A UNIFIED PROBABILISTIC FRAMEWORK FOR CLUSTERING GENES FROM GENE EXPRESSION AND PROTEIN-PROTEIN INTERACTION DATA

Xiaofeng Dai[1] *and Harri Lähdesmäki*[1,2]

[1]Department of Signal Processing,
Tampere University of Technology, Tampere, Finland
[2]Department of Information and Computer Science,
Helsinki University of Technology, Helsinki, Finland
{xiaofeng.dai,harri.lahdesmaki}@tut.fi

ABSTRACT

This paper presents a novel mixture model for clustering genes based on Gaussian and bernoulli distributed data. One typical application is to cluster genes with gene expression and protein-protein interaction (PPI) data. The underlying assumption is that genes within a cluster have on average more PPIs with a set of genes and share similar expression profiles than genes from different clusters. The proposed mixture model, GBMM, differs from its component models in its integration of different data types into a single and unified probabilistic modeling framework. Moreover, the model can be extended to other parametric distributions and, therefore, incorporate even more information in a coherent manner. We developed the expectation maximization algorithm for GBMM, and used four well-known approximation-based model selection criteria to test their performances under different scenarios. The results verify that combining expression and PPI data can greatly improve clustering accuracy compared with analyzing each single data source alone, and the more PPIs are known for a given set of genes the better performance improvement the algorithm can have.

1. INTRODUCTION

The most commonly used information in gene clustering is expression data. However, the assumption that genes in the same functional group share similar expression patterns is often violated by the varied transcriptional coherence (in response to diverse environmental stresses) and random noise contamination [1]. Therefore, the accuracy of methods relying only on gene expression data has largely been restricted by the over-dependence nature on the measured expression values of individual genes, especially when dealing with correlated genes whose expression similarity is low. Another data type that can deliver information on functional gene relationship is protein-protein interactions (PPIs). It is reported that $70 - 80\%$ of interacting protein pairs share at least one function [2]. However, it is easy to gain over-confidence on intra-functional PPIs, while neglecting inter-functional interactions. Therefore, it is necessary to incorporate information about every functional interactions to improve the accuracy of computational prediction.

Given the particular imperfection of using each single data source, gene expression and PPI data are often coupled together for different applications, such as identifying molecular pathways [3] and inferring gene functions [1]. It has been demonstrated that protein interactions are reflected in gene expression, and their relationship has been demonstrated in bacteriophage T7 and yeast- [4]. In this paper, we present a unified probabilistic framework, Gaussian-bernoulli mixture model (GBMM), for fusing data of Gaussian and bernoulli distributions. One typical application is to cluster genes using expression and PPI data with the assumption that one gene corresponds to one protein, and genes within a cluster have on average more PPIs with a set of genes and share similar expression profiles than genes from different clusters (e.g., genes that are clustered together interact with each other). This method differs from other methods in its extreme flexibility and broad applicability. Not only expression and PPI data can be modeled by this method, any data that follows Gaussian distribution (e.g., various other microarray-based measurements, and bernoulli distribution (such as promoter binary binding data, or literature-derived interactions) can be fitted into this framework. Additional information can be easily incorporated by adding more component models into this framework. Moreover, the proposed algorithm is a mixture model based method, whose probabilistic nature guarantees an efficient utilization of PPI data, avoiding the 'all-or-none' attitude that has been implicitly adopted by many people.

An expectation maximization (EM) algorithm to jointly estimate parameters of Gaussian and bernoulli distributions is developed for GBMM, which is then tested by comparing with its two component models, Gaussian mixture model (GMM) and bernoulli mixture model (BMM). Four well-known model selection criteria, Bayesian information criterion (BIC), integrated classification likelihood-BIC (ICL-BIC, called ICL for simplicity), Akaike information criterion (AIC), and modified AIC (AIC3) were tested in GBMM and its two component models in this study. Performance tests were done on GBMM with both

complete and incomplete PPI matrixes, each representing a different scenario. Complete PPI matrix is a square symmetric matrix, where information about interactions among all the interactors is available; while in incomplete PPI data only partial interactions among the interested proteins (protein products of genes that are need to be clustered) are available, and may or may not include interactions of the interested proteins with other molecules. Performances were also compared on dealing with data containing different noise levels. Besides verifying that combining gene expression and PPI data can highly improve the clustering accuracy compared with analyzing either of the single data sources alone, our results also show that the more PPIs are available the better performance the algorithm has. According to our study, AIC and AIC3 perform similar in GBMM, and work better than the other two criteria when bernoulli distributed data is too noisy.

