基于网络约束双聚类的癌症亚型分类

王 星[1] 王 峻[1] 余国先[1] 郭茂祖[2),3)]

[1)](西南大学计算机与信息科学学院 重庆 400715)

[2)](北京建筑大学电气与信息工程学院 北京 100044)

[3)](建筑大数据智能处理方法研究北京市重点实验室 北京 100044)

摘 要 癌症亚型识别在肿瘤异质性分析中具有重要意义.双聚类可以在大规模基因表达数据的基因和样本维度上同时进行聚类分析,发现部分样本在部分基因子集上表达相似的双聚类簇,进而发现相应的癌症亚型,为癌症的精准基因治疗等提供了重要的信息.双聚类算法通过结合基因相互作用网络数据,可进一步提高癌症亚型分类的准确度,但已有整合基因网络的双聚类算法通常仅基于基因的度加权选择基因,易受网络中噪声互作的干扰和缺失互作的误导.为此,该文提出了一种基于基因互作网络正则化的双聚类算法(Network Regularized Bi-Clustering algorithm,NetRBC).NetRBC 首先通过最小化聚类簇上的均方残差分别求取癌症基因表达数据矩阵上的基因簇和样本簇指示矩阵;然后利用基因网络和基因簇指示矩阵构建图正则项;最后将此正则项结合到基于均方残差的非负矩阵分解中,约束基因簇和样本簇矩阵的协同分解,以期提高癌症亚型分类的精度.在多个癌症基因表达数据上的实验结果表明,NetRBC 比已有相关方法能够更准确地区分癌症亚型.

关键词 双聚类;均方残差;非负矩阵分解;癌症亚型;基因网络

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.01274

Network Regularized Bi-Clustering for Cancer Subtype Categorization

WANG Xing[1)] WANG Jun[1)] YU Guo-Xian[1)] GUO Mao-Zu[2),3)]

[1)](*College of Computer and Information Science, Southwest University, Chongqing* 400715)

[2)](*College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing* 100044)

[3)](*Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing* 100044)

Abstract Cancer subtype identification is crucial for understanding tumor heterogeneity. Existing methods for identifying cancer subtypes have primarily focused on utilizing traditional clustering algorithms (such as k-means and hierarchical clustering) to cluster gene expression data and thus to identify subtypes. These traditional approaches, however, separately group the data from genes or samples dimension only, so they cannot discover the patterns that similar genes exhibit similar behaviors only over a subset of conditions (or samples). Bi-clustering can simultaneously group large scale gene expression data from sample and gene dimensions, and find out bi-clusters that relevant samples exhibit similar gene expression profiles over a subset of genes, and thus to identify corresponding cancer subtypes. The discovered bi-clusters bring insights for categorizing cancer subtypes and precise gene treatments. Incorporating the information of gene-gene interaction networks can further improve the quality of the discovered bi-clusters. However, current efforts generally use the networks to weight and select genes. They are often interfered by noisy interactions

收稿日期:2018-03-16;在线出版日期:2018-10-08.本课题得到国家自然科学基金(61873214,61872300,61741217,61871020,61571163,61532014)、重庆市基础与前沿研究项目(cstc2018jcyjAX0228,cstc2016jcyjA0351)资助.**王 星**,硕士研究生,中国计算机学会(CCF)学生会员,主要研究方向为机器学习和生物信息学.E-mail:wx1993cs@email.swu.edu.cn.**王 峻**(通信作者),博士,副教授,中国计算机学会(CCF)高级会员,主要研究方向为数据挖掘和生物信息学等.E-mail:kingjun@swu.edu.cn.**余国先**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为机器学习、数据挖掘和生物信息学.**郭茂祖**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为机器学习、数据挖掘和生物信息学.

and misled by missing interactions. There are many types of bi-clusters, including constant bi-cluster, constant row bi-cluster, constant column bi-cluster, coherent values additive bi-cluster and coherent value multiplicative bi-cluster. To address these limitations and explore multiple types of bi-clusters, in this paper, we introduce a gene-gene interaction Network Regularized Bi-Clustering algorithm (NetRBC) based on the Semi-Nonnegative Matrix Tri-Factorization (SNMTF). NetRBC firstly integrates the mean square residuals into SNMFT, and optimizes the gene-cluster and sample-cluster indicator matrices via minimizing the sum-squared loss of the discovered bi-clusters. Next, it constructs a graph regularization term by using the gene networks and gene-cluster indicator matrix. The core idea of the regularization term is that if a pair of genes interact with each other, these genes may co-regulate the production of one cancer subtype, so we except that these genes can be grouped into the same bi-clusters. After that, NetRBC incorporates the regularization term into a sum-squared loss based SNMTF to guide the collaborative factorization and thus to pursue gene-cluster indicator matrix and sample-cluster indicator matrix, and thus to improve the accuracy of cancer subtypes categorization. At the same time, NetRBC uses a regularization parameter to control the contribution of gene-gene interaction network. We also give an optimization technique to optimize the gene-cluster and sample-cluster indicator matrices, which uses the multiplicative updating technique to alternatively optimize one variable, while fixing the other variables, until convergence. We conduct experiments on six cancer gene expression datasets with known subtypes to comparatively study the performance of NetRBC. We further test NetRBC on two large-scale cancer gene expression datasets from The Cancer Genome Atlas (TCGA) project and use the clinical features of patients to evaluate the performance, since the true subtypes of these samples belonging to are unknown. Extensive experimental results show that NetRBC can better group patients into subtypes than competitive comparing methods, and the proposed network regularization term indeed significantly improves the cancer subtype categorization accuracy.

Keywords　bi-clustering; sum-squared residue; nonnegative matrix factorization; cancer subtypes; gene networks

1　引　言

癌症亚型的定义和发现是癌症个性化治疗的基础，它可为癌症病人精准治疗方案的设计提供非常重要的参考[1-2]. 基因组技术的发展和广泛应用使得研究者能够便捷地获取癌症病例全基因组的高通量测序数据，为其在全基因组水平上研究癌症个体间的差异创造了条件. 针对上述癌症基因组数据，研究者利用聚类算法挖掘分析其包含的遗传信息，探索癌症存在的亚型及其相应的肿瘤分子标志物，对癌症研究和治疗具有重要的现实意义[2].

基于高通量测序技术获取的基因表达数据通常可以用矩阵表示，矩阵的每一行表示一个基因，每一列表示一个样本. 矩阵中的元素表示基因在不同样本(实验条件)下的表达值，实际应用到癌症基因表达数据中时，矩阵中的元素表示对应基因在不同病人中的表达值. 基因表达数据矩阵具有基因维度高、噪声大、样本维度低的特点，同时存在不同程度的数据缺失，因此，利用这类数据进行聚类挖掘和分析时面临着诸多困难.

在早期的基因表达数据分析中，研究者使用的聚类算法都是从单一的维度进行处理，单独地对样本进行聚类或者对基因进行聚类，如 k-means[3] 和层次聚类(Hierarchical Clustering, HC)[4]，其结果反映的是基因在全部样本上表达的相关性或样本在全部基因上表达的相关性. 然而在生物体内，参与同一调控功能的仅是一部分基因，且这部分基因只在部分条件下表达. 为此，研究者提出了利用双聚类算法进行基因表达数据分析[5]. 双聚类算法可以同时从基因维度和样本维度挖掘分析基因表达数据，它解决了传统聚类方法只能在基因表达数据的基因或

者样本方向上进行聚类的问题，克服了不能发掘数据中局部信息的缺陷，可以找到在部分基因上具有较高相关性的样本子集，从而实现癌症样本亚型区分的目的.

已有双聚类算法在双聚类过程中，仅仅依靠基因表达数据，存在一定的局限性. 随着研究的深入，人们发现基因互作网络有助于更系统深入地解释某些基因与癌症亚型的相关性，越来越多的研究者开始将基因互作网络融合到基因表达数据的双聚类算法中[6-7]. 然而，这类双聚类算法一般仅基于基因互作网络中基因的度设置基因权重并进行关键基因选择，易受网络噪声的干扰，且未能较好地利用基因网络的拓扑结构信息. 为此，本文首先设计了一个基于最小化簇内均方残差和非负矩阵分解[8]的双聚类目标方程，然后利用基因互作网络构建一个图正则化项并结合到目标方程中指导基因簇和样本簇指示矩阵的协同分解，进而提出一种基于网络约束双聚类(Network Regularized Bi-Clustering，NetRBC)的癌症亚型分类算法. 在多个癌症基因表达数据集上的实验结果表明 NetRBC 比现有相关算法能够更准确地区分癌症亚型.

