrichia	Erysipelotrichales
	Coriobacteriia	Coriobacteriales
	Mollicutes	Mollicutes_RF9
	Deltaproteobacteria	Desulfovibrionales
	Melainabacteria	Gastranaerophilales
	Alphaproteobacteria	Rhodospirillales
	Synergistia	Synergistales

表 5 II 型糖尿病生物种类

Table 5 Biological types of type II diabetes

门	纲	目
Bacteroidetes	Bacteroidia	Bacteroidales
Proteobacteria	Gammaproteobacteria	Enterobacteriales
Firmicutes	Negativicutes	Selenomonadales
Actinobacteria	Clostridia	Clostridiales
Fusobacteria	unidentified_Actinobacteria	Bifidobacteriales
Verrucomicrobia	Fusobacteriia	Fusobacteriales
Tenericutes	Betaproteobacteria	Burkholderiales
Cyanobacteria	Bacilli	Lactobacillales
Proteobacteria	unidentified_Firmicutes	Pasteurellales
Bacteroidetes	Verrucomicrobiae	Verrucomicrobiales
	Erysipelotrichia	Erysipelotrichales
	Coriobacteriia	Coriobacteriales
	Mollicutes	Mollicutes_RF9
	Deltaproteobacteria	Desulfovibrionales
	Melainabacteria	Gastranaerophilales
	Chloroplast	Aeromonadales
	Alphaproteobacteria	unidentified_Chloroplast
	Gammaproteobacteria	Rhodospirillales
	Bacteroidia	Aeromonadales
		Bacteroidales

II 型糖尿病合并胃肠自主神经病变包含 9 门,占 OTUs 总门类的 32.14%;包括 17 纲,占 OTUs 总门类的 30.91%;包括 17 目,占 OTUs 总目类的15.74%。进一步分析门类发现厚壁菌门(Firmicutes)、拟杆菌门(Bacteroidetes)与变形菌门(Proteobacteria)表现明显。鉴于 Firmicutes、Bacteroidetes 比例与肥胖有密切关系,作为人体中含量较大的两种分布广泛,Firmicutes 是指细胞壁的肽聚糖含量高,一般在 50%到 80%,全部为化能营养型生物,无光能营养型,琦细胞壁厚度在 10~50 nm。Bacteroidetes 广泛具有外膜、肽聚糖层和细胞质膜,其中多糖是 Bacteroidetes 的主要能量来源。Proteobacteria 外膜由脂多糖组成,参与人体多种代谢,具有极多形状。

Firmicutes、Bacteroidetes、Proteobacteria 是影响 II 型糖尿病的主要门类,见图 9。

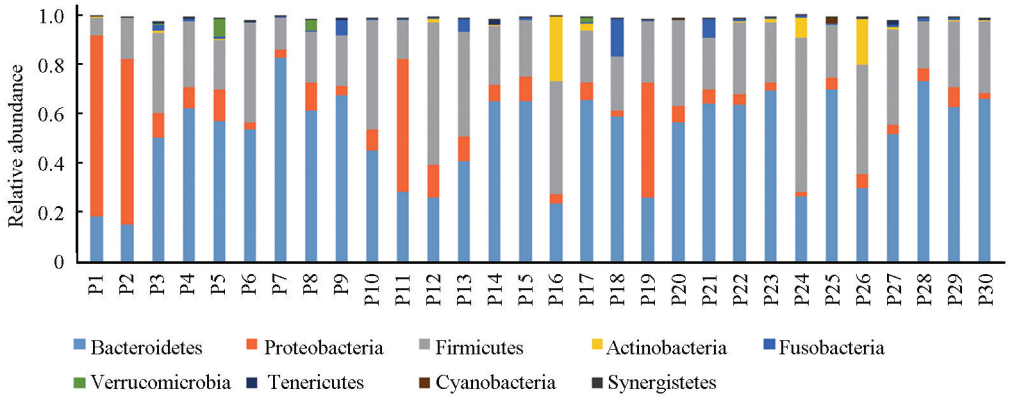

图 9　II 型糖尿病合并胃肠自主神经病变纲类中主要细菌的相对丰度分布

Fig.9　Relative abundance distribution of main bacteria in the class of type II diabetes mellitus with gastrointestinal autonomic neuropathy

