A NON-PARAMETRIC BAYESIAN CLUSTERING FOR GENE EXPRESSION DATA

Liming Wang and Xiaodong Wang

Department of Electrical Engineering
Columbia University, New York, NY 10027
{liming, wangx}@ee.columbia.edu

ABSTRACT

Clustering is an important data processing tool for interpreting microarray data and genomic network inference. In this paper, we propose a non-parametric Bayesian clustering algorithm based on the hierarchical Dirichlet processes (HDP). The proposed clustering algorithm captures the hierarchical features prevalent in biological data such as the gene express data by introducing a hierarchical structure in the model. We develop a Gibbs sampling algorithm based on the Chinese restaurant metaphor. We conduct experiments on the yeast galactose datasets and yeast cell cycle datasets by comparing our clustering results to the standard results. The proposed clustering algorithm is shown to outperform several popular clustering algorithms by revealing the underlying hierarchical structure of the data. The experiments also show that the proposed clustering algorithm provides more information and reduces the unnecessary clustering fragments than the clustering algorithm based on Dirichlet mixture model.

Index Terms— Hierarchical Dirichlet processes, Dirichlet processes, clustering, microarray data

1. INTRODUCTION

With the development of the microarray technology, one often needs various algorithms to investigate the gene functions and regulation relations contained in the high-volume microarray data. Clustering is considered to be an important tool for analyzing the biological data [1]. The aim of clustering is to group the data into disjoint subsets, where in each subset the data show certain similarities to each other.

Numerous clustering methods have been proposed. One large category can be characterized as the distance-based algorithm. That is, a distance is first defined for clustering purpose and then the clusters are formed based on the distances of the data. Typical examples in this category include the K-means algorithm and the self-organizing map algorithm. These algorithms are based on simple rules, and they often suffer from robustness issue, i.e., they are sensitive to noise which is extensive in biological data.

Another important category of clustering methods is the model-based algorithms. Specifically, data are assumed to be generated by some mixture distribution. Each component of the mixture corresponds to a cluster. However, the separated estimations of the number of clusters and the mixture parameters make this approach sensitive to noise. In order to cope with the above sensitivity problem of the finite-mixture model, one may set the prior by using the Dirichlet processes [2]. Such kind of methods is often called the non-parametric approach.

Hierarchical clustering is yet another more advanced approach, which groups together the data with similar features based on the underlying hierarchical structure.

It is well-known that many genes share different levels of functionalities. The resemblances of different genes are commonly represented at different levels of perspectives, e.g., at the cluster level instead of individual gene level. In this case, we desire to have a hierarchical clustering algorithm recognizing the gene resemblances not at the single gene level but at the higher cluster level, to avoid unnecessary fragmental clusters that impede the proper interpretation of the biological information.

In this paper, we propose a non-parametric Bayesian clustering algorithm for gene expression data based on the hierarchical Dirichlet process (HDP) [3]. The HDP model incorporates the merits of both the infinite-mixture model and the hierarchical clustering. The hierarchical structure is introduced to allow sharing data among related clusters. Moreover, as a non-parametric approach, our clustering method does not need to assume a fixed number of clusters *a priori*.

2. SYSTEM MODEL AND PROBLEM FORMULATION

We assume that the gene expression data are random samples from some underlying distributions. All data in one cluster are generated by the same distribution. Suppose that for the mircoarray data, there are N genes in total. For each gene we conduct M experiments. Let g_{ji} denote the expression of the i-th gene in the j-th experiment, $1 \leq i \leq N$ and $1 \leq j \leq M$. For each g_{ji}, we associate a latent membership variable z_{ji}, which indicates the cluster membership of g_{ji}. That is, if genes g_{ji} and $g_{j'i'}$ are in the same cluster, we have $z_{ji} = z_{j'i'}$. Note that z_{ji} is supported on a countable set such as $\mathbb{N}$ or $\mathbb{Z}$.

