



A new method for mining information of gut microbiome with probabilistic topic models

Xin Xiong^{1,2} · Minrui Li¹ · Yuyan Ren¹ · Xusheng Yao¹ · Yuhui Du⁴ ·
Qingsong Huang^{1,2} · Xiangyang Kong⁴ · Jianfeng He^{1,3}

Received: 28 April 2021 / Revised: 17 March 2022 / Accepted: 12 September 2022 /

Published online: 13 October 2022

© The Author(s) 2022

Abstract

Microbiome is closely related to many major human diseases, but it is generally analyzed by the traditional statistical methods such as principal component analysis, principal coordinate analysis, etc. These methods have shortcomings and do not consider the characteristics of the microbiome data itself (i.e., the “probability distribution” of microbiome). A new method based on probabilistic topic model was proposed to mine the information of gut microbiome in this paper, taking gut microbiome of type 2 diabetes patients and healthy subjects as an example. Firstly, different weights were assigned to different microbiome according to the degree of correlation between different microbiome and subjects. Then a probabilistic topic model was employed to obtain the probabilistic distribution of gut microbiome (i.e., per-topic OTU (operational taxonomic units, OTU) distribution and per-patient topic distribution). Experimental results showed that the output topics can be used as the characteristics of gut microbiome, and can describe the differences of gut microbiome over different groups. Furthermore, in order to verify the ability of this method to characterize gut microbiome, clustering and classification operations on the distributions over topics for gut microbiome in each subject were performed, and the experimental results showed that the clustering and classification performance has been improved, and the recognition rate of three groups reached 100%. The proposed method could mine the information hidden in gut microbiome data, and the output topics could describe the characteristics of gut microbiome, which provides a new perspective for the study of gut microbiome.

Keywords Probabilistic topic model · Latent Dirichlet Allocation · Gut microbiome · Type 2 diabetes mellitus

✉ Xiangyang Kong
1745982615@qq.com

✉ Jianfeng He
jfhenghe@foxmail.com

Extended author information available on the last page of the article

1 Introduction

Microbiome has been linked to many major human diseases, including obesity, diabetes, autism, allergies, inflammatory bowel disease, cardiovascular disease, many types of cancer and depression and so on [28]. Therefore, human microbiome may become the latest therapeutic intervention targets and thus play an important role in the diagnosis, analysis and treatment of these diseases [42]. For example, correlation analysis of gut microbiome has been successfully applied in clinical assessment and patient diagnosis of diabetes [38]. At present, microbiome research has not only fully demonstrated its great value in clinical medicine and personalized medicine [12], but also penetrated into many fields such as Marine science [27], environmental science [51], agricultural science [51] and earth science [17]. According to the official website of the International Diabetes Federation (IDF) in November 2021, the number of adult diabetic patients in the world has reached 537 million, accounting for 10.5% of the total population of the world; China's diabetes mellitus has grown to 140 million, ranking first in the world (<https://diabetesatlas.org/>). Diabetes has become a major public health problem that seriously affects people's physical and mental health. Therefore, taking microbiome of patients with type 2 diabetes as an example in this paper, probabilistic topic model was employed to mine hidden information in microbiome and then infer the probability topics related to type 2 diseases, which will provide a new perspective for the study of microbiome, and may provide new targeted microbiological treatment for type 2 diabetes.

2 Related work

The study of microbiome usually analyzes the community composition and diversity of bacteria, which is used to study one of the basic problems of microbial ecology: how many different *taxa* or *OTUs* (*operational taxonomic units*) are present? Usually, multivariate statistics or pattern recognition methods are employed to identify different structural patterns in gut microbiome, such as principal component analysis (PCA) [32, 33, 38], principal coordinate analysis (PCoA) [13, 39, 49], partitioning around medoid (PAM) clustering [2, 57], etc. However, microbial metagenomics data is characterized by high diversity but sparseness. These methods have some inherent shortcomings and cannot deal with microbiome data well. The probabilistic topic model is not sensitive to highly sparse and noisy data, which is not only widely used in the field of document mining, but also used in microbiome data analysis to mine hidden topics.

2.1 Traditional methods

Both PCA and PCoA are dimension reduction techniques, as shown in Table 1. The advantages of PCA and PCoA are simple and easy to use, low cost, easy to understand results, and no parameter restrictions. The disadvantages are: (1) the data information cannot be retained well in the case of complete ignorance of the data. For example, PCA needs data preprocessing and standardization. The usual way to standardize is to divide by the standard deviation. There may be a problem here. If the standard deviation is very small and close to zero, especially for the data polluted by noise, the standard deviation of noise has a more significant effect on data amplification, while the data that is not polluted by noise has less amplification effect. (2) The final number of reduced dimension, that is, the number of potential latent variables, cannot be

Table 1 The principles, advantages and disadvantages of PCA, PCoA, PAM, and probabilistic topic models

	Principles	Advantages	Disadvantages
PCA and PCoA	<p>PCA transforms data into a new coordinate system through linear transformation; then the first largest variance of the data projection is on the first coordinate (called the first principal component), and the second largest variance is on the second coordinate (the second principal component).</p> <p>PCoA is similar to PCA. After sorting through a series of eigenvalues and eigenvectors, the most important coordinates in the distance matrix are found by selecting the first few eigenvalues.</p>	Simple and easy to use; low cost; easy to understand results; no parameter restrictions	In the case of complete ignorance of data, the data information cannot be retained well. The final number of reduced dimension, that is, the number of potential latent variables, cannot be well estimated.
PAM	<p>PAM is a kind of clustering partition algorithm, also known as K-medoid algorithm, which refers to using the center point to represent a cluster. The basic idea of PAM algorithm: the purpose is to divide N objects in data set D into K clusters, and then randomly select one member in each cluster as the center point. Then, in each step, the members who are not the center point are compared one by one according to the difference or distance between them and the center point to see whether they can become center point.</p>	Low sensitivity to outliers	It is needed to specify the K value; it is very effective for small data, but it does not have good scalability for large data sets.
Probabilistic topic models	<p>Probabilistic topic model is a generative model, in which not as representing the community, but the sample is treated as having being generated by sampling from the community. The basic idea is that a document is regarded as a mixture of latent topics, each of which is expressed by a distribution on words (these items are related to document mining, but in other fields “document” and “word” have different meanings).</p>	Not sensitive to highly sparse and noisy data	Unsupervised, it cannot use the prior information of the data; assuming that the topics are independent of each other, the relationship between the topics cannot be obtained.

well estimated. PAM is a kind of clustering partition algorithm, also known as K-medoid algorithm, which refers to using the center point to represent a cluster. The advantage of PAM is that the sensitivity to outliers is greatly reduced, because the class center it selects is a

specific point rather than a geometric center (such as K-means). The disadvantages are: (1) it is needed to specify the K value; (2) it is very effective for small data sets, but it does not have good scalability for large data sets.

2.2 Probabilistic topic model

Microbial metagenomics data is characterized by high diversity but sparseness, and most taxa appear only in a few samples with low abundance. In addition, the samples vary in reads: a small sample will inherently be noisier than a larger one. Therefore, PCA, PCoA and PAM do not work well for such data sets. The probabilistic topic model is not sensitive to highly sparse and noisy data, so it is more suitable for microbial metagenomics data. According to the probabilistic topic model, not as representing the community, but the sample is treated as having being generated by sampling from the community, in which the most natural assumption to make is sampling with replacement, so that the likelihood of an observed sample is a multinomial distribution with a parameter vector where a given item represents the probability that a read is from a given taxa [25]. The natural prior of polynomial distribution parameters is Dirichlet. This is the widely used probabilistic topic model - Latent Dirichlet Allocation (LDA). The basic idea of LDA is that a document is regarded as a mixture of latent topics, each of which is expressed by a distribution on words (these items, such as document, topics and words, are related to document mining because this method was first applied in the field of natural language processing, and in other fields “document” and “word” have different meanings). LDA employs two Dirichlet-Multinomial distributions to model the relationship between documents and topics, and the relationship between topics and words respectively [22, 36]. Approximate methods, such as variational inference [6] and Markov chain Monte Carlo (MCMC) [29], are commonly used in LDA to calculate the posterior probabilities. The calculated probability distributions are employed to make inference about the topics and documents.

LDA has been widely used in document mining [5, 6, 19, 20, 23] and image retrieval and annotation [14, 31, 45]. It also has been applied in bioinformatics for various purposes, such as protein structure representation [43], drug labeling [4], and next generation sequence [59]. However, the study of applying probabilistic topic model to gut microbiome is scarce. Zhang et al. exploited LDA to boost metagenomic reads binning [58]. Chen et al. showed that the configuration of functional groups in meta-genome samples can be inferred by probabilistic topic modeling (LDA) [10]. Holmes et al. applied Dirichlet multinomial mixtures (DMM) model to gut microbiome of the fat and thin twins [25]. Stewart et al. used DMM to model the 16 S rRNA gene sequencing and metagenomic sequencing data of children gut microbiome [49]. Wang et al. used LDA to study gut microbiome of patients with mild hepatic encephalopathy and the efficacy of rifampicin combined with probiotics [53, 54]. Woloszynek et al. evaluated a topic model approach for parsing microbiome data structure [56]. Abe et al. proposed a new probabilistic model for microbial association analysis, because traditional probabilistic modeling cannot distinguish between the bacterial differences derived from enterotype and those related to a specific disease [1]. Okui et al. proposed a Bayesian nonparametric topic model for microbiome data using subject attributes [35]. These studies indicate that there are some meaningful findings in the analysis of gut microbiome by probabilistic topic model, which is different from the traditional statistical methods. To the best of our knowledge, there is no research on the combination of surveillance information (i.e. the degree of correlation between different microbiome and patients) and probabilistic topic

model. In this study, gut microbiome data of patients with type 2 diabetes were taken as a case. In addition, previous studies only used probabilistic topic models to cluster gut microbiome, or to find topics, but did not make full use of the characteristics of gut microbiome itself and in-depth analysis of the output of probabilistic topic model.