如图 9 分析纲类发现,Bacteroidia,Clostridia,Negativicutes 与 unidentified_Actinobacteria 表现明显。拟杆菌纲(Bacteroidia)作为人体肠道中优势厌氧菌,对维持微生态体系具有重要作用。梭菌纲(Clostridia)芽孢直径往往会大于菌体,菌体整体成梭形,其中少数为致病菌。厚壁菌纲(Negativicutes)与放线菌门未确认纲(Unidentified_Actinobacteria)均表现为个体案例占比高,见图 10-图 11。

如图 10 分析目类发现,Bacteroidales、Clostridiales、Selenomonadales 与 Bifidobacteriales 表现明显,其分别为拟杆菌目、梭菌目、硒单目和双歧杆菌目。

2 型糖尿病包含 10 门,占 OTUs 总门类的 35.71%;19 纲,占 OTUs 总门类的 34.55%;20 目,目占 OTUs 总目类的 18.52%。详见图 12-图 13。

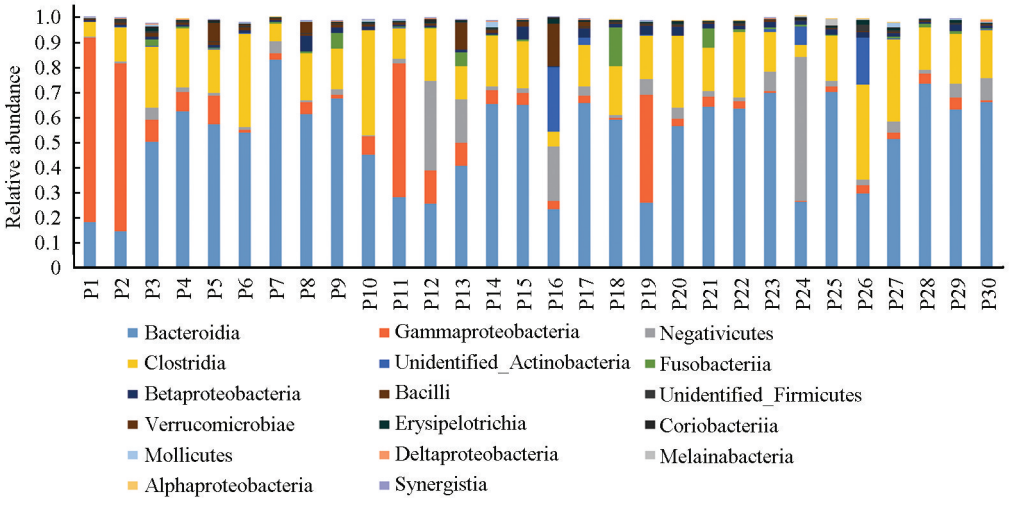

图 10　II 型糖尿病合并胃肠自主神经病变目类中主要细菌的相对丰度分布

Fig.10　Relative abundance distribution of main bacteria in the order of type II diabetes mellitus with gastrointestinal autonomic neuropathy