本文第 2 节将介绍相关的研究工作，包括单聚类算法、双聚类算法和结合基因互作网络的双聚类算法；第 3 节将具体介绍本文提出的基于网络正则化的双聚类算法；第 4 节将进行实验和结果分析；最后在第 5 节总结本文并对未来工作进行展望.

2 相关工作

早期面向基因表达数据的分析算法主要是传统的聚类算法. Phillips 等人[9]利用 k-means 划分基因表达数据中的样本，该算法首先随机初始化 k 个簇中心点，将每个样本分别划分到最近中心点所在簇中，再根据簇中元素更新簇中心点，如此迭代重复上述过程直到簇中心点稳定. 但是该算法随机初始化簇中心点使得聚类结果不稳定，且对噪声敏感，在高维数据上的性能有限. Monti 等人[10]利用一致性聚类(Consensus Clustering，ConC)分析基因表达数据，通过整合多次聚类的结果提高聚类精度. Eisen 等人[11]通过在基因维度上进行层次聚类，试图说明功能相似的基因在表达值上也相似. Shai 等人[12]也利用层次聚类算法在神经胶质瘤基因表达数据上从样本的维度进行聚类，从而找到该肿瘤的亚型. 与 k-means 类似，层次聚类在高维数据上也易受噪声的干扰[13]. Vesanto 等人[14]利用自组织映射网络(Self-organizing Map)对基因表达数据进行聚类分析，该方法首先将高维基因映射到一维或者二维结构中，然后在低维结构中将距离较近的节点构成相邻的类，从而对样本进行聚类. 这些单聚类方法得到的聚类簇反映的是一组基因在全部样本下的表达一致性，或者一组样本在全部基因下的表达一致性. 然而现有研究证明部分样本只在部分基因下具有表达一致性，在基因表达数据中寻找具有局部相似性的簇是发现癌症亚型的关键[15]. 传统单聚类方法无法发现基因表达数据中的局部簇结构，难以准确的区分癌症亚型.

为解决上述问题，研究者提出利用双聚类方法分析基因表达数据. Cheng 和 Church[5]首次在基因表达数据分析中应用双聚类算法(Cheng and Church，CC)，CC 定义了一个均方残差的度量评价双聚类簇的质量，均方残差越小，簇越稳定、越好；再设定一个阈值，随机地删除或者添加矩阵中的行或列寻找均方残差小于设定阈值的簇. 之后，各种类型的双聚类方法陆续提出并被应用到基因表达数据分析中，如基于图理论的双聚类[16-17]、基于谱分析的双聚类[18-19]、基于启发式搜索的双聚类[20-21]和基于矩阵分解的双聚类[22-23]等. Tanay 等人[16]提出的 SAMB(Statistical-Algorithmic Method for Bicluster analysis)是一种基于图理论的双聚类算法. SAMB 将原始的基因表达数据转化为二分图，图中两组节点分别表示基因和样本，节点之间边权重利用一个概率模型分配，再通过计算该图中的最大权重子图发现双聚类簇. 因为癌症病人通常仅患同一癌症的单种亚型，所以 Kluger 等人[18]假设基因表达数据中各个双聚类簇构成了一个棋盘状结构，不存在重叠，在此基础上提出了基于谱分析的 SBC(Spectral Bi-Clustering)算法. SBC 首先从基因和样本维度求解特征值和特征向量来分析两个维度的结构，重构原始数据矩阵，再基于奇异值分解进行双聚类[24]. Cho 等人[20]提出基于最小化均方残差的双聚类算法 MSSRCC(Minimum Sum-Squared Residue Co-Clustering)，MSSRCC 将文献[5]中定义的均方残差设为目标方程，在行和列上通过类似 k-means 的方法迭代求解基因簇指示矩阵和样本簇指示矩阵. Lee 等人[22]基于稀疏的奇异值分解对基因表达数据进行双聚类分析，将原始矩阵通过低秩矩阵近似表示，并基于基因表达数据引入惩罚项约束基因簇和样本簇指示矩阵的稀疏分解. 这类基于矩阵分解的方法与

QUBIC[25]和 SGFA[26]不同，后两种算法可以找到重叠的双聚类簇，但它们并不会将每个基因或样本划分到簇中. 以上方法均仅基于基因表达数据进行分析，没有关注基因之间的互作关系，聚类结果的可解释性不足.

生物分子网络在理解和研究复杂疾病机制中具有非常重要的作用，如基因/蛋白质互作网络直接影响基因表达数据的模式[7]，近年来，越来越多的研究者将生物分子网络信息引入到基于基因表达数据的癌症亚型分类研究中. Hanisch 等人[27]利用酵母菌基因表达数据和新陈代谢调控网络进行聚类研究，他们首先基于新陈代谢网络中基因的相似度和该基因在表达谱中的相似度定义了一种距离度量，并依此进行层次聚类分析. Nguyen 等人[28]提出的 PINS(Perturbation clustering for data Integration and disease Subtyping)，Wang 等人[29]提出的 SNF(Similarity Network Fusion for aggregating data types on a genomic scale)和 Mo 等人[30]提出的 iClusterPlus(Integrative Clustering of multiple genomic data sets)利用多源的基因组数据分别计算样本相似度，然后通过不同方式整合多种相似度矩阵来提升癌症亚型的分类精度，但它们都是单聚类集成方法，没有同时对基因进行聚类分析. Liu 等人[31]提出了一种基于网络辅助的双聚类算法 NCIS(Network-assisted bi-clustering)，该算法首先在基因互作网络上利用 GeneRank[32]算法求得每个基因的权重，再利用三元非负矩阵分解[8]求基因簇和样本簇指示矩阵(指示矩阵中对应的值分别表示每个基因和每个样本属于某个簇的概率，一般选择指示矩阵中概率最大的值作为对应基因和样本所属的簇). Yu 等人[33]基于 MSSRCC 的目标方程优化多个簇上的均方残差和 NCIS 中基于网络的基因加权方式，提出了一种用于发现癌症亚型的双聚类算法 NetBC(Network-aided Bi-Clustering)，不同于 NCIS，NetBC 不仅可以发现常量模式的簇，还可以发现共演变模式的簇.

这些基于网络的双聚类算法通常都是直接利用基因节点的度加权基因表达数据中的基因，这种方式设置的权重易受基因互作网络中噪声互作的干扰，影响聚类精度. 此外，基因互作网络中存在大量缺失互作[34]，也影响聚类精度. 而现有研究发现间接互作的基因间也很可能共享相同的功能，构成生物学通路，影响癌症的发生和发育[35]. 在分析已有研究工作的基础上，本文提出一种基于网络正则化的双聚类算法(Network Regularized Bi-Clustering，NetRBC). 不同于已有结合基因网络的双聚类算法，NetRBC 无需对基因进行加权，它不仅能够利用基因间的间接互作和网络拓扑结构，还可以得到更具有生物学意义和耦合性更好的基因簇和样本簇指示矩阵. 在多个癌症基因表达数据上的实验结果表明 NetRBC 比其它相关方法[8,18,20,31,33]能够更准确地区分癌症亚型.

3　基于网络约束的双聚类算法

3.1　最小化均方残差

癌症亚型在基因表达数据上一般对应五种类型的簇：常量簇、行常量簇、列常量簇类、加法模型簇和乘法模型簇[24]. NCIS 利用三元非负矩阵分解求解簇的行指示矩阵和列指示矩阵[31]. 类似地，文献[22]和文献[23]中的方法均利用矩阵的低秩表示重构原始矩阵，没有结合均方残差，所以该类方法只能挖掘常量簇，不能发现加法和乘法模型的簇；此外，它们均基于基因表达数据引入惩罚项约束基因簇和样本簇指示矩阵的稀疏分解，并未结合生物网络数据. 为挖掘上述 5 种类型的簇，受到 CC[5]和 MSSRCC[20]的启发，本文以均方残差为基准度量簇中元素的相关性.

假设基因表达数据矩阵为 $\boldsymbol{S}\in\mathbb{R}^{m\times n}$，$m$ 为基因数，n 表示样本数. $\boldsymbol{S}$ 中每个元素 s_{ij} 的均方残差可定义为

$$h_{ij}=(s_{ij}-\bar{s}_{Ij}-\bar{s}_{iJ}+\bar{s}_{IJ})^2 \tag{1}$$

其中 $\bar{s}_{iJ}$ 表示的是 s_{ij} 所在簇的行均值：

$$\bar{s}_{iJ}=\frac{1}{|J|}\sum_{j\in J}s_{ij} \tag{2}$$

$\bar{s}_{Ij}$ 表示的是 s_{ij} 所在簇的列均值：

$$\bar{s}_{Ij}=\frac{1}{|I|}\sum_{i\in I}s_{ij} \tag{3}$$

$\bar{s}_{IJ}$ 表示的是 s_{ij} 所在簇中所有元素的均值：

$$\bar{s}_{IJ}=\frac{1}{|I||J|}\sum_{i\in I,j\in J}s_{ij} \tag{4}$$

其中 I 和 J 分别表示簇中基因和样本的集合，$|I|$ 和 $|J|$ 表示集合中基因和样本的个数，h_{ij} 的值越小，则簇的相关性越高，理想的残差为 0.