For each g_{ji}, we associate a coefficient $\theta_{z_{ji}}$ whose index is determined by its membership variable z_{ji}. In order to have a Bayesian approach, we also assume that for each coefficient θ_k is drawn independently from a prior distribution G_0

$$\theta_k \sim G_0. \tag{1}$$

The membership variable $\mathbf{z} = \{z_{ji}\}_{j,i}$ has a discrete distribution

$$\mathbf{z} \sim \pi. \tag{2}$$

We assume the each g_{ji} is drawn independently from a distribution $F(\theta_{z_{ji}})$

$$g_{ji} \sim F(\theta_{z_{ji}}), \tag{3}$$

where $\theta_{z_{ji}}$ is a coefficient associated with g_{ji} and F is a distribution family such as the Gaussian distribution family. In summary, we have the following model for the expression data

$$\begin{aligned} \theta_k &\sim G_0 \\ \mathbf{z} &\sim \pi \\ g_{ji}|z_{ji}, \theta_k &\sim F(\theta_{z_{ji}}). \end{aligned} \tag{4}$$

Instead of assuming a fixed number of clusters *a priori*, one can assume infinite number of clusters to avoid the estimation accuracy problem on the number of clusters as we mentioned earlier. Correspondingly in (4), the prior π is an infinite discrete distribution. Again as in the Bayesian fashion, we will introduce priors for all parameters. The Dirichlet process is one such prior.

Recall that the Dirichlet distribution $\mathcal{D}(u_1, \ldots, u_K)$ of order K on a $(K-1)$-simplex in $\mathbb{R}^{K-1}$ with parameter $u_1, \ldots, u_K$ can be viewed as a random measure in the finite discrete probability space. Let (X, σ, μ_0) be a probability space. A Dirichlet process $\mathcal{DP}(\alpha_0, \mu_0)$ with parameter $\alpha_0 > 0$ is defined as a random measure: for any non-trivial finite partition $(\chi_1, \ldots, \chi_r)$ of X with $\chi_i \in \sigma$, we have the random variable

$$(\mathcal{DP}(\chi_1), \ldots, \mathcal{DP}(\chi_r)) \sim \mathcal{D}(\alpha_0\mu_0(\chi_1), \ldots, \alpha_0\mu_0(\chi_r)). \tag{5}$$

One remarkable property of the Dirichlet process is that although it is generated by a continuous process, it is discrete (countably many) almost surely [2]. In other words, almost every sample distribution drawn from the Dirichlet process is a discrete distribution. As a consequence, the Dirichlet process is suitable to serve as a non-parametric prior of the infinite mixture model.

The Dirichlet mixture model uses the Dirichlet process as a prior. The model in (4) can then be represented as follows:

$$g_{ji}|z_{ji}, \theta_k \sim F(\theta_{z_{ji}}); \tag{6}$$

θ_k is generated by the measure μ_0

$$\theta_k \sim \mu_0; \tag{7}$$

$\{z_{ji}\}$ is generated by a Dirichlet process $D(\alpha_0, \mu_0)$

$$\{z_{ji}\} \sim D(\alpha_0, \mu_0). \tag{8}$$

Recall that $D(\alpha_0, \mu_0)$ is discrete almost everywhere, which corresponds to the indices of the clusters.

As we mentioned before, biological data such as the expression data often exhibit hierarchical structures. Recall that in the statistical model (8), the clustering effect is induced by the Dirichlet process $D(\alpha_0, \mu_0)$. If we need to take into account different level of clusters, it is natural to introduce a prior with clustering effect to the base measure μ_0. Again in this case, the Dirichlet process can serve as such prior. We simply set the prior to the base measure μ_0 as

$$\mu_0 \sim D_1(\alpha_1, \mu_1), \tag{9}$$

where $D_1(\alpha_1, \mu_1)$ is another Dirichlet process. In this paper, we use the same letter for the measure, the distribution it induces and the corresponding density function as long as it is clear from the context.

In summary, we have the following hierarchical Dirichlet process model for the data:

$$\begin{aligned} \mu_0 &\sim D_1(\alpha_1, \mu_1) \\ \{z_{ji}\}|\mu_o, \alpha_0 &\sim D(\alpha_0, \mu_0) \\ \alpha_0, \alpha_1 &\sim \Gamma(a, b) \\ \theta_k &\sim \mu_1 \\ g_{ji}|z_{ji}, \theta_k &\sim F(\theta_{z_{ji}}), \end{aligned} \tag{10}$$

where $\Gamma(a, b)$ is the Gamma distribution with fixed parameters a, b. We assume that F and μ_1 are conjugate priors. In this paper, F is assumed to be the Gaussian distribution and μ_1 is the inverse Gamma distribution.