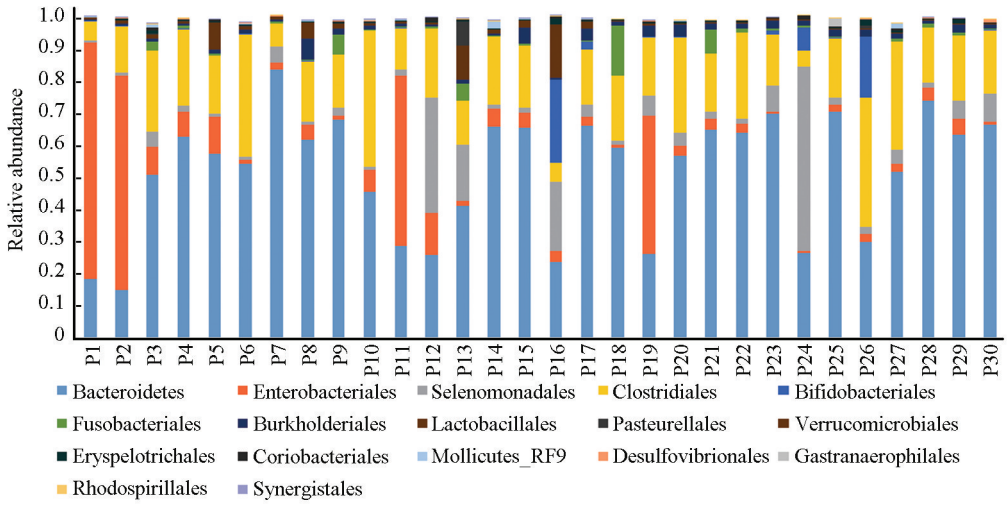

图 11 II 型糖尿病合并胃肠自主神经病变科类中主要细菌的相对丰度分布

Fig.11 Relative abundance distribution of main bacteria in the family of type II diabetes mellitus with gastrointestinal autonomic neuropathy