在此基础上，本文定义基因簇指示矩阵 $\boldsymbol{R}$ 和样本簇指示矩阵 $\boldsymbol{C}$. $\boldsymbol{R}\in\mathbb{R}^{m\times k}\left(\sum_{k'=1}^{k}\boldsymbol{R}_{ik'}=1\right)$，其中 k 表示基因簇的个数，如果基因 i 属于第 k' 个簇则 $\boldsymbol{R}_{ik'}=1$，

否则 $\boldsymbol{R}_{ik'}=0$；$\boldsymbol{C}\in\mathbb{R}^{n\times l}\left(\sum_{l'=1}^{l}\boldsymbol{C}_{jl'}=1\right)$，其中 l 表示样本簇的个数，如果样本 j 属于第 l' 个簇，则 $\boldsymbol{C}_{jl'}=1$，否则 $\boldsymbol{C}_{jl'}=0$.

假设第 k' 个基因簇包含 $m_{k'}$ 个基因，$I_{k'}$ 表示第 k' 个簇所含基因的集合，则有 $\sum_{k'=1}^{k}|m_{k'}|=\sum_{k'=1}^{k}|I_{k'}|=m$，$\tilde{s}_{Ij}$ 可表示为

$$\begin{aligned}\tilde{s}_{Ij}&=\sum_{k'=1}^{k}\boldsymbol{R}_{ik'}\frac{1}{m_{k'}}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{S})_{k'j}\\&=\sum_{k'=1}^{k}\boldsymbol{R}_{ik'}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{-1}_{k'k'}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{S})_{k'j}\\&=(\boldsymbol{R}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{-1}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{S})_{ij}=\Psi_1(\boldsymbol{R})_{ij}\end{aligned}\tag{5}$$

其中 $(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{S})_{k'j}=\sum_{i'\in I_{k'}}\boldsymbol{S}_{i'j}$.

假设第 l' 个样本簇包含 $n_{l'}$ 个样本，$I_{l'}$ 表示第 l' 个簇的样本集合，则有 $\sum_{l'=1}^{l}|n_{l'}|=\sum_{l'=1}^{l}|I_{l'}|=n$，$\tilde{s}_{iJ}$ 可表示为

$$\begin{aligned}\tilde{s}_{iJ}&=\sum_{l'=1}^{l}(\boldsymbol{SC})_{il'}\frac{1}{n_{l'}}\boldsymbol{C}^{\mathrm{T}}_{l'j}\\&=\sum_{l'=1}^{l}(\boldsymbol{SC})_{il'}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})^{-1}_{l'l'}\boldsymbol{C}^{\mathrm{T}}_{l'j}\\&=(\boldsymbol{SC}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})^{-1}\boldsymbol{C}^{\mathrm{T}})_{ij}=\Psi_2(\boldsymbol{C})_{ij}\end{aligned}\tag{6}$$

其中 $(\boldsymbol{SC})_{il'}=\sum_{j\in I_{l'}}\boldsymbol{S}_{i'j}$.

类似地 $\tilde{s}_{IJ}$ 可表示为

$$\Psi_3(\boldsymbol{R},\boldsymbol{C})_{ij}=(\boldsymbol{R}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{-1}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{SC}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})^{-1}\boldsymbol{C}^{\mathrm{T}})_{ij}=\tilde{s}_{IJ}\tag{7}$$

结合式(1)和式(5)～(7)，$\boldsymbol{S}$ 中 mn 个元素上的均方残差总和为

$$\begin{aligned}\Psi(\boldsymbol{R},\boldsymbol{C})&=\|\boldsymbol{S}-\Psi_1-\Psi_2+\Psi_3\|^2\\&=\|(\boldsymbol{I}_m-\boldsymbol{R}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{-1}\boldsymbol{R}^{\mathrm{T}})\boldsymbol{S}(\boldsymbol{I}_n-\boldsymbol{C}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})^{-1}\boldsymbol{C}^{\mathrm{T}})\|^2\end{aligned}\tag{8}$$

其中 $\boldsymbol{R},\boldsymbol{C}\geq 0$，$\boldsymbol{I}_m\in\mathbb{R}^{m\times m}$ 和 $\boldsymbol{I}_n\in\mathbb{R}^{n\times n}$ 均为单位矩阵.

3.2 结合基因互作网络信息

考虑到基因表达数据矩阵中基因个数普遍大于样本数，样本之间的相似度很难衡量[13]. 部分算法在聚类之前会先对基因进行筛选，如 MSSRCC 基于基因表达谱的方差筛选基因，保留方差变化较大的基因，因为这样的基因更有可能是致病基因，再在筛选后的表达数据上定义均方残差损失函数. 但是这样的筛选很有可能将那些具有重要生物学意义的基因筛掉，因为表达值变化很小的基因同样可能导致疾病. 基因/蛋白质互作网络对理解和研究癌症等复杂疾病的模式具有重要作用[6]. 因此，一些双聚类算法结合基因互作网络辅助关键基因的选择，以期提升癌症亚型的分类精度[27,31,33]. 如 NCIS 和 NetBC 均通过结合基因互作网络和基因表达谱特征加权选择基因，约束基因簇和样本簇矩阵优化求解. 然而，已有研究发现基因互作网络数据存在大量的缺失互作和噪声互作[35]，因此，利用基因网络选择加权关键基因的可靠性有限，相应获取的聚类簇质量有限，影响癌症亚型分类的精度.

研究发现存在间接互作的基因也很有可能共享相同的功能、构成通路，影响疾病的发生，这些基因也应该划分到同一个簇中[35]. 为此，本文引入正则项如下：

$$\begin{aligned}&\frac{1}{2}\sum_{i=1,j=1}^{m}\|\boldsymbol{R}_{i\cdot}-\boldsymbol{R}_{j\cdot}\|\boldsymbol{P}_{ij}\\&=\sum_{i=1,j=1}^{m}\boldsymbol{R}_{i\cdot}\boldsymbol{P}_{ij}\boldsymbol{R}^{\mathrm{T}}_{i\cdot}-\sum_{i=1,j=1}^{m}\boldsymbol{R}_{i\cdot}\boldsymbol{P}_{ij}\boldsymbol{R}^{\mathrm{T}}_{j\cdot}\\&=\sum_{i=1,j=1}^{m}\boldsymbol{R}_{i\cdot}\boldsymbol{P}_{ii}\boldsymbol{R}^{\mathrm{T}}_{i\cdot}-\sum_{i=1,j=1}^{m}\boldsymbol{R}_{i\cdot}\boldsymbol{P}_{ij}\boldsymbol{R}^{\mathrm{T}}_{j\cdot}\\&=\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{P})\boldsymbol{R})=\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{R})\end{aligned}\tag{9}$$

不同于 Gu 等人[36]在流形假设的基础上利用样本特征构建 k 近邻图再定义正则化项，这种正则项易受高维样本噪声特征的影响，本文直接利用基因互作网络信息定义正则化项. 式(9)中 $\boldsymbol{P}$ 表示基因互作网络的邻接矩阵，如果基因 i 与基因 j 存在互作，则 $\boldsymbol{P}_{ij}=1$，否则 $\boldsymbol{P}_{ij}=0$；tr( )为矩阵的迹运算，$\boldsymbol{D}_{ii}=\sum_j\boldsymbol{P}_{ij}$ 是一个对角矩阵，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{P}$ 是图拉普拉斯矩阵[37]. 最小化式(9)可以使得具有较强交互($\boldsymbol{P}_{ij}>0$)的成对基因(基因 i 和基因 j)对应的基因簇指示矩阵 $\boldsymbol{R}_{i\cdot}$ 和 $\boldsymbol{R}_{j\cdot}$ 更可能相似，这对基因更可能划分到同一个簇中，因为存在互作的基因更可能共享相同的功能，构成生物学通路. 需要说明的是，基因表达数据预处理阶段的方差分析能够识别出表达值明显异常的致病基因，这类基因通常是对疾病产生存在较大影响的单个基因，这类基因在双聚类时会被划分到与其表达最相似的基因簇中. 值得指出的是，式(9)还考虑了基因间的间接互作，若 $\boldsymbol{P}_{ij}>0$ 和 $\boldsymbol{P}_{jk}>0$ 但 $\boldsymbol{P}_{ik}=0$，最小化式(9)也会使得对应的 $\boldsymbol{R}_{i\cdot}$ 和 $\boldsymbol{R}_{k\cdot}$ 相似，因为间接互作的基因也可能参与到同一个生物学通路中，影响癌症的发生和发展. 这表明式(9)定义的基于基因网络的正则项不仅能够利用网络的拓扑结构信息，还能在一定程度上克服缺失互作的影响和减少噪声互作的误导.