The aim for clustering is to determine the posterior probability of the latent membership variables given the observed gene expressions

$$P(\mathbf{z}|\mathbf{g}), \tag{11}$$

where $\mathbf{g} = \{g_{ji}\}_{j,i}$. Once we have the inference result in (11), we can apply the maximum *a posterior* (MAP) criterion to obtain an estimate of membership variable $\hat{z}_{\cdot i}$ for the i-th gene as

$$\hat{z}_{\cdot i} = \arg_a \max \sum_j P(z_{ji} = a|\mathbf{g}). \tag{12}$$

3. INFERENCE ALGORITHM

In this section, we develop a Gibbs sampling algorithm to solve the inference problem (11) by employing the Chinese restaurant metaphor [4], which is a visualized characterization for interpreting the Dirichlet process. We refer to [4] for the proof and other details of the equivalence between the Chinese restaurant metaphor and the Dirichlet processes.

In the Chinese restaurant metaphor for the HDP model (10), we view $\{z_{ji}\}$ as customers entering a restaurant sequentially. The restaurant has infinite number of rows and columns of tables which are labeled by t_{ji}. Each z_{ji} will associate to one and only one table in the j-th row. We use $\phi(z_{ji})$ to denote the column index of the table in the j-th row taken by z_{ji}, i.e., z_{ji} will sit at table $t_{j\phi(z_{ji})}$. If it is clear from the context, we will use ϕ_{ji} in short for $\phi(z_{ji})$. The index of the random variable θ_k in (10) is characterized by a menu containing various dishes. Each table picks one and only one dish from the menus $\{m_k\}_{k=1,2,\dots}$, which are drawn independently from the base measure μ_1. g_{ji} is drawn independently according to the dish it chooses through the distribution $F(\cdot)$ as in (10). We denote $\lambda(t_{ji})$ as the index of the dish taken by table t_{ji}, i.e., table t_{ji} chooses dish $m_{\lambda(t_{ji})}$. As before, we may write λ_{ji} in short of $\lambda(t_{ji})$. The HDP is reflected in this metaphor such that the customers choose the tables as well as the dishes in a Dirichlet process fashion. The customers sitting at the same table are classified into one cluster. Moreover, the customers sitting at different tables but ordering the same dish will also be clustered into the same group. Hence the clustering effect is performed at the cluster level, i.e., we allow "clustering among clusters". We also introduce two useful counter variables: c_{ji} denotes the number of customers sitting at table t_{ji}; d_{jk} counts the number of tables in row j serving dish m_k.

Using the Chinese restaurant metaphor, instead of inferring z_{ji}, we can directly infer ϕ_{ji} and λ_{ji}. We will sample $\boldsymbol{\phi} = \{\phi_{11}, \phi_{12}, \dots\}$ and $\boldsymbol{\lambda} = \{\lambda_{11}, \lambda_{12}, \dots\}$ from the posterior distribution $P(\boldsymbol{\phi}, \boldsymbol{\lambda}|\mathbf{g})$. We can calculate the related conditional probabilities as follows.

If a is a value that has been taken before, the conditional probability of $\phi_{ji} = a$ is given by

$$P(\phi_{ji} = a|\boldsymbol{\phi}^c_{ji}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \alpha_1, \alpha_0, \mu_1, \mathbf{g}) \propto c_{ja} f_{\lambda_{ja}}(g_{ji}|\mathbf{g}^c_{ji}), \tag{13}$$

where $\boldsymbol{\theta} = \{\theta_{ji}\}_{j,i}$ and $\boldsymbol{\lambda} = \{\lambda_{ji}\}_{j,i}$. The superscript c denotes the complement of the variables in its category, i.e., $\mathbf{g}^c_{ji} = \{g_{j'i'}\}_{(j',i')\neq(j,i)}$ and $\boldsymbol{\phi}^c_{ji} = \{\phi_{j'i'}\}_{(j',i')\neq(j,i)}$. $f_{\lambda_{ja}}(g_{ji}|\mathbf{g}^c_{ji})$ denotes the conditional density of g_{ji} given all other data generated according to menu $m_{\lambda_{ja}}$, which can be calculated as

$$f_{\lambda_{ja}}(g_{ji}|\mathbf{g}^c_{ji}) = \frac{\int \prod_{\lambda_{j'\phi_{j'i'}}=\lambda_{ja}} F(g_{j'i'}|\theta)\mu_1(\theta)d\theta}{\int \prod_{j'i'\neq ji, \lambda_{j'\phi_{j'i'}}=\lambda_{ja}} F(g_{j'i'}|\theta)\mu_1(\theta)d\theta}. \tag{14}$$