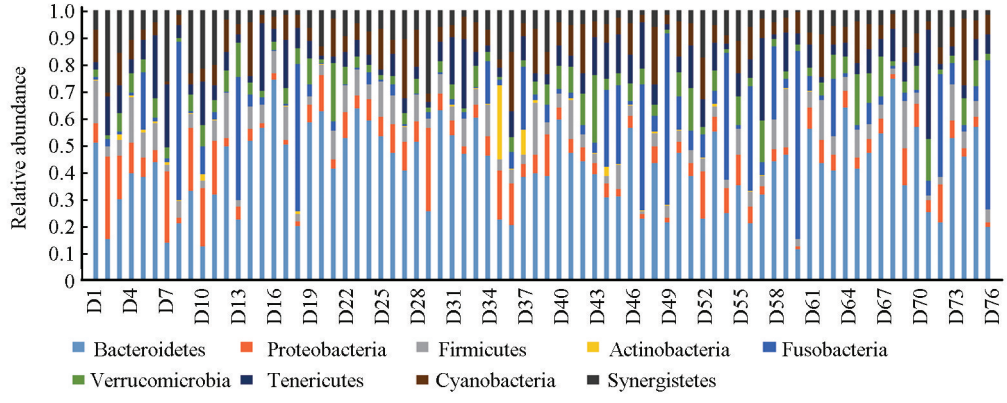

图 12 II 型糖尿病门类中主要细菌的相对丰度分布

Fig.12 Relative abundance distribution of main bacteria in type II diabetes

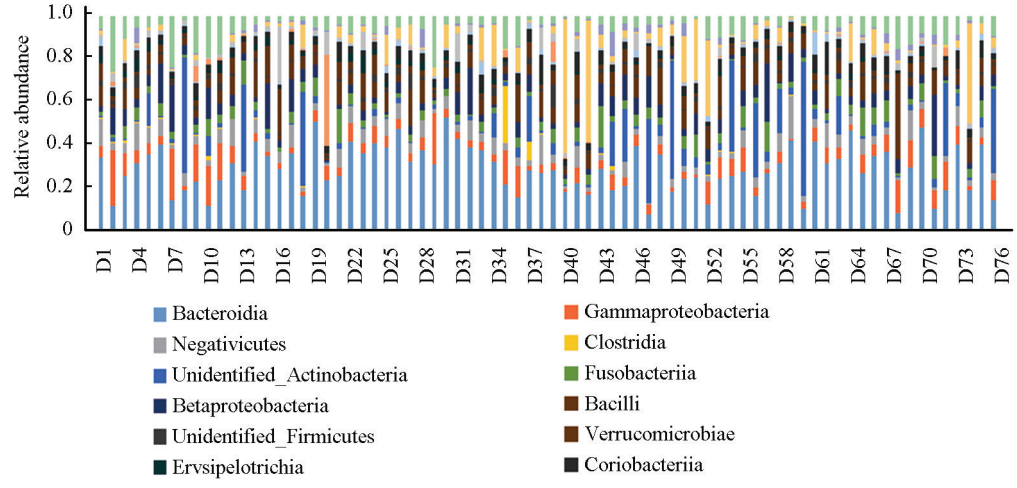

图 13 II 型糖尿病纲类中主要细菌的相对丰度分布

Fig.13 Relative abundance distribution of main bacteria in type II diabetes

Bacteroidetes、Firmicutes、Proteobacteria、Cyanobacteria、Fusobacteria 表现突出与 II 型糖尿病合并胃肠自主神经病变不完全相同。没有减少新的门类，影响 II 型糖尿病的细菌门类包含了影响胃肠自主神经病变的门类。单是缺少了 Cyanobacteria、Fusobacteria。蓝菌门（Cyanobacteria）也值得注意，蓝菌就是旧分

类法中的蓝藻门(Cyanophyta)因为细胞生物学研究发现它们属于原核生物,理论上是一种细菌。梭杆菌门(Fusobacteria)是一个小类群的革兰氏阴性细菌。其中梭杆菌属常见于消化道,是口腔菌群之一,也可导致一些疾病。

Bacteroidia、Chloroplast、Gammaproteobacteria 表面明显。Bacteroidia、Chloroplast 与 II 型糖尿病合并胃肠自主神经病变相同,单是 Gammaproteobacteriaγ 变形菌纲是所知的细菌中种类最多的一纲典型代表如沙门氏菌属(肠炎和伤寒)、耶尔辛氏菌属(鼠疫)、弧菌属(霍乱)、绿脓杆菌。

Bacteroidales、Aeromonadales 表现明显。Bacteroidales 与 II 型糖尿病合并胃肠自主神经病变相同,Aeromonadales 为气单胞菌目可引起人类腹泻。

4.4　DPMM 混合下聚类分析

采用 DPMM 改进聚合方式对给定 K 值进行遍历。K 值的表达观测数值开始增大,DPMM 迭代次数越多,抽样迭代观测量越大。具体流程如下:

1)导入清晰后 OTUs 数据,将 OTU 对所有样本案例相关性均为 0 数据删除;

2)判断导入的 OTUs 数据集一行数据的和是否为 0,为 0 终止程序并报错,不为 0 继续执行;

3)通过 OTUs 对模型进行初始化,并生成符合正态分布的函数;

4)对 OTUs 进行抽样;

5)给定权重,计算抽样拟合结果;

6)将拟合结果由好到差不断填充到列表中,并按权重加权拟合结果,此权重可以满足无限权重的和为 1;

7)重复 4、5、6,直到满足终止条件。

本实验以 K-means 算法最大聚类 10 类为终止条件,以 AIC/BIC 进行评估,反馈聚类的种类个数见图 14。

根据图 14,首先排除分 1 类的无意义结果,发现当 $K=3$ 时,AIC/BIC 值最小,所以 AIC/BIC 最佳,并且 AIC 与 BIC 的走势整体一致说明,观察观测参数变化良好。分析每一个 OTU 的贡献情况贡献最突出的 OTU 为 OTU963、OTU363、OTU1390、OTU19、OTU359 分别对应物种种类为:

OTU963k _ _ Bacteria; p _ _ Firmicutes; c _ _ Clostridia;o__Clostridiales

OTU363k _ _ Bacteria; p _ _ Firmicutes; c _ _ Clostridia;o__Clostridiales;f__Ruminococcaceae