Therefore, the contributions of this paper include: (1) a new model based on probabilistic topic model was proposed to analyze gut microbiome, which could mine the information hidden in gut microbiome data, and the output topics could describe the characteristics of gut microbiome; (2) the distributions generated by LDA model could be combined with various data mining algorithms as new features, which is helpful for us to well understand the structural differences of gut microbiome among different groups.

3 Methods

Weights of *OTUs* were calculated firstly, and then different weights were assigned to different microbiome, and then employed LDA to obtain the distribution of gut microbiome in different groups. Finally, the distribution over topics for gut microbiome in each patient (i.e., per-patient topic distributions) generated by LDA were clustered and classified to verify its ability to characterize gut microbiome. The flow chart of the proposed method is shown in Fig. 1. Firstly, the data set (relative abundance of gut microbiome in Fig. 1a) is acquired, and then the weight of each *OTU* (Fig. 1b) is calculated, and the procedure to obtain the weights is shown in Fig. 2; secondly, the LDA model is employed (Fig. 1c), represented by two distributions: the distribution over *OTUs* for each topic (per-topic *OTU* distributions) and the distribution over topics for gut microbiome in each patient (per-patient topic distributions); thirdly, Gibbs collapsed sampling [37] is employed to determine the optimal number of topics (Fig. 1d), and the analysis results are visualized in a tree graph [15] (Fig. 1e); finally, the per-patient topic distributions generated by LDA are clustered and classified to verify their ability to characterize the gut microbiome data (Fig. 1f).

3.1 Calculating weights of *OTUs*

In the field of document mining, it is necessary to first convert the terms in document into the DocumentTermMatrix (DTM), that is, the frequency of each term (word or vocab) in each document. For gut microbiome data, relative abundance reflects the proportion of different bacteria in the samples, that is, corresponding to the DocumentTermMatrix (DTM). Probabilistic topic model was originally designed for document analysis, which assumes that the importance of each word in document is the same. However, this assumption is not perfect. Wallach et al. pointed out that the high frequency stopwords had a great influence on the topic inference of probabilistic topic model [52]. In the human intestines, the distributions of gut microbiome are also different, among which the dominant floras are *Bacteroidetes* and *Firmicutes* [46], which may have an impact on the inference of disease-related topics. In addition, when the distributions generated by LDA were directly used to construct classifiers, it is found that different *OTUs* of different groups played different roles. Therefore, according to the different importance of *OTUs*, the weights of *OTUs* were calculated and multiplied by relative abundance, so as to adjust the proportion of different microbiome. The flow chart of weight calculation is shown in Fig. 2.

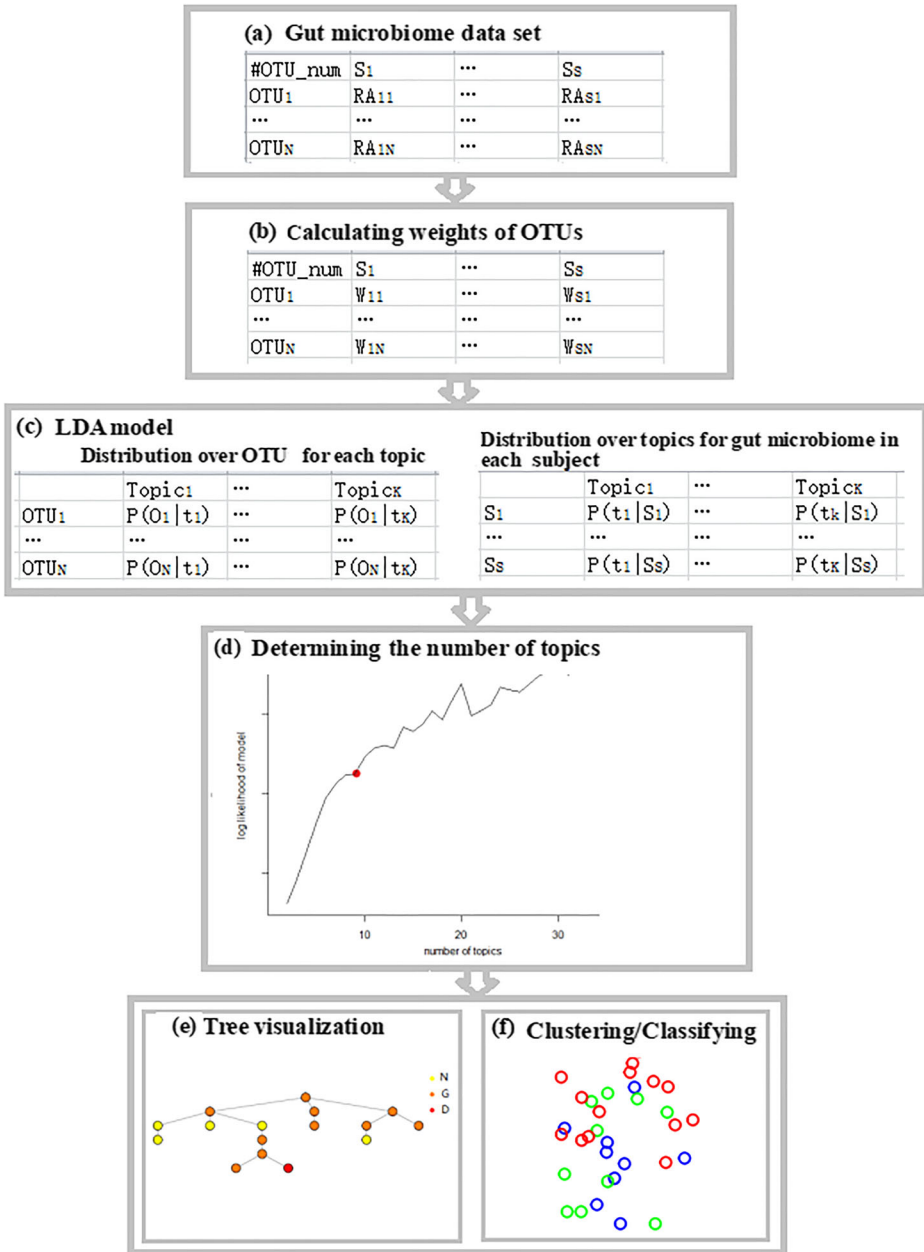


Fig. 1 A flowchart of the proposed method. **a** shows the gut microbiome data set, where OTU₁, ..., OTU_N denote the name of OTUs, S_1 , ..., S_s represent subjects, and RA represents relative abundance. **b** shows the weights of each OUT, which are calculated according to formula (1)–(4). **c** shows LDA modeling after calculating weights, where t_k is the k-th topic, O_N is the n-th OTU, $P(O_N | t_k)$ is the conditional probability; S_s is the s-th subject, $P(t_k | S_s)$ is the conditional probability. **d** shows the determination of the optimal number of topics in LDA model. **e** shows the tree graph of three groups according to the results of LDA model. Yellow, orange and red circles represent normal health subjects (abbreviated as N), T2DM with genetic autonomic neuropathy (abbreviated as G) and T2DM (abbreviated as D) respectively. **f** shows clustering and classification operations based on the results of LDA model