2. METHODS

2.1. Mixture model based clustering

In model-based clustering methods, each observation $\mathbf{o}_j$, where $j = 1, \ldots, n$ and n is the number of genes, is drawn from a finite mixture distribution with the prior probability π_i, component-specific distribution $f_i^{(g)}$ and its parameters θ_i. The formula is given as [5]

$$f(\mathbf{o}_j|\theta) = \sum_{i=1}^{g} \pi_i f_i^{(g)}(\mathbf{o}_j|\theta_i), \tag{1}$$

where $\theta = \{(\pi_i, \theta_i) : i = 1, \ldots, g\}$ is used to denote all the unknown parameters, with the restriction that $0 < \pi_i \leq 1$ for any i and $\sum_{i=1}^{g} \pi_i = 1$. Note that g is the number of components in this model. In the following texts, superscript (g) is ignored for simplicity.

2.1.1. GBMM

Define $\theta = [\pi, \theta_1, \theta_2]^T, \theta_1 = \left[\mu_{11}, \ldots, \mu_{gp_1}, \sigma_1^2, \ldots, \sigma_{p_1}^2\right]^T$, $\theta_2 = [q_{11}, \ldots, q_{gp_2}]^T$, and $\pi = [\pi_1, \ldots, \pi_g]^T$, where p_1 and p_2 each represents the dimension of the observations in Gaussian and bernoulli mixture model, respectively. Denote X and Y as the Gaussian and bernoulli distributed random variables, respectively, function f of $\mathbf{x}$, $\mathbf{y}$ as their density function, and $\mathbf{o} = [\mathbf{x}^T, \mathbf{y}^T]^T$.

GBMM is a joint mixture model of Gaussian and bernoulli distributions, with the assumption that, for each component i, data of both distributions are independent, i.e., $f_i(\mathbf{o}) = g_i(\mathbf{x})h_i(\mathbf{y})$. In the GMM part, each component is assumed to be the product of p_1 independent Gaussian distributions, whose probability density function is defined as

$$g_i(\mathbf{x}|\theta_{1i}) = \frac{1}{(2\pi)^{\frac{p_1}{2}}|V|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T V^{-1}(\mathbf{x}-\mu_i)\right), \tag{2}$$

where $\theta_{1i} = [\mu_i, V], \mu_i = [\mu_{i1}, \ldots \mu_{ip_1}], V = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_{p_1}^2)$ and $|V| = \prod_{u=1}^{p_1} \sigma_u^2$. In the BMM part, each component is modeled as bernoulli distribution, with the probability density function for each gene defined as

$$h_i(\mathbf{y}|\theta_{2i}) = \prod_{v=1}^{p_2} q_{iv}^{\mathbf{y}_v}(1-q_{iv})^{(1-\mathbf{y}_v)}, \tag{3}$$

where $\theta_{2i} = [q_{i1}, \ldots, q_{ip_2}]$. Note that in complete PPI matrix $p_2 = n$, and in incomplete case often $p_2 < n$.

We assume diagonal covariance matrix in GMM since it significantly reduces the number of parameters and thus the complexity, which is useful in dealing with high-dimensional data.

2.2. EM algorithms

Standard EM algorithm is applied to estimate the parameters θ iteratively, where the data log-likelihood (natural logarithm) can be written as

$$\log L(\theta) = \sum_{j=1}^{n} \log\Big(\Big[\sum_{i=1}^{g} \pi_i f_i(\mathbf{o}_j|\theta_i)\Big]\Big), \tag{4}$$

given $O = \{\mathbf{o}_j : j = 1, ..., n\}$, whose direct maximization, however, is difficult. The problem is thus solved by maximizing the complete data log-likelihood as shown in Equation 5

$$\log L_c(\theta) = \sum_{j=1}^{n}\sum_{i=1}^{g} \chi(c_j = i) \log\left(\pi_i f_i(\mathbf{o}_j|\theta_i)\right), \tag{5}$$

where $c_j \in \{1, \ldots, g\}$ is the clustering membership of $\mathbf{o}_j$, and $\chi(c_j = i)$ is the indicator function of whether o_j is from the ith component or not.