OTU139k _ _ Bacteria; p _ _ Proteobacteria; c _ _ Betaproteobacteria; o _ _ Burkholderiales; f _ _ Comamonadaceae;g__Tepidimonas;s_

OTU19k _ _ Bacteria; p _ _ Bacteroidetes; c _ _ Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__Bacteroides_caccae

OTU359k _ _ Bacteria; p _ _ Firmicutes; c _ _ Clostridia;o__Clostridiales;f__Clostridiales_vadinBB60_group;g__;s__

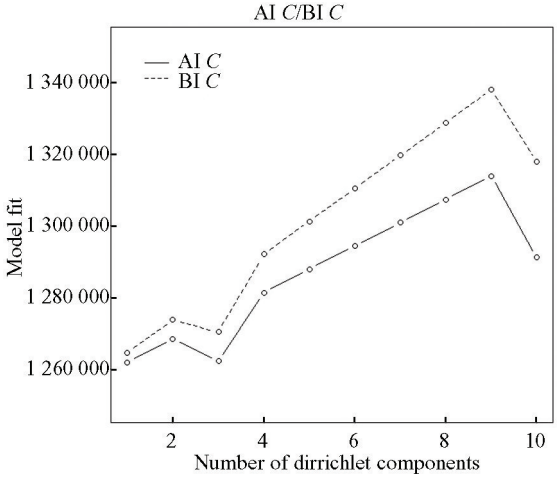

图 14　DPMM 混合聚类模型拟合效果

Fig.14　Fitting effect of DPMM mixed clustering model

进一步对比 DPMM 优势,采用热力图分析发,将样本数据分为 4 类,OTUs 用 K-means 分别聚成 2 类、4 类、1 796 类,用 DPMM 改进的 K-means 进行聚类,聚类效果见图 15-图 18。

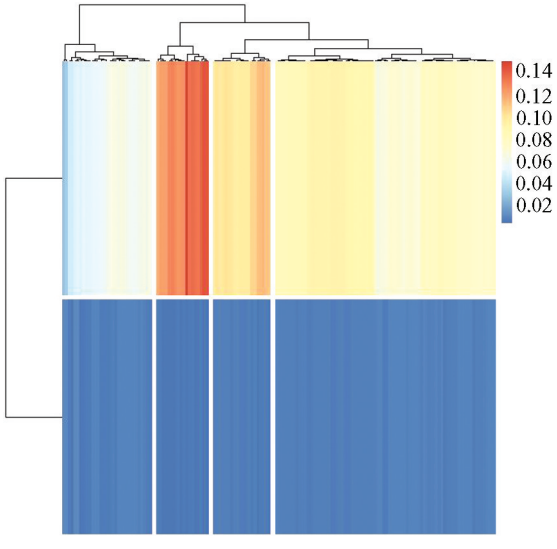

图 15　K-means 2 聚类热图

Fig.15　K-means 2 clustering heat map

图 15 是将 K-means 的 K 值设定为 2;图 16 是将 K-means 的 K 值设定为 4;图 17 是将 K-means 的 K 值设定为1 796,从热力图中可以得出每个 OTU 的

分布情况;图 18 是将 K-means 是运用 DPMM+K-means 进行结合得出来的,DPMM 通过不断选取 K 值,最后选出了 $K=3$ 时为最佳,热力图可以看出聚成三类后分层效果明显,得出结果与传统生物信息学分析一致。