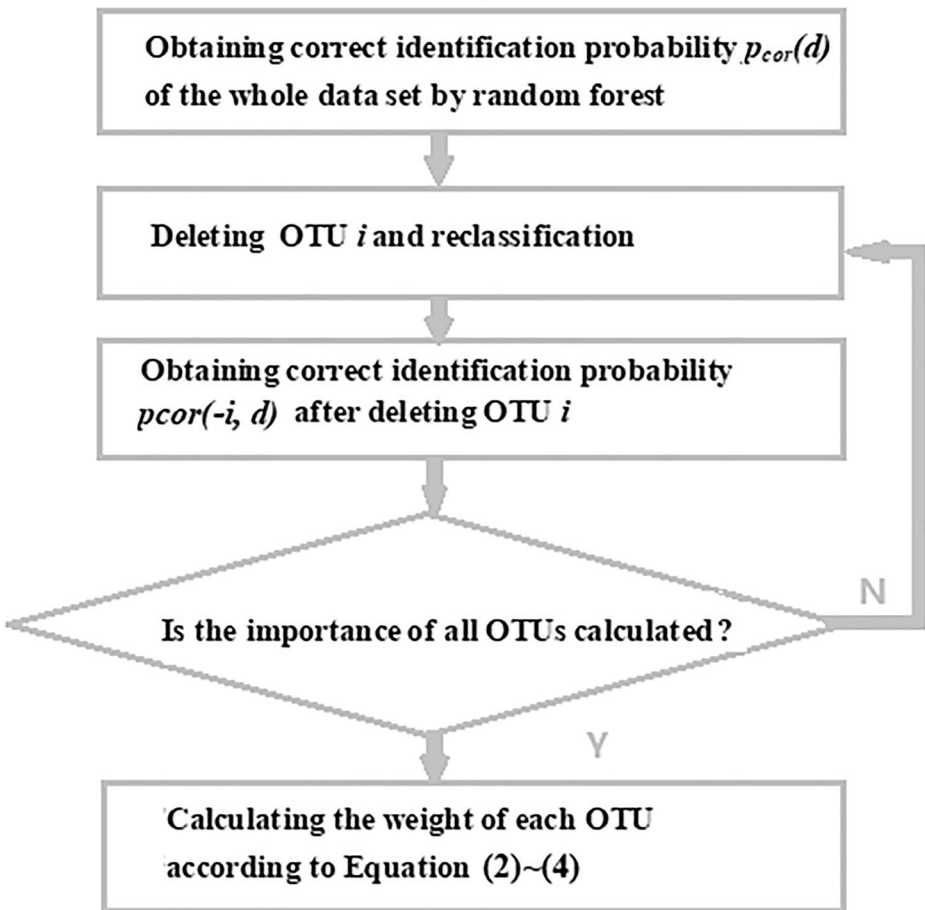


Fig. 2 A flowchart of weight calculation

The importance of missing OTUs can be measured by the ratio of the recognition rate of data set after deleting an OTU to the recognition rate of the whole data set, as shown in formula (1):

$$I = \log \frac{p(d')}{p(d)} \quad (1)$$

Where $p(d)$ is the correct identification probability of the whole data set, $p(d')$ is the correct identification probability of missing an OTU. Obviously, the greater the difference between $p(d)$ and $p(d')$ is, the greater the absolute value of I is. The $p(d)$ is determined for given data set, so the larger the value of I is, the greater the change of classification accuracy after deleting an OTU is, the higher the importance of the OTU is; conversely, if the value of I is smaller, it shows that the classification accuracy changes slightly after deleting the OTU, and the effect of the OTU on classification is relatively small. In this paper, random forest is employed to calculate the correct identification probability, as shown in Fig. 2.

In order to prevent the weights of some OTUs to be too large, the formula (1) is slightly modified according to the references [21, 47], which is replaced by the absolute value of the subtraction of $\log \frac{p_{err}(-i,d)}{p_{err}(d)}$ and $\log \frac{p_{cor}(-i,d)}{p_{cor}(d)}$, in which the normalization factor is added, as shown in formula (2)~(3). Therefore, $I(i)$ is defined as the change of identification accuracy after deleting the OTU i , \bar{I} is the average value of $I(i)$, α is the number of OTUs 1551, and $weight(i)$ is the weight of the OTU i , as follows:

$$I(i) = \left| \log \frac{p_{err}(-i,d)}{p_{err}(d)} - \log \frac{p_{cor}(-i,d)}{p_{cor}(d)} \right| \tag{2}$$

$$\bar{I} = \frac{1}{\alpha} \sum_{i=1}^{\alpha} \left| \log \frac{p_{err}(-i,d)}{p_{err}(d)} - \log \frac{p_{cor}(-i,d)}{p_{cor}(d)} \right| \tag{3}$$

$$weight(i) = \frac{I(i)}{\bar{I}} \tag{4}$$

Where $p_{err}(-i, d)$ is the error identification probability after the missing OTU i , $p_{cor}(-i, d)$ is the correct identification probability after the missing OTU i , $p_{err}(d)$ is the error identification probability of the whole data set, similarly, $p_{cor}(d)$ is the correct identification probability of the whole data set.

3.2 Latent Dirichlet Allocation model

In this study, 140 subjects were recruited, and the gut microbiome of each subject included 1551 OTUs. According to LDA model, each patient’s gut microbiome was treated as one document and each OTU as one word, so that the data was composed of 140 documents and each document was composed of 1551 words. The algorithm is as follows [6]:

1. for each topic k , where k in $\{1 \dots K\}$, pick a distribution over OTUs $\varphi_k \sim Dir(\beta)$;
2. for each patient P_m , where m in $\{1 \dots M\}$,
 - a. pick a distribution over topics $\theta_m \sim Dir(\theta)$;
 - b. for each OTU O_n with n in $\{1 \dots N\}$,
 - (1) pick a topic $z \sim Multinomial(\theta_m)$;
 - (2) pick OTU $O_n \sim Multinomial(\varphi_z)$;

Where, implied variables θ and φ can be estimated according to Eqs. (5) and (6):

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^K (n_m^{(k)} + \alpha_k)} \tag{5}$$

$$\phi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^V (n_k^{(t)} + \beta_t)} \tag{6}$$

Where, φ_k is a distribution over OTUs for topic k , θ_m is a distribution over topics for patient m , $n_m^{(k)}$ represents the number of OTUs with topic k in patient m , $n_k^{(t)}$ denotes the number of OTUs with topic k in the OTU t , and V denotes the total number of OTUs without repetition. *Dir* represents a Dirichlet distribution and *Multinomial* represents a multinomial distribution. The distribution of OTUs for topics and the distribution of topics for patients are viewed as random variables obeying Dirichlet distributions with parameters β and α , respectively.

The initial value of α is $50/k$, where k is the number of topics and the initial value of β is 0.1 [55]. In the original LDA model published by Blei et al. [6, 24], variational EM algorithm was used to estimate unknown parameters $\theta_{m,k}$ and $\phi_{k,t}$, and later researchers found that Gibbs sampling was also a good method to infer unknown parameters [37].

3.3 Choosing the number of topics

The main parameter of LDA is to determine the number of topics k (optimal values for other hyper-parameters (i.e., α and β) are automatically picked by the different fitting methods). The generally-recommended method to select the number of topics is to use cross-validation with different values of k , looking at the likelihood for each topic number [15]. However, the computation time for such a method may be prohibitive on large data sets and large range of topic numbers. In addition, a large number of topics (and therefore a more complex statistical model) may lead to over fitting. Therefore, it is preferable to use the smallest possible number that provides a good explanation of the data. However, because of the loose significance of the concept of ‘topics’ in the context of gut microbiome, it is difficult to give a reliable estimate of the ideal number based on biological knowledge alone. Three fitting methods are provided in the Celltree package [15], namely Gibbs, VEM and maptpx. In Gibbs method, Collapsed Gibbs Sampling method [37] is used to infer the parameters of the Dirichlet distributions for a given number of topics. It gives high accuracy but is very time-consuming to run on a large amount of data sets. In VEM method, Variational Expectation-Maximisation [24] is used, which tends to converge faster than Gibbs collapsed sampling, but with lower accuracy. In Maptpx method, the method described in [44] is used, which estimates the parameters of the topic model for increasing number of topics (using previous estimates as a starting point for larger topic numbers). In this study, maptpx method was adopted firstly and it is found that the optimal number of topics was more than 100, which was obviously inappropriate and not well explained our data. Therefore, the Gibbs method was finally adopted. For more related information, please see the Section 5 of this paper.

3.4 Clustering analysis

In this study, the distributions generated by LDA (the per-patient topic distributions) were regarded as the features of gut microbiome [59], and then the conventional clustering method (k-means) was adopted for verifying cluster performance. Since the data sets included three groups (N, G, D groups), the number of clusters was set as 3 (N, G and D 3 groups) or 2 (G and D 2 groups) in the k-means method. The per-patient topic distributions were equivalent to perform a dimensionality reduction on the original data. The traditional PCA method was also used to reduce dimensionality, so that the number of obtained principal component from PCA was equal to the number of topics from the per-patient topic distributions to facilitate comparison. Clustering results were evaluated by Adjusted Rand Index (ARI) [26], with values ranging from 0 to 1. Generally, the higher the value is, the better the clustering performance is.

3.5 Classification analysis

To further evaluate the ability of the per-patient topic distributions to characterize gut microbiome, the distributions were employed to construct classifiers, such as support vector machine (SVM) [48] and random forest (RF) [7], to identify 3 or 2 groups of patients. 70% of each group was used as the training set, and the remaining 30% was used as the test set. In order to illustrate the performance of the proposed method in classification, the weights and the per-patient topic distribution were calculated on the training set (70% of the data set), and the performance was verified on the test set (30% of the data set). In this study, the function “svm” (with “Polynomial” kernel and optimized values of parameters γ and $cost$ under different classification tasks) in R package “e1071” and function “randomForest” (with number of trees setting as 500 and default values of other parameters) in R package “randomForest” were utilized to train the classifiers.

4 Experiment results

In this study, 140 cases of gut microbiome data were collected from the Department of Endocrinology in Yunnan First People’s Hospital, China, from 2015 to 2017, 74 cases of patients with T2DM (abbreviated as D), 27 cases of patients with T2DM with gastrointestinal autonomic neuropathy (abbreviated as G) and 39 cases of the normal healthy subjects (abbreviated as N). All subjects signed the informed consent, and the experiment was approved by the ethics committee of Kunming University of Science and Technology. No antibiotics, probiotics or lactose were used for all subjects within one month before sampling.

4.1 Experimental environment

The experiments in this paper are run on the computer of Intel(R) Core(TM) Ci9-9900k CPU @ 3.60 GHz and 32G RAM. And R 4.0.0 (<https://www.r-project.org/>) is employed for all data processing and plotting in this study. The LDA modeling and result visualization are completed by celltree software package [15] based on R language.