In the EM algorithm, E step computes the expectation of the complete data log-likelihood, $Q(\theta|\theta^{(m)})$, as

$$Q(\theta|\theta^{(m)}) = \sum_{j=1}^{n}\sum_{i=1}^{g} \tau_{ji}^{(m)} \log(\pi_i f_i(\mathbf{x}_j|\theta_{1i}) f_i(\mathbf{y}_j|\theta_{2i})), \tag{6}$$

where

$$\tau_{ji}^{(m)} = \frac{\pi_i^{(m)} f_i(\mathbf{x}_j|\theta_{1i}^{(m)}) f_i(\mathbf{y}_j|\theta_{2i}^{(m)})}{\sum_{i'=1}^{g} \pi_{i'}^{(m)} f_{i'}(\mathbf{x}_j|\theta_{1i'}^{(m)}) f_{i'}(\mathbf{y}_j|\theta_{2i'}^{(m)})}, \tag{7}$$

according to Bayes' rule and $\theta^{(m)}$ represents the parameters estimated in the m^{th} iteration. Note that $\tau_{ji}^{(m)}$ is the estimated posterior probability of $\mathbf{o}_j$ coming from component i at iteration m, and we can assign each $\mathbf{o}_j$ to a component based on $\{i_0|\tau_{ji_0} = \max_i \tau_{ji}\}$. Equation 6 shows that our assumption of the Gaussian and bernoulli distributed data being independent carries over to the expected log-likelihood as well.

In the EM algorithm of GBMM, the parameters of the Gaussian part, μ_i's and σ_u^2's, can be estimated by the standard EM algorithm of GMM with diagonal covariance matrix, which works by iterating over

$$\hat{\mu}_i^{(m+1)} = \sum_{j=1}^{n} \tau_{ji}^{(m)} \mathbf{x}_j / \sum_{j=1}^{n} \tau_{ji}^{(m)}, \tag{8}$$

$$\hat{\sigma}_u^{2,(m+1)} = \sum_{j=1}^{n}\sum_{i=1}^{g} \tau_{ji}^{(m)} (x_{ju} - \mu_{iu}^{(m)})^2 / n. \tag{9}$$

The parameters of bernoulli part, q_{iw}'s, are updated by

$$\hat{\mathbf{q}}_i^{(m+1)} = \frac{\sum_{j=1}^{n} \tau_{ji}^{(m)} \mathbf{y}_j}{\sum_{j=1}^{n} \tau_{ji}^{(m)}}$$

and π's are updated by

$$\hat{\pi}_i^{(m+1)} = \sum_{j=1}^{n} \tau_{ji}^{(m)} / n, \tag{10}$$

where $\tau_{ji}^{(m)}$ is calculated from Equation 7. Note that $\{u = 1, \ldots, p_1\}$ and $\{v = 1, \ldots, p_2\}$.

From the above equations, it is easy to see that the standard EM for GBMM will reduce to the standard EM for Gaussian and bernoulli distribution, respectively, when the dimensions of the other two distributions go to zero.

The algorithm is run multiple times with different initial values. The initial parameters μ_{iu}'s, σ_u^2's and q_{iv}'s are obtained from the randomly initialized fuzzy c-means clustering results, and π_i's are initialized with the uniform probability $1/g$. For each data set, we run each EM algorithm 100 times with different initial values, and for all the tested models, we set the convergence threshold ($|Q|$ is used to monitor the convergence) and the maximum number of iterations to 0.0001 and 100, respectively. All the simulations have reached convergence according to the statistics stored during the simulations.

2.3. Model selection

Four well-known approximation-based model selection criteria, BIC [6], ICL [7], AIC [8, 9], and AIC3 [8, 10] are compared in GBMM, according to which the best-performing criterion is chosen. Calculations for the above criteria can be found in [11]. Note that the number of free parameters in GBMM is $p_1 + p_1 g + p_2 g + g - 1$.

3. RESULTS

The data sets designed for the performance test are shown in Table 1, where each data type falls into two categories with respect to different noise levels. In Gaussian distributed data, the noise is determined by the mean and variance, and in bernoulli distributed data, it is defined as the ratio between the number of false interactions and true interactions. Specifically, the designed data are good Gaussian ('gG'), bad Gaussian ('bG'), good bernoulli ('gB') and bad bernoulli ('bB'). For noisy Gaussian distributed data, we also consider different noise sources, which are close means and large variances (denoted as 'bG$_m$' and 'bG$_v$', respectively), and for bernoulli distributed data, different data structures are also taken into account for both good and bad data, which are complete ($p_2 = n$ and symmetric) PPI matrixes (denoted as 'gB$_p$' and 'bB$_p$'), and incomplete (normally $p_2 < n$) PPI data (denoted as 'gB$_b$' and 'bB$_b$'). In our design, $n = 100$, $p_1 = 4$, and $p_2 = 6$ for B$_b$. All the simulations are repeated 10 times with randomly generated data sets. The designed expression data are listed in Table 1, and the sparsity patterns of two kinds of PPI matrixes are shown in Fig. 1.