On the other hand, if a is a new value then we have

$$P(\phi_{ji} = a|\boldsymbol{\phi}^c_{ji}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{g}) \propto \alpha_0 [\sum_{k=1}^{K_{ja}} \frac{\sum_j d_{jk}}{\sum_{jk} d_{jk} + \alpha_1} f_k(g_{ji}|\mathbf{g}^c_{ji}) + \frac{\alpha_1}{\sum_{jk} d_{jk} + \alpha_1} \int F(g_{ji}|\theta)\mu_1(\theta)d\theta]. \tag{15}$$

We also have the following conditional probabilities for λ_{ji}. If a is used before, we have

$$P(\lambda_{j\phi_{ji}} = a|\boldsymbol{\phi}, \boldsymbol{\lambda}^c_{j\phi_{ji}}, \boldsymbol{\theta}, \alpha_1, \alpha_0, \mathbf{g}) \propto (\sum_j d_{ja}) f_a(g_{ji}|\mathbf{g}^c_{ji}); \tag{16}$$

otherwise we have

$$P(\lambda_{j\phi_{ji}} = a|\boldsymbol{\phi}, \boldsymbol{\lambda}^c_{j\phi_{ji}}, \boldsymbol{\theta}, \alpha_1, \alpha_0, \mathbf{g}) \propto \alpha_1 \int F(g_{ji}|\theta)\mu_1(\theta)d\theta. \tag{17}$$

We now summarize the Gibbs sampling algorithm for the HDP inference as follows.

- Initialization with randomly assignments of all the variables.
- For $l = 1, 2, \dots, l_0 + L$,
  - Draw samples of $\{\phi^{(l)}_{ji}\}$ from their posteriors
    $$P(\phi^{(l)}_{ji} = a|\boldsymbol{\phi}^{(l-1)c}_{ji}, \boldsymbol{\lambda}^{(l-1)}, \alpha_1^{(l-1)}, \alpha_0^{(l-1)}, \mathbf{g}) \tag{18}$$
    given by (13) and (15) using the Metropolis-Hastings (M-H) algorithm [5].
  - Draw samples of $\{\lambda^{(l)}_{j\phi^{(l)}_{ji}}\}$ from their posteriors
    $$P(\lambda^{(l)}_{j\phi^{(l)}_{ji}} = a|\boldsymbol{\phi}^{(l)}, \boldsymbol{\lambda}^{(l-1)c}_{j\phi^{(l)}_{ji}}, \alpha_1^{(l-1)}, \alpha_0^{(l-1)}, \mathbf{g}) \tag{19}$$
    given by (16) and (17) using M-H algorithm.
  - Since $P(\alpha_0|\boldsymbol{\phi}, \boldsymbol{\lambda}, \alpha_1, \mathbf{g}) = P(\alpha_0)$ and $P(\alpha_1|\boldsymbol{\phi}, \boldsymbol{\lambda}, \alpha_0, \mathbf{g}) = P(\alpha_1)$, simply draw samples of $\alpha_0^{(l)}$ and $\alpha_1^{(l)}$ from their prior Gamma distributions.
- Using the samples after the "burn-in" period $\{\boldsymbol{\phi}^{(l)}, \boldsymbol{\lambda}^{(l)}\}_{l=l_0+1}^{l_0+L}$ to calculate $\hat{P}(\boldsymbol{\phi}, \boldsymbol{\lambda}|\mathbf{g})$, which is given by
  $$\hat{P}(\phi_{ji} = a, \lambda_{j\phi_{ji}} = b) = \frac{\sum_{l=l_0+1}^{l_0+L} \mathbf{1}\{\phi^{(l)}_{ji} = a, \lambda^{(l)}_{j\phi^{(l)}_{ji}} = b\}}{L}, \tag{20}$$
  where $\mathbf{1}(\cdot)$ is the indicator function. Determine the membership distribution $P(\mathbf{z}|\mathbf{g})$ from the inferred joint distribution $\hat{P}(\boldsymbol{\phi}, \boldsymbol{\lambda}|\mathbf{g})$ by $P(z_{ji} = a|\mathbf{g}) = \sum_b \hat{P}(\lambda_{jb} = a|\mathbf{g}, \phi_{ji} = b)\hat{P}(\phi_{ji} = b|\mathbf{g})$.
- Calculate the estimation of clustering index $\hat{z}_{\cdot i}$ for the i-th gene by $\hat{z}_{\cdot i} = \arg_a \max \sum_j P(z_{ji} = a|\mathbf{g})$.