在上述分析的基础上，结合式(8)和(9)，可得 NetRBC 的目标方程：

$$\Phi(\boldsymbol{R},\boldsymbol{C})=\Psi(\boldsymbol{R},\boldsymbol{C})+\lambda\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{R}) \tag{10}$$

其中 $\lambda\geq 0$ 为正则化系数，从式(10)可以发现基因网络可以调节基因簇指示矩阵 $\boldsymbol{R}$ 的优化分解，进而约束指导样本簇指示矩阵 $\boldsymbol{C}$ 的优化. 后文的实验会证明引入式(9)定义的基于基因网络的正则项可显著地提升双聚类的性能和癌症亚型分类的精度.

3.3　目标方程求解

为求解 $\boldsymbol{R}$ 和 $\boldsymbol{C}$，NetRBC 采用一种类似期望最大化[38]的方法分别固定 $\boldsymbol{R}$(或 $\boldsymbol{C}$)，再迭代交替更新 $\boldsymbol{C}$(或 $\boldsymbol{R}$)，直至达到指定迭代次数或均方残差小于设定的阈值. 目标方程求解的收敛性分析参见附录 1. 具体求解方案如下：

令 $\boldsymbol{X}_1=\boldsymbol{S}(\boldsymbol{I}_n-\boldsymbol{C}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})^{-1}\boldsymbol{C}^{\mathrm{T}})$，$\boldsymbol{X}_2=(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{-1}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{X}_1$，假定 $\boldsymbol{C}$ 已知，则原目标方程可化简为以 $\boldsymbol{R}$ 为参数的目标函数：

$$\begin{aligned}Q(\boldsymbol{R})&=\|\boldsymbol{X}_1-\boldsymbol{R}\boldsymbol{X}_2\|^2+\lambda\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{R})\\&=\mathrm{tr}(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1-2\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{R}\boldsymbol{X}_2+\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{X}_2)+\\&\quad\lambda\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{R})\end{aligned} \tag{11}$$

$$\text{s.t. } \boldsymbol{R}\geq 0,\ \sum_{k'=1}^{k}\boldsymbol{R}_{i,k'}=1$$

令 $\boldsymbol{\alpha}\in\mathbb{R}^{m\times k}$ 为 $\boldsymbol{R}$ 的拉格朗日乘子，则以 $\boldsymbol{R}$ 为参数的拉格朗日函数为

$$L(\boldsymbol{R},\boldsymbol{\alpha})=Q(\boldsymbol{R})-\mathrm{tr}(\boldsymbol{\alpha}\boldsymbol{R}^{\mathrm{T}}) \tag{12}$$

令 $\dfrac{\partial L(\boldsymbol{R})}{\partial\boldsymbol{R}}=0$ 得到式(13)：

$$\boldsymbol{\alpha}=2\lambda\boldsymbol{L}\boldsymbol{R}-2\boldsymbol{A}+2\boldsymbol{R}\boldsymbol{B} \tag{13}$$

其中 $\boldsymbol{A}=\boldsymbol{X}_1\boldsymbol{X}_2^{\mathrm{T}}$，$\boldsymbol{B}=\boldsymbol{X}_2\boldsymbol{X}_2^{\mathrm{T}}$. 由 Karush-Kuhn-Tucker (KKT)[39]条件，令 $\boldsymbol{\alpha}_{ij}\boldsymbol{R}_{ij}=0$ 可得：

$$(\lambda\boldsymbol{L}\boldsymbol{R}-\boldsymbol{A}+\boldsymbol{R}\boldsymbol{B})_{ij}\boldsymbol{R}_{ij}=0 \tag{14}$$

令 $\boldsymbol{L}=\boldsymbol{L}^{+}+\boldsymbol{L}^{-}$，$\boldsymbol{A}=\boldsymbol{A}^{+}+\boldsymbol{A}^{-}$，$\boldsymbol{B}=\boldsymbol{B}^{+}+\boldsymbol{B}^{-}$，其中 $\boldsymbol{A}^{+}=(|\boldsymbol{A}|+\boldsymbol{A})/2$，$\boldsymbol{A}^{-}=(|\boldsymbol{A}|-\boldsymbol{A})/2$，$\boldsymbol{B}^{+}$，$\boldsymbol{B}^{-}$ 类似的定义. 可得式(15)：

$$(\lambda\boldsymbol{L}^{+}\boldsymbol{R}-\lambda\boldsymbol{L}^{-}\boldsymbol{R}-\boldsymbol{A}^{+}+\boldsymbol{A}^{-}+\boldsymbol{R}\boldsymbol{B}^{+}-\boldsymbol{R}\boldsymbol{B}^{-})\boldsymbol{R}=0 \tag{15}$$

由此可得 $\boldsymbol{R}$ 的更新公式：

$$\boldsymbol{R}_{ik'}=\boldsymbol{R}_{ik'}\sqrt{\frac{[\lambda\boldsymbol{L}^{-}\boldsymbol{R}+\boldsymbol{A}^{+}+\boldsymbol{R}\boldsymbol{B}^{-}]_{ik'}}{[\lambda\boldsymbol{L}^{+}\boldsymbol{R}+\boldsymbol{A}^{-}+\boldsymbol{R}\boldsymbol{B}^{+}]_{ik'}}} \tag{16}$$

令 $\boldsymbol{X}_3=(\boldsymbol{I}_m-\boldsymbol{R}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{-1}\boldsymbol{R}^{\mathrm{T}})\boldsymbol{S}$，$\boldsymbol{X}_4=\boldsymbol{X}_3\boldsymbol{C}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})^{-1}$. 假定 $\boldsymbol{R}$ 已知，可得以 $\boldsymbol{C}$ 为参数的目标函数：

$$\begin{aligned}Q(\boldsymbol{C})&=\|\boldsymbol{X}_3-\boldsymbol{X}_4\boldsymbol{C}^{\mathrm{T}}\|^2+\lambda\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{R})\\&=\mathrm{tr}(\boldsymbol{X}_3^{\mathrm{T}}\boldsymbol{X}_3-2\boldsymbol{X}_3^{\mathrm{T}}\boldsymbol{X}_4\boldsymbol{C}^{\mathrm{T}}+\boldsymbol{C}\boldsymbol{X}_4^{\mathrm{T}}\boldsymbol{X}_4\boldsymbol{C}^{\mathrm{T}})+\\&\quad\lambda\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{R})\end{aligned} \tag{17}$$

$$\text{s.t. } \boldsymbol{C}\geq 0,\ \sum_{l'=1}^{l}\boldsymbol{C}_{j,l'}=1$$

令 $\boldsymbol{\beta}\in\mathbb{R}^{n\times l}$ 为 $\boldsymbol{C}(\boldsymbol{C}\geq 0)$ 的拉格朗日乘子，则以 $\boldsymbol{C}$ 为参数的拉格朗日函数为

$$L(\boldsymbol{C},\boldsymbol{\beta})=Q(\boldsymbol{C})-\mathrm{tr}(\boldsymbol{\beta}\boldsymbol{C}^{\mathrm{T}}) \tag{18}$$

令 $\dfrac{\partial L(\boldsymbol{C})}{\partial\boldsymbol{C}}=0$ 可得：

$$\boldsymbol{\beta}=-\boldsymbol{D}+\boldsymbol{C}\boldsymbol{F} \tag{19}$$

其中 $\boldsymbol{D}=\boldsymbol{X}_3^{\mathrm{T}}\boldsymbol{X}_4$，$\boldsymbol{F}=\boldsymbol{X}_4^{\mathrm{T}}\boldsymbol{X}_4$. 由 KKT 条件，令 $\boldsymbol{\beta}_{ij}\boldsymbol{C}_{ij}=0$ 可得：

$$(\boldsymbol{D}-\boldsymbol{C}\boldsymbol{F})_{ij}\boldsymbol{C}_{ij}=0 \tag{20}$$

令 $\boldsymbol{D}=\boldsymbol{D}^{+}+\boldsymbol{D}^{-}$，$\boldsymbol{F}=\boldsymbol{F}^{+}+\boldsymbol{F}^{-}$，其中 $\boldsymbol{D}^{+}=(|\boldsymbol{D}|+\boldsymbol{D})/2$，$\boldsymbol{D}^{-}=(|\boldsymbol{D}|-\boldsymbol{D})/2$，$\boldsymbol{F}^{+}$ 和 $\boldsymbol{F}^{-}$ 类似定义，可得：

$$(\boldsymbol{D}^{+}-\boldsymbol{D}^{-}+\boldsymbol{C}\boldsymbol{F}^{+}-\boldsymbol{C}\boldsymbol{F}^{-})_{ij}\boldsymbol{C}_{ij}=0 \tag{21}$$

由此可得 $\boldsymbol{C}$ 的更新公式：

$$\boldsymbol{C}_{jl'}=\boldsymbol{C}_{jl'}\sqrt{\frac{[\boldsymbol{D}^{+}+\boldsymbol{C}\boldsymbol{F}^{-}]_{jl'}}{[\boldsymbol{D}^{-}+\boldsymbol{C}\boldsymbol{F}^{+}]_{jl'}}} \tag{22}$$

在上述分析和算法优化求解的基础上，本文给出 NetRBC 算法的一般流程，如算法 1 所示.