4.2 Weights of OTUs

The weight of each OTU was calculated in two cases: 3 classifications for N, G and D groups and 2 classifications for G and D groups. The weight calculation process is shown in Fig. 2. The weights of OTUs for 2- and 3-classification are shown in Fig. 3a and b. There are six grades, i.e., 3.10, 2.45, 1.75, 0.98, 0.88 and 0.11 for 3-classification in Fig. 3a. There are five OTUs with the largest weight 3.10, namely OTU108 (*Ruminiclostridium*), OTU365 (*Mollicutes*), OTU855 (*Lachnospiraceae*), OTU1586 (*Nitrosomonadaceae*) and OTU1793 (*Clostridiales*). The mean relative abundances of these five OTUs in N, G and D groups are shown in Fig. 4. There are five grades, i.e., 5.21, 3.19, 2.24, 1.70, 0.06 for 2-classification in Fig. 3b. There is only one OTU with the largest weight 5.21, OTU253 (*Gemella*), and only one OTU with the second largest weight 3.19, OTU857 (*Prevotella*). The mean relative abundances of the two OTUs in G and D groups are shown in Fig. 5a. For the other three smaller grades, the mean relative abundances of OTU7, OTU35 and OTU1 are plotted in Fig. 5b. It can be found that the plots of OTU253 and OTU857 with larger weight have a significant

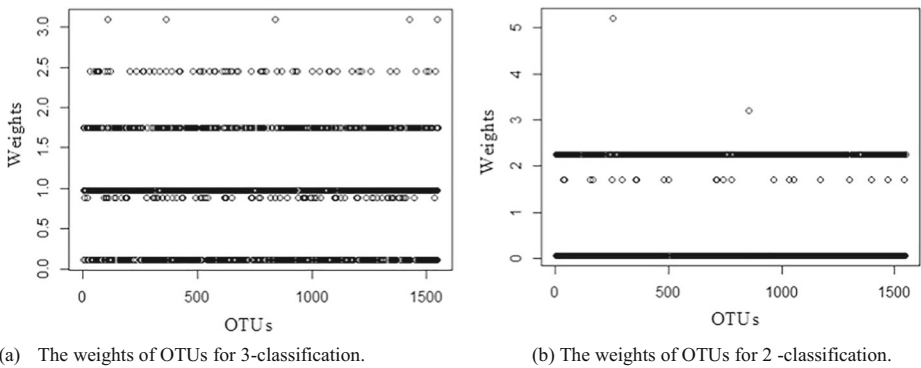


Fig. 3 The weights of OTUs for 3-classification (N, G and D groups) and 2-classification (G and D groups). N - Normal healthy subjects, D - Patients with T2DM, G - Patients with T2DM with gastrointestinal autonomic neuropathy patients

difference in G and D groups, while the plots of OTU7, OTU35 and OTU1 with smaller weight have a small difference in G and D groups

4.3 Topic analysis

The size of gut microbiome data of three groups inputted into LAD model is 1551*140, and the optimal number of topics is 12, as shown Fig. 11. The size of per-patient topic distributions is 140*12, whose heat map is shown in Fig. 6, in which three groups are shown in red, blue and green on the right side. The size of per-topic OTU distributions is 1551*12, whose heat map is shown in Fig. 7. The top 10 OTUs with high probability of the 12 topics are listed in Table 2, in which the names of OTUs at the generic level are indicated. These OTUs in each topic are arranged in descending order of probability. The size of gut microbiome of two groups (G, D groups) is 1551*101, and the number of topics is also 12. The size of per-patient topic

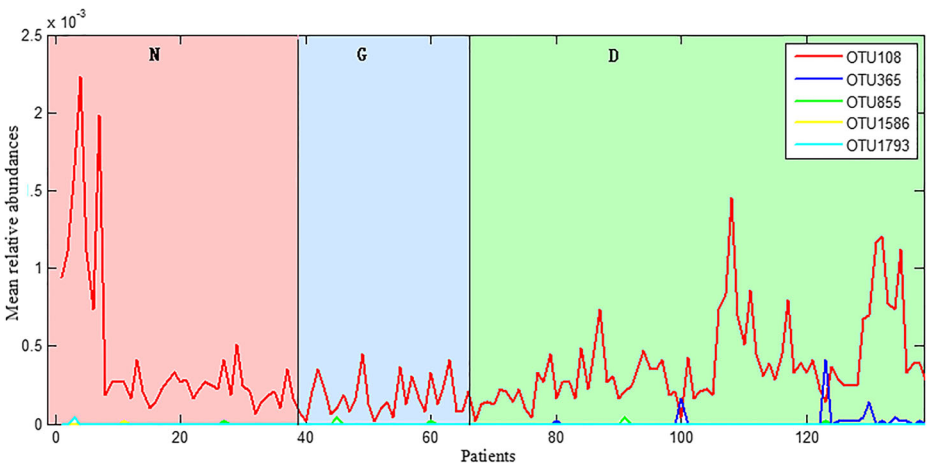
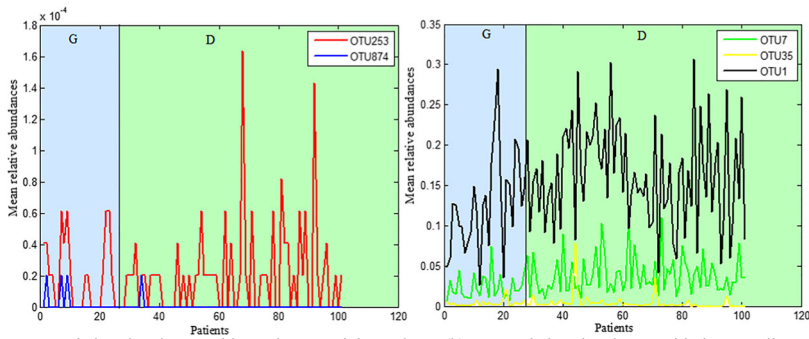


Fig. 4 Five mean relative abundances with the largest weight for 3-classification. N - Normal healthy subjects, D - Patients with T2DM, G - Patients with T2DM with gastrointestinal autonomic neuropathy patients



(a) Mean relative abundances with two larger weight grades. (b) Mean relative abundances with three smaller weight grades.

Fig. 5 Mean relative abundances with five weight grades for 2-classification, N - Normal healthy subjects, D - Patients with T2DM, G - Patients with T2DM with gastrointestinal autonomic neuropathy patients

distributions is 101×12 , as shown in Fig. 8. The size of per-topic OTU distributions is also 1551×12 , as shown in Fig. 9, of which the top 10 OTUs with high probability are listed in Table 3.

As shown in Fig. 6, topic 6 is mainly spread among N group, and this topic covers most healthy subjects. Topics 5, 8 and 12 are mainly spread among D group, and these three topics account for about one third of D group respectively. Topics 4 and 7 are mainly spread among G group. Two-thirds of topic 7 is spread among G group, and the other one-third is spread

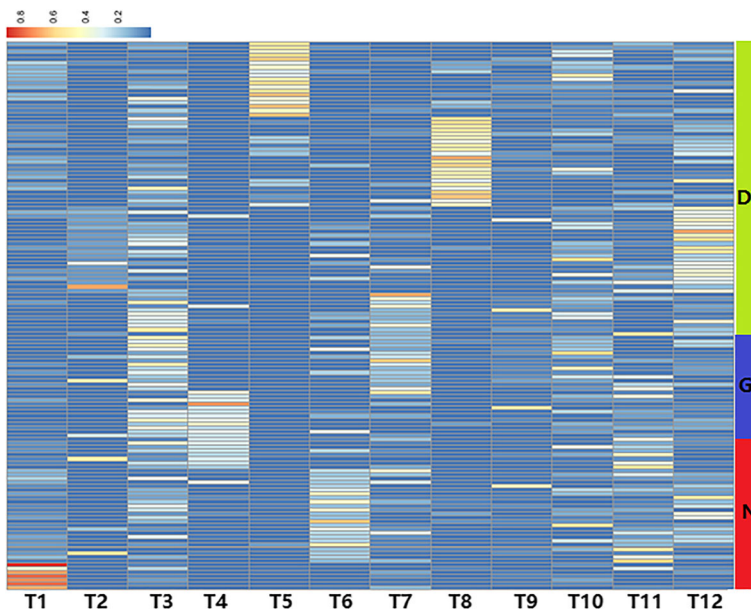


Fig. 6 Heat map of the per-patient topic distributions of N, G and D groups when the topic number is 12. The color histogram from blue to red shows the value of the topic probability of patient ranged from 0 to 1. On the right side of the graph, three groups of 140 subjects are shown in red, blue and green. Topic 6 is mainly spread among N group, topics 5, 8 and 12 are mainly spread among D group, and topics 4 and 7 are mainly spread G group. N - Normal healthy subjects, D - Patients with T2DM, G - Patients with T2DM with gastrointestinal autonomic neuropathy patients