cluster 1				cluster 2				cluster 3			
5	-8	20	15	10	1	-20	0	-10	8	5	15
1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
3	15	5	11	2	13	6	9	1	14	7	10
1	2	3	2.5	1	2	3	2.5	1	2	3	2.5
5	-8	20	15	10	1	-20	0	-10	8	5	15
10	20	30	25	10	20	30	25	10	20	30	25

Note: '||' and '|' separate the clusters and the dimensions within each cluster, respectively. The three 2×12 boxes downwards list the parameters of scenarios 'gG', 'bG$_m$' and 'bG$_v$', respectively. In each box the 1^{st} line shows μ and the 2^{nd} line shows σ.

Table 1. Parameters of expression data.

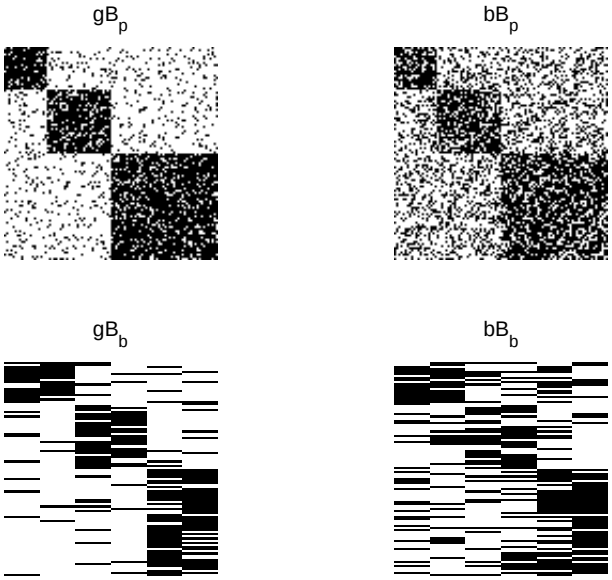


Figure 1. Sparsity patterns of PPI data. The noise level is 1/9 in 'gB$_p$' and 'gB$_b$', and is 1/2 in 'bB$_p$' and 'bB$_b$'.

We employed the same scoring system as developed in [11], denoted as 'E score', for performance evaluation in simulations, which not only records the clustering accuracy but also reflects the influence of the criterion for model selection.

We first compared different model selection criteria under different scenarios, whose results are shown in Table 2. It is seen that the choice of model selection criterion is case dependent, and generally, AIC and AIC3 work slightly better than the other two criteria in GBMM.

The performance of GBMM was compared with GMM and BMM for all the possible combinations of our test data sets, whose results are shown in Fig. 2. It is seen that GBMM outweighs both of its component models under all the tested scenarios, which demonstrates the power of utilizing multiple data sources in gene clustering. Moreover, in BMM and GBMM, the larger the second dimension (given a fixed noise level) of PPI matrix is, the more PPI information are utilized, and the more accurate the clustering is.

Model	Scenario	AIC	AIC3	BIC	ICL
GBMM	gG+gB$_p$	**1**	**1**	**1**	**1**
	gG+bB$_p$	**1**	**1**	**1**	**1**
	bG$_m$+gB$_p$	**0.999**	**0.999**	**0.999**	**0.999**
	bG$_m$+bB$_p$	**0.975**	**0.975**	0.787	0.787
	bG$_v$+gB$_p$	**1**	**1**	**1**	**1**
	bG$_v$+bB$_p$	**0.965**	**0.965**	0.803	0.803
	gG+gB$_b$	0.9990	**1**	**1**	**1**
	gG+bB$_b$	**1**	**1**	**1**	**1**
	bG$_m$+gB$_b$	**0.963**	**0.963**	0.942	0.942
	bG$_m$+bB$_b$	0.645	**0.648**	0.600	0.600
	bG$_v$+gB$_b$	**0.957**	**0.957**	0.947	0.947
	bG$_v$+bB$_b$	0.693	**0.774**	0.727	0.727
GMM	gG	**1**	**1**	**1**	**1**
	bG$_m$	0.508	**0.512**	0.506	0.506
	bG$_v$	**0.597**	0.577	0.581	0.581
BMM	gB$_p$	**0.998**	**0.998**	**0.998**	**0.998**
	bB$_p$	**0.814**	**0.814**	0.800	0.800
	gB$_b$	**0.922**	**0.922**	0.906	0.906
	bB$_b$	**0.604**	**0.604**	**0.604**	**0.604**
Note: The best criterion with respect to the highest average E scores and used in drawing Fig. 2 are shown in bold face. All values are rounded to three decimal points.					