算法 1.　NetRBC.

输入：基因表达矩阵 $\boldsymbol{S}$；基因簇个数 k

　　样本簇个数 l；基因互作网络 $\boldsymbol{P}$

输出：基因簇指示矩阵 $\boldsymbol{R}$，样本簇指示矩阵 $\boldsymbol{C}$

1. $\boldsymbol{L}\leftarrow$基于互作网络 $\boldsymbol{P}$ 计算拉普拉斯矩阵
2. $\boldsymbol{R},\boldsymbol{C}\leftarrow\mathrm{Init}(\boldsymbol{R},\boldsymbol{C},k,l)$ //初始化 $\boldsymbol{R},\boldsymbol{C}$
3. WHILE(not convergence)
4. 　$\boldsymbol{R}_{ik'}\leftarrow update(\boldsymbol{R}_{ik'})$ //利用式(16)更新 $\boldsymbol{R}$
5. 　$\boldsymbol{C}_{jl'}\leftarrow update(\boldsymbol{C}_{jl'})$ //利用式(22)更新 $\boldsymbol{C}$
6. END WHILE

4　实验分析

4.1　在癌症亚型已知的基因表达数据集上的实验

本实验采用 6 个已知癌症亚型的基因表达数据集进行实验，表 1 中列出了这些数据集的相关信息. 其中，Breast 为乳腺癌数据集，包含 3 种亚型，各样本数量为 34、44、20；DLBCL 是弥漫大 B 细胞淋巴瘤数据集，包含了 4 种亚型，各样本数量为 22、39、25、50；Multi 是混合了 4 种不同癌症亚型的数据集，各样本数量为 26、26、23、28；ALL 是急性淋巴细胞白血病数据集，包含了 3 种亚型，各样本数量为 19、11、8；TOX 为混合了 4 种癌症亚型的基因表达数据，各样本数量为 45、45、39、42；Liver 是肝癌数据集，包含 4 种亚型，各样本数目为 67、32、21、2.

表 1 癌症亚型已知的基因表达数据集

数据集	来源	亚型种类	样本个数	基因个数
Breast	[41]	3	98	1213
DLBCL	[42]	4	129	3795
Multi	[42]	4	103	5565
ALL	[43]	3	38	5571
TOX	[44]	4	170	5748
Liver	[45]	4	122	8799

由于上述数据集中基因与具体癌症亚型的关联信息未知，所以本文针对样本的聚类结果进行评价分析. 本文从 BioGRID 数据库[40]（访问日期：2017-4-30）下载了初始的基因互作网络，并根据不同数据集中的相关基因，筛选出其对应的基因互作网络 $\boldsymbol{P}$.

本文采用聚类中常用的度量指标 Rand Index (RI)[46]、$F1$-measure[47]、$Accuracy$ 和 NMI[48] 评价癌症亚型分类的效果. 为了确定基因的聚类个数，本文在每种个数设置下独立运行了 50 次 k-means 算法，再基于一致性聚类算法[10]选择其表型相关系数[49]最大时的 k 作为基因的聚类个数，样本簇个数 l 取真实的亚型个数，k 和 l 的取值如表 2.

表 2 基因簇个数 k 和样本簇个数 l

数据集	k	l
Breast	5	3
DLBCL	7	4
Multi	8	4
ALL	5	3
TOX	6	4
Liver	6	4

为了分析引入网络正则化项的贡献，本文首先对系数 λ 设置搜索空间为 $10^{-2}\sim10^{8}$，从 10^{-2} 开始每次扩大 10 倍. 本文在 6 个数据集上对比 $\lambda=0$，即不加入网络正则化项和 NetRBC 的运行结果. 值得指出的是，NetRBC($\lambda=0$)与 NetBC 不同，NetBC 加权基因，而 NetRBC($\lambda=0$)没有. 图 1 是在评价度量 RI(10 次独立运行的均值)下的实验结果；图 2 是在评价度量 $F1$-measure 下的实验结果. 从图中可以看出，随着 λ 增大到 10^{5} 左右，NetRBC 不仅比 NetRBC($\lambda=0$)获得更好的聚类结果，其方差也逐步减小. 这是因为存在的直接互作或间接互作的基因更有可能共享相同的功能、构成通路，影响疾病的发生，而 λ 的增大使得基因互作网络对基因簇指示矩阵和样本簇指示矩阵的约束指导作用更大. 从图 1 和图 2 中的结果可知在基因表达数据上结合基因互作网络，可以降低基因表达数据中噪声的干扰，提升双聚类性能，进而提高癌症亚型的分类精度. 由于基因互作网络本身存在缺失互作和噪声互作，各种疾病相关的基因互作模式存在差异，所以对于不同的数据集，引入基因互作网络的贡献效果不同且最优的 λ 设置也不同. 例如在 Breast 和 ALL 数据集上，当 λ 取 $10^{0}\sim10^{2}$ 时，结果的精度有所下降，为此，可通过调节 λ 优化基因互作网络对双聚类效果的贡献. 与 NCIS 和 NetBC 直接利用基因网络加权基因的方式相比，本文的方法更为灵活和合理.