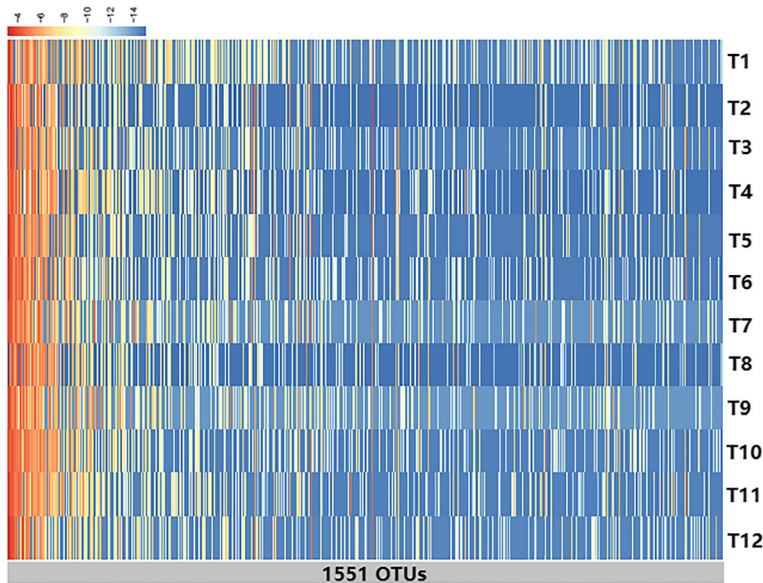


Fig. 7 Heat map of the per-topic OTU distributions of N, G and D groups when the topic number is 12. The color histogram from blue to red shows the OTU probability of topics ranged from -14 to -4 . The first 150 OTUs are with high probability among the 12 topics. N - Normal healthy subjects, D - Patients with T2DM, G - Patients with T2DM with gastrointestinal autonomic neuropathy patients

among D group. Topics 3, 10 and 11 are widely spread among three groups. Topics 2 and 9 are less spread among three groups.

In Fig. 7, the first 150 OTUs among the 12 topics are with high probability. From Table 2 at order level, the top 10 OTUs of the topic 6 of N group are *Clostridiales* (*Romboutsia*, *Pseudobutyrvibrio*, *Faecalibacteri*, *Lachnospiraceae* and *Roseburia* belong to *Clostridiales*), *Bacteroidales*, *Burkholderiales* (*Parasutterella* belong to *Burkholderiales*). The top 10 OTUs of the topics 5, 8 and 12 of D group are *Bacteroidales* (*Bacteroides*, *Prevotella*), *Lactobacillus*, *Fusobacteriaceae*, *Clostridiales* (*Ruminococcus*, *Romboutsia*, *Roseburia*), *Enterobacteriales* (*Escherichia*), *Selenomonadales* (*Phascolarctobacterium*). The top 10 OTUs of the topics 4 of G group are *Bifidobacteriales*, *Selenomonadales* (*Megamonas*), *Bacteroidales* (*Bacteroides*, *Prevotella*), *Enterobacteriales* (*Escherichia*), *Burkholderiales* (*Parasutterella*), *Lactobacillus*. The top 10 OTUs of the topic 7 are *Lactobacillus*, *Enterobacteriales* (*Escherichia*), *Bacteroidales*, *Clostridiales* (*Romboutsia*), *Selenomonadales* (*Veillonella*)

In Fig. 8, topics 3, 9 and 10 are mainly spread among D group, and these three topics account for about one third of D group respectively, similar to topics 5, 8 and 12 of D group in Fig. 6. Topic 5 is mainly spread among G group. Two-thirds of topic 1 is spread among G group, and the other one-third is spread among D group. As shown in Table 3 at order level, the top 10 OTUs of the topics 3, 9 and 10 of D group are *Bacteroidales* (*Bacteroides*, *Prevotella*, *Parabacteroides*), *Clostridiales* (*Ruminococcus*, *Pseudobutyrvibrio*, *Lachnospiraceae*), *Burkholderiales* (*Parasutterella*), *Selenomonadales* (*Megamonas*). The top 10 OTUs of the topic 5 of G group are *Lactobacillus* (*Lactobacillus*, *Streptococcus*), *Bacteroidales* (*Bacteroides*, *Prevotella*, *Parabacteroides*), *Enterobacteriales* (*Escherichia*). Compared with topic 7 in Table 2, there are fewer *Clostridiales* (*Romboutsia*) and *Selenomonadales* (*Veillonella*). This is because two-thirds of topic 7 is spread among G group

Table 2 Top 10 OTUs with high probability of topic 4, 5, 6, 7, 8, 12 of N, G and D groups

Topic ID	Topic-10 most probable OTUs									
T4	OTU32 <i>Bifidobacteriales</i>	OTU23 Megamonas	OTU5 Bacteroides	OTU4 Prevotella	OTU2 Escherichia	OTU6 Bacteroides	OTU49 Parasutterella	OTU1042 Prevotella	OTU7 Lactobacillus	OTU9 Bacteroides
T5	OTU4 Prevotella	OTU7 Lactobacillus	OTU6 Bacteroides	OTU24 Ruminococcus	OTU2 Escherichia	OTU13 Bacteroides	OTU469 Prevotella	OTU14 Bacteroides	OTU9 Bacteroides	OUT36 Prevotella
T6	OTU3 Romboutsia	OTU11 Pseudobutyrvibrio	OTU8 Faecalibacteri	OTU5 Bacteroides	OTU6 Bacteroides	OTU12 Parasutterella	OTU27 Lachnospiraceae	OTU703 Lachnospiraceae	OTU16 Roseburia	OTU71 Lachnospiraceae
T7	OTU112 Lactobacillus	OTU2 Escherichia	OTU3 Romboutsia	OTU6 Bacteroides	OTU17 Veillonella	OTU9 Bacteroides	OTU13 Bacteroides	OTU68 Veillonella	OTU1385 Veillonella	OTU21 Parabacteroides
T8	OTU4 Prevotella	OTU14 Bacteroides	OTU6 Bacteroides	OTU7 Lactobacillus	OTU5 Bacteroides	OTU18 Fusobacterium	OTU2 Escherichia	OTU469 Prevotella	OTU13 Bacteroides	OTU3 Romboutsia
T12	OTU18 Fusobacterium	OTU1 Bacteroides	OTU9 Bacteroides	OTU7 Lactobacillus	OTU2 Escherichia	OTU14 Bacteroides	OTU13 Bacteroides	OTU15 Phascolarctobacterium	OTU53 Bacteroides	OTU16 Roseburia

N Normal healthy subjects, *D* Patients with T2DM, *G* Patients with T2DM with gastrointestinal autonomic neuropathy patients

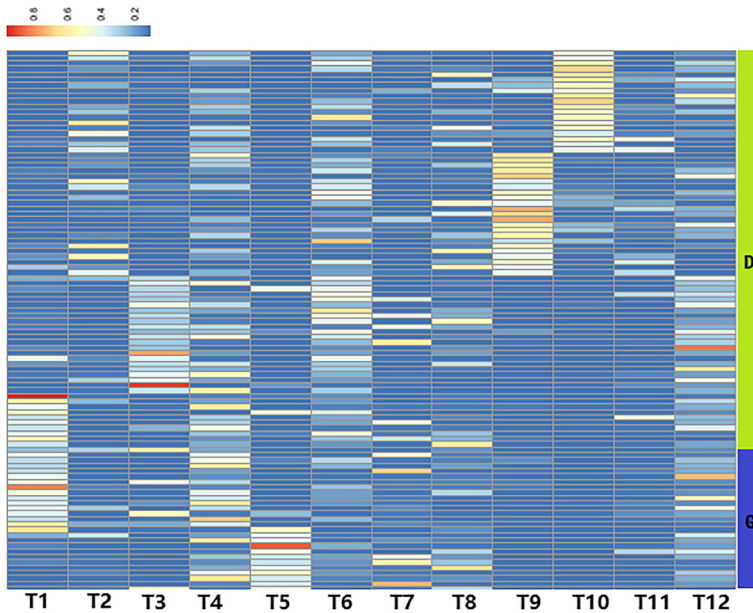


Fig. 8 Heat map of the per-patient topic distributions of G and D groups when the topic number is 12. D - Patients with T2DM, G - Patients with T2DM with gastrointestinal autonomic neuropathy patients

and the other one-third is spread among D group in Table 2, while topic 5 is all spread among G group in Table 3. The top 10 OTUs of the topic 1 are *Bacteroidales* (*Bacteroides*, *Prevotella*), *Selenomonadales* (*Megamonas*, *Phascolarctobacterium*, *Veillonella*), *Burkholderiales* (*Parasutterella*).