Table 2. Comparison of different model selection criteria in GBMM, GMM and BMM.

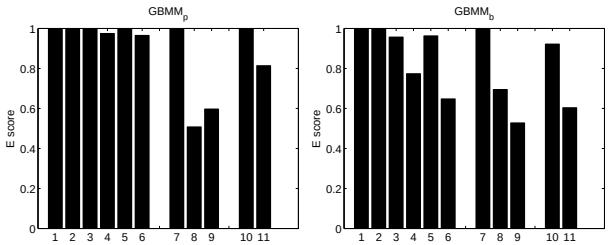
4. CONCLUSIONS

This paper presents a novel Gaussian-bernoulli mixture model, GBMM, for gene clustering from Gaussian distributed and bernoulli distributed data. One typical application is to cluster genes from expression and PPI data, assuming that one gene corresponds to one protein and genes within a cluster have on average more PPIs with a set of genes and share similar expression profiles than genes from different clusters. The results verify that combing expression and PPI data can make more efficient use of data than analyzing each single data source, and the larger the second dimension of PPI matrix is the more accurate the results are, given a fixed number of genes.

The main contribution of this paper is that we have presented an extremely flexible clustering framework which can deal with any data that follow Gaussian and bernoulli distributions. For example, besides the typical square symmetric PPI matrix, the proposed method can also be applied to incomplete bernoulli distributed data, such as binary promoter binding states. Moreover, it can be easily extended to utilize other information source by extending the current framework to other parametric distribution.

The clustering accuracy of GBMM is shown to be highly improved under all the tested scenarios compared with its component models. However, the algorithm can be slow when dealing with large dimensional unstructured PPI matrix.

In the future, we could extend this joint clustering framework into data of other parametric distributions, and apply it to solve other problems.



(a) (b)

Figure 2. Comparison of GBMM with GMM and BMM when PPI data is (a) complete, and (b) incomplete. $1 \sim 6$ each represents 'gG+gB', 'gG+bB', 'bG$_m$+gB', 'bG$_m$+bB', 'bG$_v$+gB', 'bG$_v$+bB' in GBMM; $7 \sim 9$ each stands for 'gG', 'bG$_m$' and 'bG$_v$' in GMM; $10 \sim 11$ represent 'gB' and 'bB', respectively.

5. REFERENCES

[1] K. Tu, H. Yu, and Y. X. Li, "Combing gene expression profiles and protein-protein interaction data to infer gene functions," *Journal of Biotechnology*, vol. 124, pp. 475–485, 2006.

[2] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global protein function prediction from protein-protein interaction networks," *Nature Biotechnology*, vol. 21, pp. 697–700, 2003.

[3] E. Segal, H. Wang, and D. Koller, "Discovering molecular pathways from protein interaction and gene expression data," *Bioinformatics*, vol. 19, pp. i264–i272, 2003.

[4] A. Grigoriev, "A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage t7 and the yeast *Saccharomyces cerevisiae*," *Nucleic Acids Research*, vol. 29, no. 17, pp. 3513–3519, 2001.

[5] G. Mclachlan and D. Peel, *Finite mixture models*, John Wiley & Sons, New York, 2000.

[6] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[7] Y. Ji, C. Wu, P. Liu, J. Wang, and R. K. Coombes, "Applications of beta-mixture models in bioinformatics," *Bioinformatics*, vol. 21, no. 9, pp. 2118–2122, 2005.

[8] C. Biernacki and G. Govaert, "Choosing models in model-based clustering and discriminant analysis," *Journal of Statistical Computation and Simularion*, vol. 64, no. 1, pp. 49–71, 1999.

[9] H. Akaike, "A new look at the statistical identification model," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[10] H. Bozdogan, "Model selection and akaike information criterion (aic): The general theory and its analytic extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[11] X. F. Dai, H. Lähdesmäki, and O. Yli-Harja, "BGMM: a Beta-Gaussian mixture model for clustering genes with multiple data sources," in *Fifth International Workshop on Computational Systems Biology*. Finland: Tampere University of Technology, 2008, pp. 25–28.