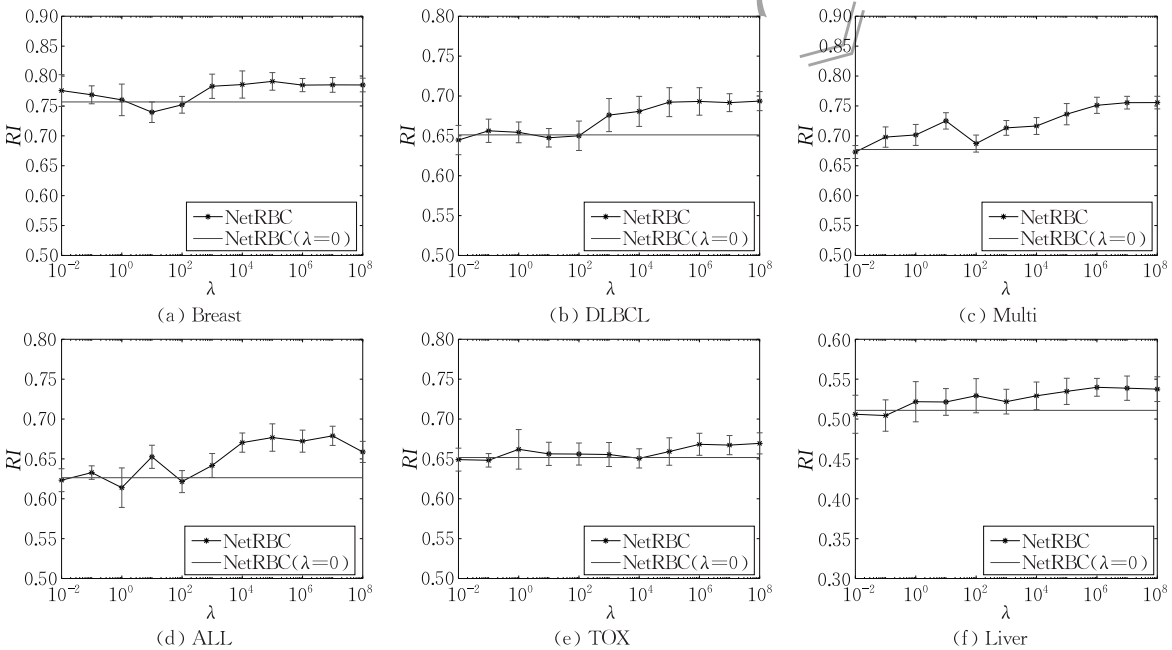

图 1 NetRBC 在不同 λ 参数值下的 RI

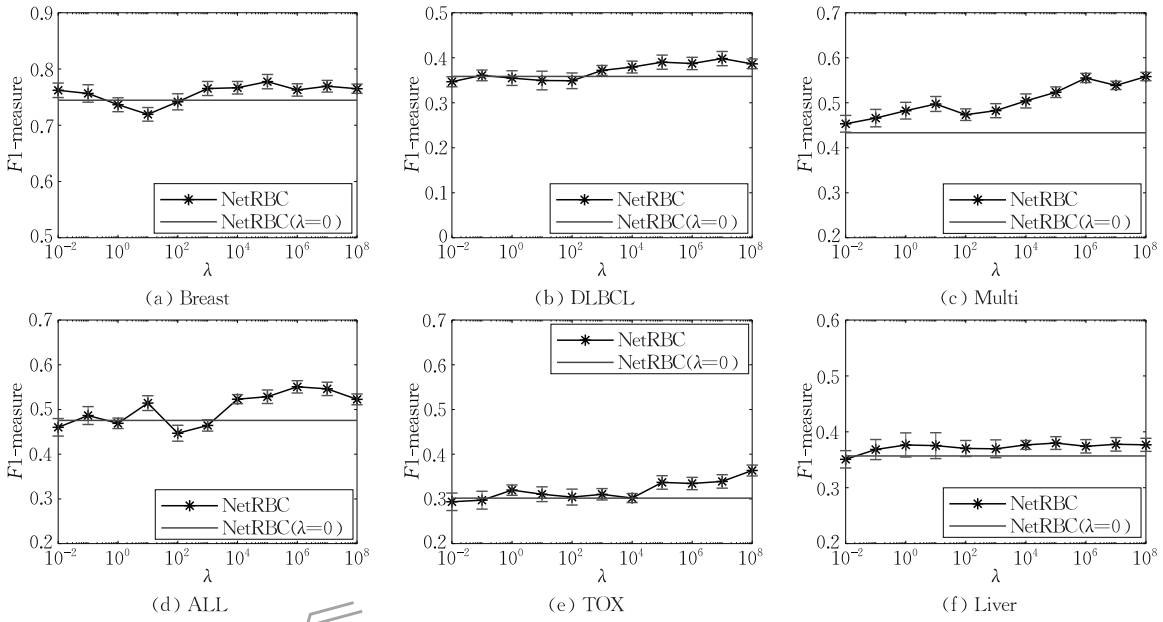

图 2　NetRBC 在不同 λ 参数值下的 $F1$-measure

为进一步对比分析 NetRBC 的性能，本文选择 NetBC[33]、NCIS[31]、NMTF[8]、MSSRCC[20]、SBC[18] 和 k-means 作为对比算法，因为 CC[5]、QUBIC[25] 和 SGFA[26] 找到的双聚类簇存在重叠，而本文采用的真实数据集中一个病人只有一个癌症亚型标记，所以本文不与这三个方法对比. 本文在每个数据集独立运行这些对比算法 10 次，并将对应结果汇报在表 3～表 6 中.

表 3　对比算法在已知癌症亚型基因表达数据上的实验结果（*RI*）

算法	Breast	DLBCL	Multi	ALL	TOX	Liver
NetRBC	**0.7911±0.011**	**0.6923±0.010**	**0.7363±0.018**	**0.6767±0.017**	**0.6593±0.017**	0.5347±0.016
NetBC	0.7203±0.018	0.6263±0.016	0.7018±0.013	0.6147±0.012	0.6437±0.013	**0.5670±0.019**
NCIS	0.6704±0.012	0.6129±0.004	0.6323±0.003	0.6296±0.009	0.6316±0.004	0.5582±0.012
NMTF	0.5982±0.018	0.6237±0.010	0.6587±0.015	0.6149±0.015	0.6271±0.003	0.5493±0.012
MSSRCC	0.7008±0.013	0.6305±0.005	0.6662±0.017	0.6238±0.015	0.6541±0.012	0.5235±0.014
SBC	0.4124±0.000	0.4179±0.000	0.6488±0.017	0.3613±0.000	0.6317±0.001	0.4111±0.000
k-means	0.7752±0.012	0.6119±0.014	0.6429±0.015	0.6110±0.019	0.6348±0.016	0.5413±0.015

表 4　对比算法在已知癌症亚型基因表达数据上的实验结果（*F1-measure*）

算法	Breast	DLBCL	Multi	ALL	TOX	Liver
NetRBC	**0.7776±0.012**	**0.3902±0.013**	**0.5235±0.018**	0.5286±0.012	0.3369±0.015	0.3742±0.011
NetBC	0.6622±0.012	0.3127±0.016	0.4341±0.010	0.4263±0.014	0.3079±0.016	0.3585±0.012
NCIS	0.6479±0.009	0.2700±0.007	0.3187±0.012	0.4711±0.012	0.2554±0.010	0.3720±0.013
NMTF	0.6578±0.009	0.2901±0.015	0.3557±0.016	0.4540±0.015	0.2499±0.006	0.3481±0.015
MSSRCC	0.6704±0.018	0.3212±0.010	0.4078±0.014	0.4481±0.013	0.3156±0.018	0.3651±0.012
SBC	0.5839±0.000	0.3739±0.000	0.4471±0.009	**0.5308±0.000**	**0.3458±0.001**	**0.4564±0.000**
k-means	0.7489±0.017	0.3085±0.014	0.3390±0.013	0.4408±0.012	0.3162±0.017	0.3735±0.015

表 5　对比算法在已知癌症亚型基因表达数据上的实验结果（*Accuracy*）

算法	Breast	DLBCL	Multi	ALL	TOX	Liver
NetRBC	**0.9373±0.010**	**0.4936±0.009**	**0.7750±0.015**	0.7106±0.014	0.4653±0.013	**0.5503±0.013**
NetBC	0.8694±0.011	0.4527±0.013	0.6313±0.012	0.5974±0.012	0.4082±0.011	0.4344±0.013
NCIS	0.7765±0.018	0.3550±0.020	0.5969±0.013	0.5079±0.012	0.3357±0.013	0.4779±0.015
NMTF	0.8357±0.013	0.3961±0.011	0.5000±0.011	0.6158±0.015	0.3287±0.013	0.4148±0.011
MSSRCC	0.7969±0.016	0.4767±0.013	0.5594±0.011	0.7158±0.012	0.4801±0.010	0.5172±0.012
SBC	0.5500±0.016	0.4690±0.015	0.5125±0.012	0.7053±0.013	**0.5708±0.003**	0.5197±0.000
k-means	0.8255±0.008	0.4318±0.014	0.5125±0.011	**0.7237±0.012**	0.4813±0.011	0.5067±0.017

表 6 对比算法在已知癌症亚型基因表达数据上的实验结果(*NMI*)

算法	Breast	DLBCL	Multi	ALL	TOX	Liver
NetRBC	**0.5806±0.008**	**0.1556±0.016**	**0.5591±0.005**	0.3503±0.011	**0.1674±0.015**	0.1374±0.015
NetBC	0.5021±0.012	0.1097±0.017	0.3765±0.018	0.2162±0.011	0.0801±0.011	0.0947±0.014
NCIS	0.2182±0.019	0.0374±0.018	0.2877±0.011	0.1357±0.017	0.0235±0.018	0.1108±0.012
NMTF	0.3364±0.013	0.0737±0.016	0.2254±0.016	0.2474±0.014	0.0198±0.010	0.0813±0.011
MSSRCC	0.5274±0.019	0.1462±0.015	0.3384±0.010	**0.5552±0.011**	0.1289±0.010	**0.1576±0.010**
SBC	0.2484±0.000	0.1476±0.000	0.4984±0.016	0.0145±0.000	0.1169±0.001	0.1498±0.012
k-means	0.5647±0.014	0.1207±0.004	0.2742±0.018	0.5402±0.017	0.1386±0.015	0.1504±0.012

从表 3 中 k-means 与 NetRBC 的 RI 结果可以看出,随着基因维度的增加,NetRBC 和 k-means 的性能均有所下降,这是由于基因数量大使得样本之间相似性的度量不准确,进而影响了这些聚类算法的精度.k-means 比 NetRBC 的性能下降速度更快的原因是它基于所有的基因度量样本之间的相似度,而 NetRBC 仅利用部分基因.从表 3 中 NetBC、NCIS 和 NMTF 的对比结果可以看出,利用了基因互作网络信息的 NetBC 和 NCIS 相对 NMTF 在大部分数据集上均有所提升,表明结合基因互作网络数据可以提升疾病亚型的分类精度.然而,提升的幅度不是很明显,这是因为 NetBC 和 NCIS 基于基因互作网络加权基因,网络本身存在缺失和噪声互作,误导了基因的加权.NetRBC 通过定义基于基因互作网络的正则项调控基因簇和样本簇指示矩阵分解,在一定程度上减轻了网络中缺失互作的影响,同时也降低了噪声互作的干扰,所以其提升效果相对 NetBC 和 NCIS 更显著.从表 4 中 $F1$-measure 的结果来看,虽然 MSSRCC 没有结合基因互作网络,但它的结果有时候优于结合基因网络的 NCIS,这是因为 MSSRCC不仅能够发现常量簇还能发现共演变的簇,而 NMTF 和 NCIS 仅能挖掘常量簇.在 $F1$-measure指标上,SBC 表现较好,这是由于 SBC 在双聚类之前先对原始矩阵进行了特征分解和重构,降低了噪声的干扰,所以它在一些数据集上获得了较其它算法更好的性能.但是由于数据集来自不同的基因芯片且经过不同的预处理,SBC 的特征分解和重构仅在部分数据集上有效,而在其它数据集上的结果仍不及 NetRBC.在 Liver 数据集上,NetRBC、NetBC、MSSRCC 和 SBC 分别在不同的评价度量上获得最好的结果,NetRBC 在整体上依然获得比这些方法更好的结果.这些评价度量从不同的角度评价聚类结果,一个算法很难在各个角度都超越另外一个算法,也很难在不同数据集上均超越另一个算法.