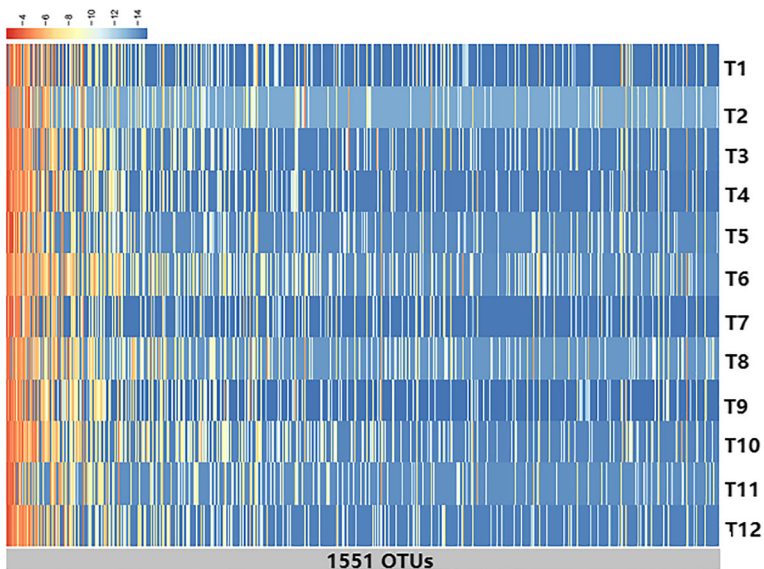


Fig. 9 Heat map of the per-topic OTU distributions of G and D groups when the topic number is 12. D - Patients with T2DM, G - Patients with T2DM with gastrointestinal autonomic neuropathy patients

Table 3 Top 10 OTUs with high probability of topic 1, 3, 5, 9, 10 of G and D group

Topic ID	Topic-10 most probable OTUs									
T1	OTU4	OTU10	OTU1042	OTU7	OTU15	OTU78	OTU53	OTU19	OTU117	OTU34
	Prevotella	Megamonas	Prevotella	Bacteroides	Phascolarctobacterium	Parasutterella	Bacteroides	Bacteroides	Bacteroides	Veillonella
T3	OTU4	OTU7	OTU39	OTU13	OTU12	OTU25	OTU24	OTU53	OTU98	OTU11
	Prevotella	Bacteroides	Prevotella	Bacteroides	Parasutterella	Parabacteroides	Ruminococcus	Bacteroides	Bacteroides	Pseudobutyrvibrio
T5	OTU112	OTU136	OTU19	OTU13	OTU2	OTU25	OTU124	OTU7	OTU4	OTU66
	Lactobacillus	Lactobacillus	Bacteroides	Bacteroides	Escherichia	Parabacteroides	Streptococcus	Bacteroides	Prevotella	Bacteroides
T9	OTU4	OTU7	OTU13	OTU24	OTU10	OTU1042	OTU11	OTU27	OTU39	OTU53
	Prevotella	Bacteroides	Bacteroides	Ruminococcus	Megamonas	Prevotella	Pseudobutyrvibrio	Lachnospiraceae	Prevotella	Bacteroides
T10	OTU4	OTU77	OTU35	OTU7	OTU10	OTU57	OTU42	OTU27	OTU11	OTU12
	Prevotella	Prevotella	Bacteroides	Bacteroides	Megamonas	Lachnospiraceae	Lachnospiraceae	Lachnospiraceae	Pseudobutyrvibrio	Parasutterella

N Normal healthy subjects, *D* Patients with T2DM, *G* Patients with T2DM with gastrointestinal autonomic neuropathy patients

4.4 Topic visualization

In order to visualize the representation of the topics generated by LDA to different groups, Celltree software package [15] is employed to visualize the generated topics with tree graph. Extracting a hierarchical structure from the lower-dimensional model follows the same general idea as other methods for dimensionality reduction (i.e., PCA or ICA): firstly computing a matrix of pairwise distance, of which the chi-square distance [9] is used to compare the topic histograms. Then this distance matrix obtained may be used with various tree building algorithms to identify the underlying tree structure. One natural way to visualize such a structure is using a minimum spanning tree (MST). As shown in Fig. 10, yellow, orange and red nodes represent N, G and D groups respectively. The left half of tree is N group and a small number of G group, and the right half are D group and the remaining G group. The plot of topics is shown in supplement material S1. Each node in the graph represents one subject, and the color sectors in the node represent 12 topics. For different subjects, the proportion of 12 topics is different. The backbone tree are shown in supplement material S2 ~ S3. Large nodes represent the trunk of tree and small nodes represent branches of tree. The tree graph and backbone tree of G and D groups are shown in supplement material S4 ~ S5.

4.5 Clustering results

In order to verify the performance of the proposed method, k-means clustering was performed on the original data, the per-patient topic distributions (12 topics) generated by LDA (LDA means that LDA model is used directly, that is, the weights of all OTUs are equal) and wLDA (wLDA means that is the proposed method in this paper, it means that the weights of all OTUs are calculated according to formula (2)~(4) and multiplied by the relative abundance, and then LDA model is employed.), 12 principal components of PCA. Clustering performance was

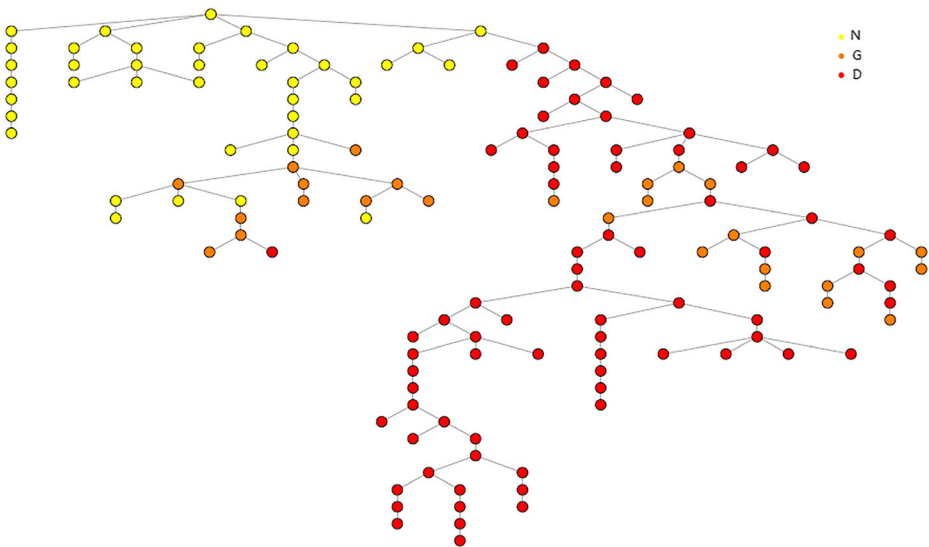


Fig. 10 Grouping tree graph of N, G and D groups. Yellow, orange and red nodes represent N, G and D groups respectively. N - Normal healthy subjects, D - Patients with T2DM, G - Patients with T2DM with gastrointestinal autonomic neuropathy patients

Table 4 Comparison of clustering performance ARI.

	2 classifications for G and D groups	3 classifications for N, G and D groups
Original data	0.9411	0.8122
PCA	0.9587	0.5679
LDA	1	1
wLDA	1	1

N Normal healthy subjects, *D* Patients with T2DM, *G* Patients with T2DM with gastrointestinal autonomic neuropathy patients

measured by ARI, as shown in Table 4. It can be seen that the ARI of LDA and wLDA is equal to 1, which is better than that of the original data, and the ARI of PCA for 3-classifications is worst.

4.6 Classification results

SVM and RF were used to train classifiers to compare the per-patient topic distributions generated by LDA and wLDA, and the classified accuracy is shown in Table 5. As can be seen from Table 5, the 3-classification accuracy of original data + SVM is low, only 0.5952, and the accuracy of LDA + SVM and wLDA + SVM is significantly improved, reaching to 0.8571 and 1 respectively. The 2-classification accuracy of LDA + SVM and LDA + RF is the same as that of original data. While the 2- and 3-classification accuracy of wLDA + SVM and wLDA + RF is significantly improved, and that of 3-classification accuracy of wLDA + SVM reaches 1. It should be noted that when calculating weights, the weights are different for different classification tasks. In order to compare with wLDA, the number of topics of LDA selected here is also 12.

5 Discussion

Using unsupervised learning or clustering methods to determine clusters of communities or envirotypes is a hot issue in the analysis of microbial community data. However, previous studies mostly adopted methods such as PCA [32, 33, 38], PCoA [13, 39, 49], PAM clustering [2, 57]. Since there are some inherent problems in microbiome data [25], new methods are needed. In this study, a new method based on probabilistic topic model was proposed to analyze gut microbiome of N, G and D groups.

Table 5 Comparison of classification accuracy

	2 classifications for G and D groups	3 classifications of N, G and D groups
Original data+SVM	0.7741	0.5952
Original data+RF	0.8709	0.8333
LDA+SVM	0.7741	0.8571
LDA+RF	0.8709	0.8333
wLDA+SVM	0.9032	1
wLDA+RF	0.8709	0.9524

N Normal healthy subjects, *D* Patients with T2DM, *G* Patients with T2DM with gastrointestinal autonomic neuropathy patients

To study the roles of different OTUs in three groups, the weight of each OTU was calculated. The results showed that these OTUs varied greatly in different groups, leading to a large change in recognition rate, and thus large weights were acquired according to the formula (2 ~ 4). It could be found that these OTUs have higher correlation with T2DM in this way, but it cannot be inferred that these OTUs are indicators to distinguish different groups and that these OTUs are dominant in a certain group.

When the probabilistic topic model was used, the topics distributions and corresponding OTUs of different groups were obtained through the per-patient topic distributions and the per-topic OTU distributions. It was easy to find out which OTUs had changed and which OTUs of topics were dominant. However, our previous experiments found that the OTUs with highest probability of topics of LDA for original relative abundance data were all *Bacteroides*. This is not surprising, because the most abundant bacteria in the human gut are *Bacteroidetes* and *Firmicutes* [46]. In addition, the per-patient topic distributions of LDA were employed as features for clustering and classification, in which the recognition rate was expected to improve further. Could OTUs with high probability associated with each group be found? Could the recognition rate of LDA be further improved? Therefore, combining the weight information with LDA model was considered.