4. EXPERIMENTAL RESULTS

In this section, we conduct experiment on the yeast galactose datasets [6] and compare the HDP algorithm to the popular MCLUST and SVM algorithms. We also perform experiment on the yeast cell cycle data [7] and compare our result to that in [1].

In order to compare the clustering performance of the HDP algorithm to other methods, we use the Rand index (RI) as the performance metric [8]. The RI is a measure of agreement between two clustering results. It takes a value between 0 and 1. The higher is the score, the higher agreements it indicates.

We conduct experiment on the yeast galactose data, which consists of 205 genes. The true number of clusters based on the functional categories is 4. We calculate the RI index between different clustering results to the result in [9], which is regarded as the standard benchmark. The performances of the algorithms under consideration are listed in Table 1.

Algorithm	Rand Index	Number of clusters
SVM	0.954	5
MCLUST	0.903	9
HDP	0.973	3.8

Table 1. Clustering performance of HDP, MCLUST and SVM on the yeast galactose data.

It is seen that the HDP algorithm performs the best among the three algorithms. Unlike the MCLUST algorithm which produces far more clusters than 4, the average number of clusters given by the HDP algorithm is very closed to the "true" value 4.

We next apply the proposed HDP clustering algorithm on the Yeast cell cycle dataset, which has been used widely for testing the performances of clustering algorithm. We resort to the MIPS database [10] to determine the functional categories for each cluster. After applying the cell-cycle selection criterion in [1], we find that there are 126 genes identified by proposed HDP algorithm but not discovered in [1]. We list in Table 2 the numbers of newly discovered genes in various functional categories.

Note that in [11] a Bayesian model with infinite number of clusters is proposed based on the Dirichlet process. The model in [11] is a special case of the HDP model proposed in this paper when there is only one hierarchy. In terms of discovering new gene functionalities, we find that the performances of the two algorithms are similar, as the method in [11] discovered 106 new genes compared to the result in [1]. However, by taking the hierarchical structure into account, the total number of clusters found by the HDP algorithm is significantly smaller than that given in [11] which is 43 clusters. The HDP clustering consolidates many fragmental clusters, which may provide an easier way to interpret the clustering results.

Function categories	Number of new genes
Cell cycle and DNA processing	20
Protein synthesis	25
Protein fate	4
Cell fate	12
Transcription	8
Unclassified protein	57

Table 2. Numbers of newly discovered genes in various functional categories by the proposed HDP clustering algorithm.

5. REFERENCES

[1] R.J. Cho et al, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mole. Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[2] T.S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[3] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[4] D. Aldous, "Exchangeability and related topics," *École d'Été de Probabilités de Saint-Flour XIII*, pp. 1–198, 1985.

[5] S. Brooks, "Markov chain Monte Carlo method and its application," *J. of the Royal Stat. Soc.*, vol. 47, no. 1, pp. 69–100, 1998.

[6] K.Y. Yeung, M. Medvedovic, and R.E. Bumgarner, "Clustering gene-expression data with repeated measurements," *Genome Biology*, vol. 4, no. 5, pp. R34, 2003.

[7] P.T. Spellman et al, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Mole. Bio. of the Cell*, vol. 9, no. 12, pp. 3273, 1998.

[8] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[9] M. Ashburner et al, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[10] H.W. Mewes et al, "MIPS: a database for genomes and protein sequences," *Nucleic Acids Research*, vol. 30, no. 1, pp. 31–34, 2002.

[11] M. Medvedovic and S. Sivaganesan, "Bayesian infinite mixture model based clustering of gene expression profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1194–1206, 2002.