为了更好地展示双聚类的结果,本文将各个算法在 Breast 数据集上的聚类结果用热图的方式在图 3

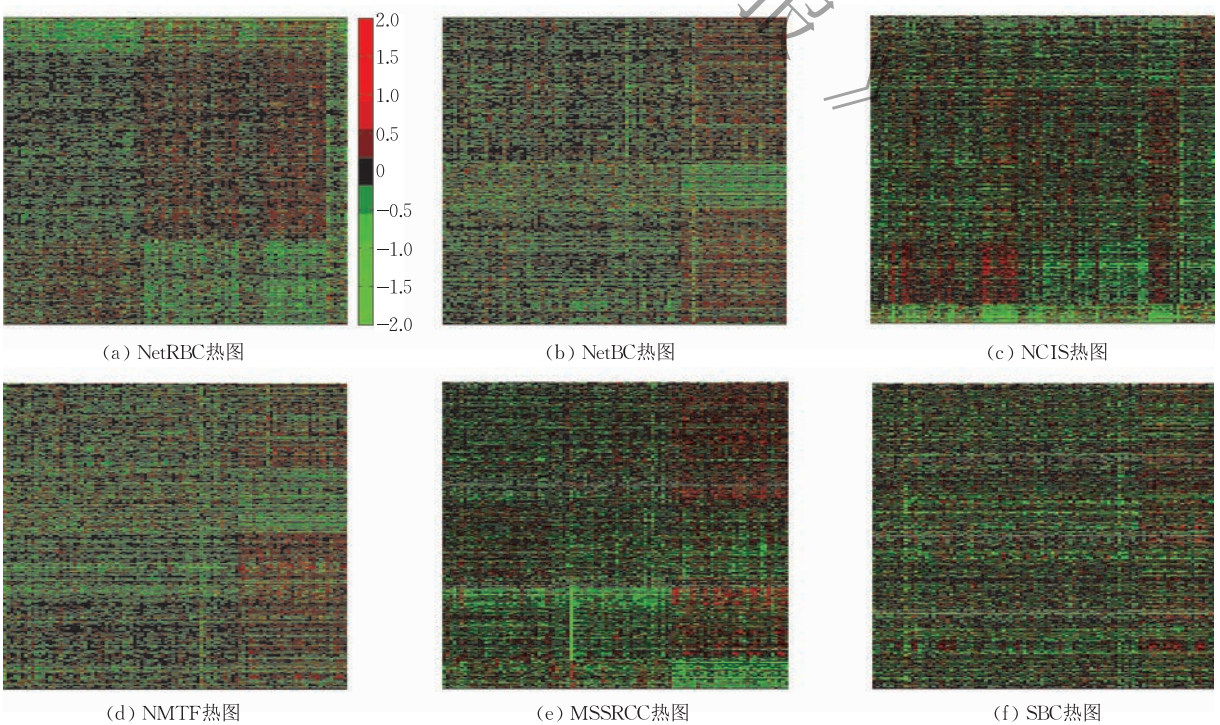

图 3 不同算法在 Breast 数据集上获取的双聚类结果的热图(行表示基因,列表示样本)

中进行展示. 图中行表示基因，列表示样本，块结构是一个双聚类簇，表示部分基因在部分样本上的表达值具有较强的相关性，从图中可以看出，NetRBC 发现的双聚类簇的块结构不仅相对其他算法更明显，而且遍布整个基因表达矩阵，而 NetBC 和 NMTF 偏向于基因表达矩阵的右侧，块结构也不明显. 值得指出的是，癌症亚型的产生和发育过程是多层面因素共同作用的结果，因此仅利用基因表达数据只能从图 3 中粗略地区分出各种癌症亚型.

为了进一步证明 NetRBC 对缺失互作和噪声互作具有较好的鲁棒性，本文在 Breast 数据集上进行如下实验：(1) 随机增加基因互作网络中的边；(2) 随机删除基因互作网络中的边；(3) 同时等量随机删除和增加基因互作网络中的边. 在以上三种设置下，比较三种算法的性能，并将结果报告在图 4 中. 从图 4 可以看出，NetRBC、NetBC 和 NCIS 在一定程度上都受到了缺失互作和噪声互作的影响，结果均有所波动，但是 NetRBC 的结果明显优于 NetBC 和 NCIS，波动更小，这说明 NetRBC 对基因互作网络的变化更鲁棒，原因是基于网络正则项的 NetRBC 能够较好的利用基因网络的拓扑结构信息，而 NetBC 和 NCIS 基于网络结构加权基因，易受噪声和缺失互作的影响. 此外，我们还发现在某些情况下随机加入或者删除一些边时，以上三种算法均出现精度上升的现象，这是因为基因互作网络本身存在缺失互作和噪声互作，随机增加边可能会填充部分缺失互作，而随机删除边可能会去除部分噪声互作，从而提升基因互作网络本身的质量，提高聚类精度.

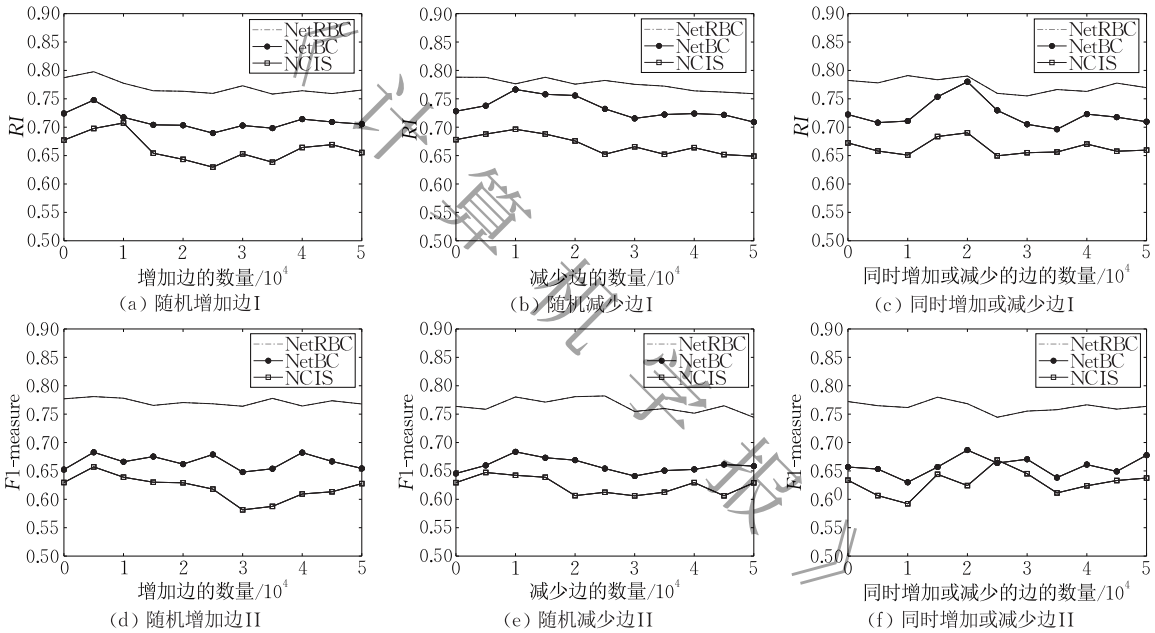

图 4　随机增加或剔除基因网络中互作边对 NetRBC、NetBC 和 NCIS 性能的影响(Breast 数据集上)

上述对比实验进一步证明，引入基于基因互作网络的正则项可更准确地发现癌症亚型；与利用基因网络加权基因的方法相比，本文的方法对双聚类算法性能的提升更为显著.