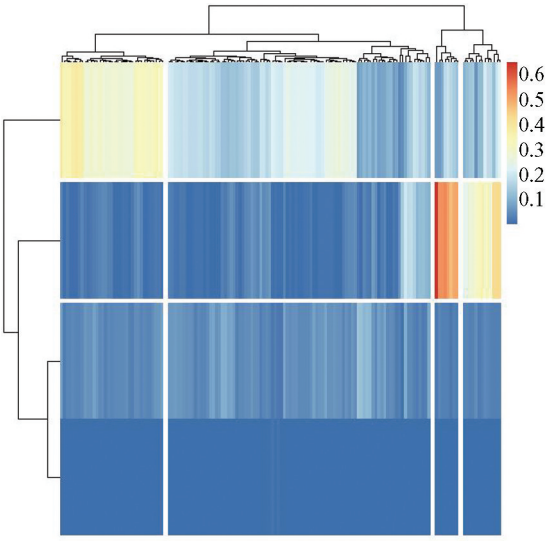

图 16 K-means 4 聚类热图

Fig.16 K-means 4 clustering heat map

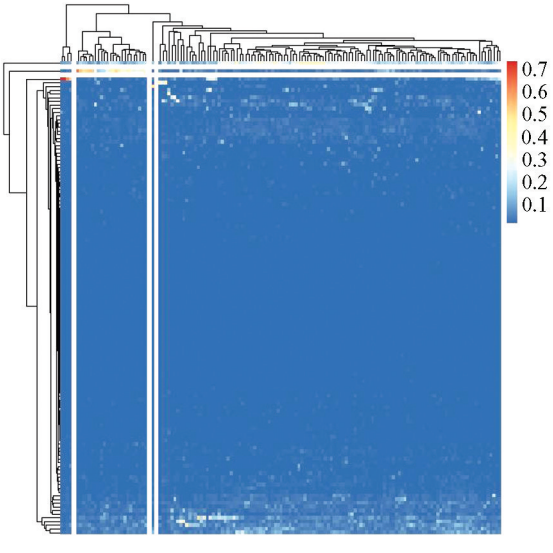

图 17 K-means 1796 聚类热图

Fig.17 K-means 1796 clustering heat map

综上,运用 DPMM+kmeans 进行分析时,发现当 $K=3$ 时,AIC/BIC 最佳,并且 AIC 与 BIC 的走势整体一致说明,观察观测参数变化良好。采用多层狄利克雷过程与传统生物信息学分析得到的结果一致,研究结果表明,DPMM+K-means 模型不仅能够有效地区分 2-型糖尿病患者样本间的相似性,而且能从中发现 DPMM+K-means 模型还能鉴定出影响菌群结构异质性最大的 OTUs 菌。