Applying the proposed method to gut microbiome of N, G and D groups, the per-patient topic distributions and the per-topic OTU distributions are shown in Figs. 6, 7, 8 and 9. *Fusobacterium*, *Enterobacteriales* and *Selenomonadales* of D group increase, and *Clostridiales* and *Burkholderiales* decrease compared with N group. *Lactobacillus*, *Bacteroidales* and *Enterobacteriales* of G group increase, and *Clostridiales* decrease. This is particularly evident in topic 5 of Table 3, which is spread throughout G group. This indicates that patients with T2DM have moderate intestinal dysregulation, which is consistent with some previous studies [30, 38, 41]. Qin et al. [38] found that in the intestinal tract of T2DM, the abundance of butyric acid bacteria such as *Clostridiales* (*Roseburia* and *Faecalibacteri*) decreased, and the abundance of some opportunistic pathogens such as *Bacteroidales* and *Enterobacteriales* increased. Sato et al. [41] found that *Lactobacillus* increased in the intestinal tract of T2DM, and some *Clostridiales* that could produce butyric acid in short-chain fatty acids (SCFAs) decreased. Karlsson et al. [30] found that the abundance of 4 *Lactobacillus* increased and that of 5 *Clostridium* decreased in T2DM. In addition, when calculating the weights of OTUs, 7 OTUs with larger change was found, among which there were no *Lactobacillus*, *Bacteroidales* and *Enterobacteriales*. While weight information is combining with probabilistic topic model, the situation becomes different. Using wLDA, topic 5 was mainly distributed in G group, and the most probable OTUs were *Lactobacillus*, *Bacteroidales* and *Enterobacteriales*. It shows that probabilistic topic model has advantages in mining hidden information, and the information mined by adding the supervised information is more targeted.

The concept of “topic” in probabilistic topic model comes from the field of document mining and is used to mine the hidden topics in the document set. It is similar to the concept of “cluster” in clustering methods. The clustering method is used to determine how many clusters are in the gut microbiome data, while the probabilistic topic model determines how many topics. The determination method is shown in the related explanation of Fig. 11 (12 topics). And these topics are reflected by the two distributions “per-patient topic distributions” and “per-topic OTU distributions”, as shown in Figs. 6 and 7. In order to further visualize the representation of the topics generated by LDA, Celltree software package is employed to visualize the generated topics with tree graph [15], as shown in supplement material S1. Grouping tree graphs can be drawn by calculating the distance according to the topics obtained, where three groups can be clearly distinguished in the grouping tree graphs.

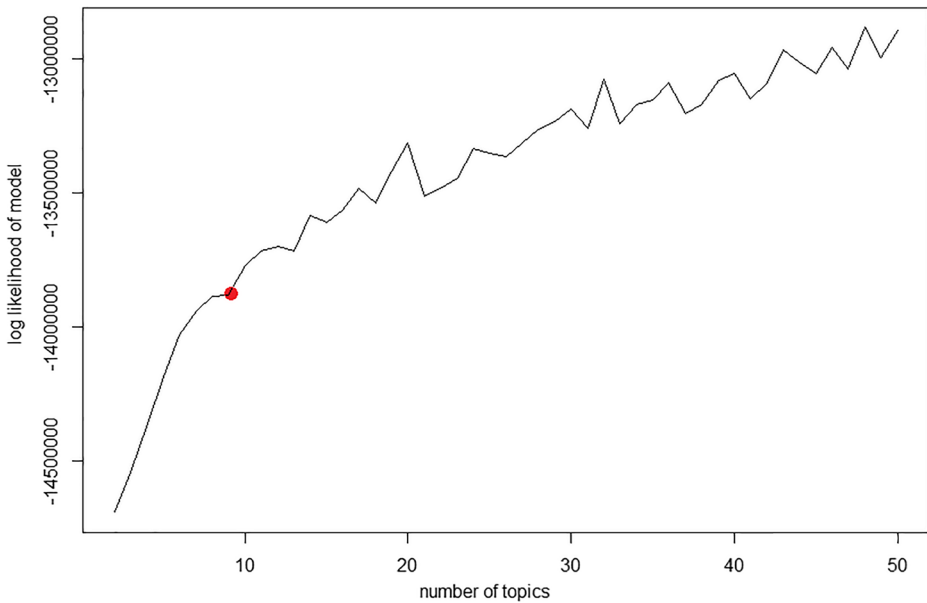


Fig. 11 A log likelihood value graph with the number of topics from 2 to 50. The red dot in the graph indicates that when K is 12, the log likelihood reaches the local minimum

The clustering effect of G group on the tree graphs is not very good (Fig. 10). In fact, from the output of wLDA (per-patient topic distributions (Fig. 6)), it could be found that topic 6 is mainly spread among N group; topics 5, 8, and 12 are mainly spread among D group; while topics 4 and 7 are mainly spread among G group. While three-quarters of topic 4 is in G group, the other quarter is in N group, and two-thirds of topic 7 is in G group, the other third is in D group. This indicates that there are no topics that belong entirely to G group. Topics 4 and 7 are shared with N and D groups, which reflect that the clustering effect of G group is not as good as that of N and D groups on the tree graph.

To objectively illustrate the effectiveness of our proposed method, the output results of wLDA, LDA and PCA were clustered and classified, and their performance was measured by two indicators, ARI and recognition rate. The ARI of LDA and wLDA reaches 1. The performance of wLDA + SVM and wLDA + RF are significantly improved. This indicates that the combining weighted information and probabilistic topic model is effective. In addition, data transformations may influence classifier performance. In subsequent studies, preprocessing the gut microbiome data, such as centered logratio transformation (CLR) will be considered, because the microbiome data itself has the problem of zero-inflated [11], and then further verify its impact on statistical results and classification performance.

About the number of topics, the maptx method provided by Celltree software package firstly was used [44], which determines the optimal number of topics by judging whether the log likelihood value monotonously decreases in three consecutive iterations. The optimal number of topics given by this method for gut microbiome of three groups is greater than 100, which obviously cannot meet our needs, because fewer and more representative topics in gut microbiome are expected. Therefore, Gibbs method [37] was employed to draw a log likelihood value graph with the number of topics from 2 to 50, as shown in Fig. 11. As the number of topics increases, the log likelihood value keeps increasing, which is similar to the

result of maptpx method. Theoretically, the higher the likelihood value is, the better the model is, but a large number of topics cannot be used to interpret our data. Therefore, the topic number 12 corresponding to the local extremum of the curve is taken as the optimal topic number, i.e., the red dot in Fig. 11. After determining the number of topics 12, the per-patient topic distributions and the per-topic OTU distributions are shown in Figs. 6 and 7. To further validate this discovery, the number of topics from 13 to 50 using Gibbs method again is taken. And the experimental results show that although the number of topics increases in the hot map of the per-patient topic distributions, the unique topics of the three groups do not change, as detailed in Appendix (K takes 13, 15, 20, 30, 40 and 50, respectively).

About the computational cost of LDA model, time complexity: $O(N_{\text{iter}}KN_pL(L-1)/2)$, where N_{iter} is the number of iterations, K is the number of topics, N_p is the number of patients, and L is the length of OTUs; Space complexity: $N_pK + WK + N_p*L$, where N_p*K is per-patient topic distribution, and $W*K$ is per-topic OTU distribution. When the number of topics is 2 ~ 50 and the number of patients is 140, the running time of maptpx method of LDA is 100.53s, which of Gibbs method is 2188.04s.

About T2DM with gastrointestinal autonomic neuropathy, according to statistics, about 4% ~ 22% of diabetes patients may have diarrhea [16, 40]. Brock et al. found that 50% of patients with long-course diabetes have severe gastrointestinal symptoms [8]. The mechanism by which diabetes induces chronic diarrhea is not well understood. However, some scholars believe that visceral autonomic neuropathy caused by diabetes leads to intestinal dysfunction, leading to irritable bowel, increased secretion and increased stool frequency and stool thinning, which constitute the basis for the onset of chronic diarrhea induced by diabetes [3, 18]. However, there are not many studies on the relationship between diabetic diarrhea and gut microbiome [34]. Virally-Monod's studies showed that the intestinal bacteria of patients with diabetes accompanied by chronic diarrhea were excessively proliferated, with the incidence of 43% [50]. In this paper, it is found that *Lactobacillus*, *Bacteroidales* and *Enterobacteriales* increased and *Clostridiales* decreased in the intestinal tract of T2DM with gastrointestinal autonomic neuropathy. However, there are only 27 cases of these patients, so it is still necessary to increase the number of these patients and conduct a large number of studies to research the relationship between T2DM with gastrointestinal autonomic neuropathy and gut microbiome.

Finally, a new model based on probabilistic topic model was proposed to analyze gut microbiome of T2DM in this study. *Fusobacterium*, *Enterobacteriales* and *Selenomonadales* of T2DM increased, and *Clostridiales* and *Burkholderiales* decreased. *Lactobacillus*, *Bacteroidales* and *Enterobacteriales* of T2DM with gastrointestinal autonomic neuropathy increased, and *Clostridiales* decreased. This provides a new perspective for us to study gut microbiome, and may provide new targeted microbiological treatment for type 2 diabetes. In addition, the distributions generated by LDA model can be combined with various data mining algorithms as new features, which will have great application potential and will be helpful for us to well understand the structural differences of gut microbiome among different populations.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11042-022-13916-7>.