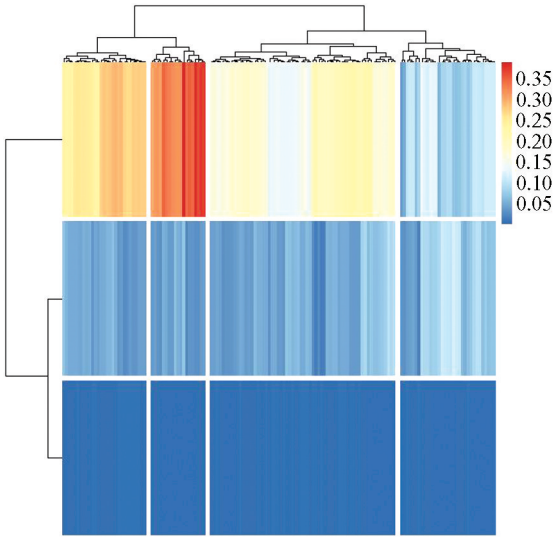

图 18 PDMM 改进K-means聚类热图

图 18 PDMM improved K-means clustering heat map

5 总结

本研究主要是基于狄利克雷过程混合模型与 K-means 聚类算法结合的 2-型糖尿病肠道菌群研究,通过对肠道菌群 OTUs 数据进行分析,建立了一个完整、系统化以及可视化的模型。将分析的结果与传统的生物信息学分析进行对比,发现当$K=3$时,AIC/BIC 最佳,并且 AIC 与 BIC 的走势整体一致说明,观察观测参数变化良好。从热力图也可以看出 $K=3$时,第一,二,三类对样本的反映效果更加明显。算法将菌群数据分为三个族,并且能够找出贡献最突出的 OTU,具有实际意义,为研究肠道菌群提供了一种新的方法。分析每一个 OTU 的贡献情况。能够找出贡献最突出的 OTU 分别为贡献最突出的 OTU 为 OTU963、OTU363、OTU1390、OTU19、OTU359,为进一步对比 DPMM 优势,采用热力图分析,将样本数据分为 4 类,对 OTUs 用 K-means 分别聚成 2 类、4 类、1 796 类,用 DPMM 改进的 K-means 进行聚类,得出结果与传统生物信息学分析一致。

参考文献(References)

[1] HILL-BURNS E M, DEBELIUS J W, MORTON J T, et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome[J]. Movement Disorders, 2017, 32(5): 739-749. DOI: 10.1002/mds.26942.

[2] KOREN O, KNIGHTS D, GONZALEZ A, et al. A guide to enterotypes across the human body: meta-analysis of micro-

bial community structures in human microbiome datasets[J]. PLoS Computational Biology, 2013, 9(1): e1002863.DOI: 10.1371/journal.pcbi.1002863.

[3]HONG B Y, FURTADO ARAUJO M V, STRAUSBAUGH L D, et al. Microbiome profiles in periodontitis in relation to host and disease characteristics[J]. Plos One, 2015, 10(5): e0127077. DOI:10.1371/journal.pone.0127077.

[4]HOLMES I, HARRIS K, QUINCE C. Dirichlet multinomial mixtures: generative models for microbial metagenomics[J]. PLoS ONE, 2012, 7(2): e30126. DOI: 10.1371/journal.pone.0030126.

[5]DONG Mei, LI Longhai, CHEN Man, et al. Predictive analysis methods for human microbiome data with application to Parkinson's disease[J]. PLoS One, 2020, 15(8): e0237779.DOI: 10.1371/journal.pone.0237779.

[6]CHEN Jun, LI Hongzhe. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis[J]. The Annals of Applied Statistics, 2013, 7(1):418-442. DOI:10.1214/12-AOAS592.

[7]YANG Yuqing, WANG Xin, XIE Kaikun, et al. Inferring multiple metagenomic association networks based on variation of environmental factors[J/OL]. https://www.biorxiv.org/content/10.1101/2020.03.04.976423v1, 2020. DOI:10.1101/2020/03.04.976423.

[8]SHI Yushu, ZHANG Liangliang, DO K A, et al. Bayesian approaches for flexible and informative clustering of microbiome data[J/OL]. https://arxiv.org/abs/2007.15812.2020. DOI: 10.48550/arXiv.2007.15812.

[9]IRANNIA Z B, CHEN T. TACO: Taxonomic prediction of unknown OTUs through OTU co-abundance networks[J]. Quantitative Biology, 2016, 4(3):149-158.DOI:10.1007/s40484-016-0073-2.

[10]MAO Jialiang, MA Li. Dirichlet-tree multinomial mixtures for clustering microbiome compositions[J]. The Annals of applied statistics, 2020, 16(3): 1476 - 1499. DOI: 10.48550/arXiv.2008.00400.

[11]NICCOLAI E, BALDI S, RICCi F, et al. Evaluation and comparison of short chain fatty acids composition in gut diseases[J]. World Journal of Gastroenterology, 2019, 25(36):5543-5558.DOI:10.3748/wjg.v25.i36.5543.

[12]HARRIS K, PARSONS T L, IJAZ U Z, et al. Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical dirichlet process[J]. Proceedings of the IEEE, 2017, 4(3): 516 - 529. DOI:10.1109/JPROC.2015.2428213.

[13]PENG Y, SHAW C A. An efficient algorithm for accurate computation of the Dirichlet-multinomial log-likelihood function[J]. Bioinformatics, 2014(11):1547-1554.DOI: 10.1093/bioinformatics/btu079.