Funding This study was funded by Grant Nos. 82060329, 11265007, 81860318 from National Natural Science Foundation of China (NSFC), No. 2020J0052 from Scientific Research Fund Project of Yunnan Education Department of China.

Declarations

Ethics approval and consent to participate The experiment was approved by the ethics committee of Kunming University of science and technology.

Competing interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Abe K, Hirayama M, Ohno K, Shimamura T (2019) ENIGMA: an enterotype-like unigram mixture model for microbial association analysis. *BMC Genom* 20(Suppl 2):191
2. Arumugam M, Raes J (2011) Eric Pelletier. Enterotypes of the human gut microbiome. *Nature* 473(7346): 174–180
3. Azpiroz F, Malagelada C (2015) Diabetic neuropathy in the gut: pathogenesis and diagnosis[J]. *Diabetologia* 59(3):1–5
4. Bisgin H, Liu Z, Kelly R, Fang H, Xu X, Tong W (2012) Investigating drug repositioning opportunities in FDA drug labels through topic modeling. *BMC Bioinformatics* 13(15):S6
5. Blei D, Jordan M (2003) Modeling annotated data. *The Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp 127–134
6. Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
7. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
8. Brock C (2013) Diabetic autonomic neuropathy affects symptom generation and brain-gut axis. *Diabetes Care* 36:3698–3705
9. Chardy P, Glemarec M, Laurec A (1974) Application of inertia methods of benthic marine ecology: Practical implications of the basic options. *Estuar Coast Mar Sci* 4:179–205
10. Chen X, He T, Hu X (2012) Estimating functional groups in human gut microbiome with probabilistic topic models. *IEEE Trans Nanobiosci* 11(3):203–215
11. Chen L, Reeve J, Zhang L (2018) GMPR: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6(2):e4600
12. Costello E, Stagaman K, Dethlefsen L (2012) The application of ecological theory toward an understanding of the human microbiome. *Science* 336(6086):1255–1262
13. Cotillard A, Kennedy S, Kong L (2013) Dietary intervention impact on gut microbial gene richness. *Nature* 500:585–588
14. Datta R, Joshi D, Li J, Wang J (2008) Image retrieval: ideas, influences, and trends of the new age. *ACM Comput Surv* 40(2):5
15. Duverle A, Yotsukura S, Nomura S (2016) CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics* 17(1):363
16. Ebert EC (2005) Gastrointestinal complications of diabetes mellitus. *Dis Mon* 51(12):620–663
17. Falkowski P, Fenchel T, Delong E (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* 320(5879):1034–1039
18. Gould M, Sellin JH (2009) Diabetic diarrhea[J]. *Curr Gastroenterol Rep* 11(5):354–359
19. Griffiths T, Steyvers M (2004) Finding scientific topics. *Proc Natl Acad Sci USA* 101(suppl. 1):5228–5235
20. Ha C, Iran D, Van N, Than K (2019) Eliminating overfitting of probabilistic topic models on short and noisy text: the role of dropout. *Int J Approx Reason* 112(SEP.):85–104
21. Hao J, Xie J, Su J, Xu X, Han X (2016) An unsupervised approach for sentiment classification based on weighted latent dirichlet allocation. *CAAI Trans Intell Syst* 11(4):539–545

22. Hofmann T (1999) Probabilistic latent semantic indexing. Annual international ACM SIGIR conference on Research and development in information retrieval, pp 50–57
23. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn* 42:177–196
24. Hoffman M, Blei D, Bach F (2010) Online learning for latent dirichlet allocation. In: Lafferty J, Williams CKI, Shawe-Taylor J, Zemel R, Culotta A (Eds) *Advances in neural information processing systems*, 23, pp 856–864
25. Holmes I, Harris K, Quince C (2012) Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* 7(2):e30126
26. Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2(2–3):193–218
27. Iverson V, Morris R, Frazar C (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. *Science* 335(6068):587–590
28. Jiang X, Hu X (2015) Big data research in microbiome. *Math Model Appl* 4(3):6–18
29. Jordan M (1999) *Learning in graphical models*. MIT Press, Cambridge
30. Karlsson F, Tremaroli V, Nookaew I (2013) Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498(7452):99–103
31. Laib L, Allili S, Ait-Aoudia S (2019) A probabilistic topic model for event-based image classification and multi-label annotation. *Sig Process Image Commun* 76:283–294
32. Lambeth S, Carson T, Lowe J (2015) Composition, diversity and abundance of gut microbiome in prediabetes and type 2 diabetes. *J Diabetes Obes* 2(3):1–7
33. Larsen N, Vogensen F, van den Berg F (2010) Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE* 5(2):e9085
34. Li X, Wang Y, Li Z et al. (2015) The Correlation between intestinal flora and diabetes: research progress. *Chin J Microecol* 27(10):1224–1228
35. Okui T (2020) A Bayesian nonparametric topic model for microbiome data using subject attributes. *IPSPJ Trans Bioinf* 13:1–6
36. Papadimitriou C, Tamaki H, Raghavan P, Vempala S (1998) Latent semantic indexing: a probabilistic analysis. *ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp 159–168
37. Phan X, Nguyen L, Horiguchi S (2008) Learning to classify short and sparse text & web with hidden topics from large-scale data collections. *Proceedings of the 17th international conference on world wide web*. ACM
38. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490(7418):55–60
39. Rajpal D, Klein J, Mayhew D (2015) Selective spectrum antibiotic modulation of the gut microbiome in obesity and diabetes rodent models. *PLoS ONE* 10(12):e0145499
40. Rayner CK et al (2001) Relationships of upper gastrointestinal motor and sensory function with glycemic control. *Diabetes Care* 24(2):371–381
41. Sato J, Kanazawa A, Ikeda F (2014) Gut dysbiosis and detection of “live gut bacteria” in blood of Japanese patients with type 2 diabetes. *Diabetes Care* 37(8):2343–2350
42. Sekirov I, Finlay B (2009) The role of the intestinal microbiota in enteric infection. *J Physiol* 587(17):4159–4167
43. Shivashankar S, Srivathsan S, Ravindran B, Tendulkar A (2011) Multi-view methods for protein structure comparison using latent dirichlet allocation. *Bioinformatics* 27(13):161–168
44. Taddy M (2012) On estimation and selection for topic models. In: *AISTATS*, pp 1184–1193
45. Tian D, Shi Z (2020) A two-stage hybrid probabilistic topic model for refining image annotation. *Int J Mach Learn Cybernet* 11(2):417–431
46. Tremaroli V, Backhed F (2012) Functional interactions between the gut microbiota and host metabolism. *Nature* 489(7415):242–249
47. Turney D, Littman L (2003) Measuring praise and criticism: inference of semantic orientation from association. *ACM Trans Inform Syst* 21(4):315–346
48. Vapnik V (1995) *The nature of statistical learning theory*. Springer, New York
49. Vatanen T, Franzosa E, Schwager R (2018) The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* 562(7728):589–594
50. Vially-Monod M, Tielmans D, Kevorkian JP et al (1999) Chronic diarrhoea and diabetes mellitus: prevalence of small intestinal bacterial overgrowth[J]. *Diabet Metab* 24(6):530–536
51. von Mering C, Hugenholtz P, Raes J (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315(5815):1126–1130
52. Wallach H (2006) Topic modeling: beyond bag-of-words. *International conference on machine learning*. ACM

53. Wang X, Zuo Z, Zhou L (2017) Microbial flora structure based on probability topic model. *Sci Sin Vitae* 47: 1220–1234
54. Wang X, Zuo Z, Fan H (2018) Study of the structure of intestinal microflora in patients with mild hepatic encephalopathy based on probability topic model. *Acta Microbiol Sinica* 58(7):1274–1286
55. Wei X, Croft W (2006) LDA-based document models for Ad-hoc retrieval. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp 178–185
56. Woloszynek S, Zhao Z, Simpson G, O'Connor P, Mell G (2017) Evaluating a topic model approach for parsing microbiome data structure. *bioRxiv*, pp 176412–17636
57. Wu G, Chen J, Hoffmann C (2011) Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science* 334(6052):105–108
58. Zhang R, Cheng Z, Guan J, Zhou S (2015) Exploiting topic modeling to boost metagenomic reads binning. *BMC Bioinformatics* 16(5):S2
59. Zhao W, Chen J, Perkins R (2016) A novel procedure on next generation sequencing data analysis using text mining algorithm. *BMC Bioinformatics* 17(1):301

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Xin Xiong^{1,2} · Minrui Li¹ · Yuyan Ren¹ · Xusheng Yao¹ · Yuhui Du⁴ ·
Qingsong Huang^{1,2} · Xiangyang Kong⁴ · Jianfeng He^{1,3}

Xin Xiong
xiongxin840826@163.com

Minrui Li
2389020966@qq.com

Yuyan Ren
1713397685@qq.com

Xusheng Yao
527864290@qq.com

Yuhui Du
429819408@qq.com

Qingsong Huang
1912443688@qq.com

¹ Faculty of Information Engineering and Automation, Kunming University of Science and Technology, 650500 Kunming, China

² Computer Technology Application Key Laboratory of Yunnan Province, Kunming University of Science and Technology, 650500 Kunming, China

³ Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, 650500 Kunming, China

⁴ Faculty of Medicine, Kunming University of Science and Technology, 650500 Kunming, China