4.2　在 TCGA 癌症基因表达数据集上实验

本文还从 TCGA[50] 数据库下载了乳腺癌(Breast invasive Carcinoma，BRCA)和肺癌(Lung cancer，LUNG) 的大规模癌症基因表达数据和相关的临床数据，并从 BioGRID 下载了相关基因之间的互作数据. 这 2 个数据集中的癌症亚型未知，参照 NCIS 和 NetBC，本文通过临床特征来评估聚类结果的质量. 原始的 BRCA 数据集包含 1215 个样本，20530 个基因，本文选择 4 个临床特征：ACJJ stage、Node Code、Tumor 百分比和 Survival. 有些样本的临床信息是缺失的，本文剔除这些样本后，基于 779 个样本和 20530 个基因进行实验. 原始 LUNG 数据集包含 1124 个样本，20530 个基因，本文选择 4 个临床特征：Necrosis 百分比、Tumor Cell 百分比、Tumor Nuclei 百分比和 Survival. 剔除临床特征缺失的样本，最终基于 901 个样本和 20530 个基因进行实验. 为了确定基因和样本的簇个数，本文选择 $k\in[3,12]$，$l\in[3,8]$，对于不同的 k，l 组合独立运行 k-means 50 次，再基于一致性聚类算法选择共表型相关系数[49] 最大时的 k 与 l 组合. 具体地，在 BRCA 数据集上本文设置 $k=8$，$l=4$；在 LUNG 数据集上设置 $k=6$，$l=4$.

在完成上述实验设置后，本文同样分别在 BRCA 和 LUNG 数据集上独立运行 6 个相关的对比算法，

这些算法参数设置与前一节一样.此外,本文还从TCGA下载了mRNA和DNA甲基化数据,并与ConC[10]、PINS[28]、SNF[29]和iClusterPlus[30]这四个单聚类集成方法进行实验对比,表7、表8分别汇报了这些算法在不同临床指标下的检验值.

表7 对比算法在TCGA乳腺癌数据集上临床特征的p-value

算法	ACJJ	NodeCode	Tumor	Survial
NetRBC	1.5228E-03	**8.8798E-05**	**3.7278E-06**	3.1656E-02
NetBC	4.4106E-02	8.9104E-03	2.9006E-02	1.1293E-01
NCIS	6.3862E-04	6.3862E-04	5.2137E-06	7.0297E-02
NMTF	4.1909E-04	1.0408E-04	3.6913E-03	7.5142E-02
MSSRCC	7.5845E-04	1.2392E-04	3.3069E-05	4.4514E-02
SBC	4.3565E-01	4.7699E-02	1.4606E-01	1.2668E-01
k-means	6.3862E-04	1.0408E-04	4.1909E-04	7.0297E-02
ConC	4.7237E-03	3.5925E-04	5.6835E-04	6.6948E-02
PINS	**9.1937E-05**	7.5404E-02	8.8535E-04	5.3756E-03
SNF	9.8682E-04	8.4982E-03	7.6131E-03	**2.8047E-03**
iClusterPlus	3.3759E-03	6.8815E-03	7.0593E-03	8.8634E-03

表8 对比算法在TCGA肺癌数据集上临床特征的p-value

算法	Necrosis	TumorCells	TumorNuclei	Survial
NetRBC	**4.7290E-02**	**7.3426E-03**	2.8373E-02	1.1376E-02
NetBC	1.7894E-01	3.2969E-02	3.4210E-02	3.2978E-02
NCIS	9.5730E-02	6.1266E-02	3.4331E-01	1.7886E-01
NMTF	1.8964E-01	4.2285E-02	1.8879E-01	3.4663E-01
MSSRCC	8.6407E-02	6.6362E-02	2.3440E-03	5.3018E-02
SBC	1.0916E-01	2.3191E-02	8.5662E-01	7.1746E-01
k-means	7.0385E-02	1.6234E-02	**2.9211E-03**	2.3522E-02
ConC	7.0936E-02	6.5510E-02	5.9744E-02	7.5127E-02
PINS	7.5469E-02	9.2612E-03	3.4039E-02	**5.5095E-03**
SNF	7.6025E-01	1.1900E-02	5.8527E-03	3.0596E-02
iClusterPlus	6.7970E-02	4.9836E-02	2.2381E-02	4.9908E-02

本文对临床特征Survival做对数秩检验,其他临床特征做卡方检验,并将卡方检验得到的p-value用Benjamini和Hochberg方法[51]进行了调整.不同癌症亚型的临床特征存在明显的差异性,通过检验之后得到的p-value越小,表明聚类结果越显著,聚类效果越好.从表中可以看出NetRBC在多个临床特征上p-value为最小,且小于显著性水平0.05,表明NetRBC相对于其它算法能更有效地利用基因互作关系区分癌症亚型,聚类效果相对其他算法更为显著.不同于NetRBC使用基因互作信息作为先验知识指导聚类,PINS、SNF和iClusterPlus融合了多源数据进行癌症亚型分类.从表7、表8可知,PINS和SNF在BRCA数据集上分别在ACJJ和生存分析的临床特征上取得最好结果,PINS在LUNG数据集的生存分析中同样取得最好结果,说明整合多源数据可以提高预测精度.但在其他临床特征上NetRBC仍优于这些算法,是因为这些集成方法在整合多源数据的过程中仍然存在一定的不足.iClusterPlus通过选择基因减少时间复杂度,但使得其结果依赖于挑选的基因;SNF基于图的融合和核方法聚类集成整合数据,但它对参数比较敏感;PINS没有考虑不同癌症基因组数据在亚型分类时的差异贡献,而且它为了增加算法鲁棒性,针对单源数据运行多次基础聚类算法,增加了时间耗费[28].此外,上述基于单聚类集成的方法无法同时对基因和样本进行聚类.

综合上述实验结果,本文认为引入基因互作网络指导双聚类可以显著地提升双聚类的性能,NetRBC能较已有相关双聚类算法更准确地发现癌症亚型.

4.3 时间复杂度分析

结合3.3小节可知,NetRBC计算$\boldsymbol{X}_1$和$\boldsymbol{X}_2$对应的时间复杂度分别为$O(nl^2+n^2l+n^2+mn^2)$,$O(mk^2+mnk)$,计算$\boldsymbol{A}$和$\boldsymbol{B}$总共的时间复杂度为$O(mnk+nk^2)$,更新$\boldsymbol{R}$的时间复杂度为$O(m^2k+mk+mk^2)$;计算$\boldsymbol{X}_3$和$\boldsymbol{X}_4$对应的时间复杂度分别为$O(mk^2+m^2k+m^2+m^2n)$,$O(nl^2+mnl)$,计算$\boldsymbol{D}$和$\boldsymbol{F}$总共的时间复杂度为$O(mnl+ml^2)$,更新$\boldsymbol{C}$的时间复杂度为$O(nl+nl^2)$.因为$k\ll m$,$l\ll n$且m通常都远大于n,假设迭代次数为T,NetRBC总的时间复杂度为$O(T(2mn(k+l)+mn(m+n)+m^2k+nl^2))$.

本文记录了6种对比算法在8个数据集上的运行时间,并报告在图5和图6中.所有算法均基于Matlab2014a(64 bit)实现.在前6个已知癌症亚型的数据集上,实验运行平台配置为:Windows 7,8 GB RAM,Inter(R) Core(TM) i5-4590;在TCGA数据集上的运行平台配置为:CentOS release 6.9,Intel Xeon E5-2678 v3,256 GB RAM.

NetRBC、NetBC和NCIS均是结合了基因网络的双聚类算法,NetRBC相对NetBC和NCIS平均用时少了约10%~20%,但精度提升了至少5%以上.这是因为NetBC和NCIS利用基因网络对基因进行加权,在迭代加权的过程中多次使用$\boldsymbol{P}$,而NetRBC仅在更新$\boldsymbol{R}$时使用$\boldsymbol{P}$.NMTF为基本的三元非负矩阵分解,没有考虑加权基因和网络正则化项,所以时间花费较NetRBC、NetBC和NCIS都小.MSSRCC采用了类似k-means的方法分别优化基因簇和样本簇指示矩阵,所以其速度较快,仅次于k-means.SBC由于需要通过奇异值分解分析原始数据,随着矩阵规模的增加,时间耗费非常大,所以它的时间花费最大.k-means本身时间复杂度较低,

且只对癌症亚型样本进行聚类分析，所以运行时间最小. 上述实验结果表明 NetRBC 不仅能够保持较高的效率，还能比相关算法获得更好的癌症亚型分类效果.

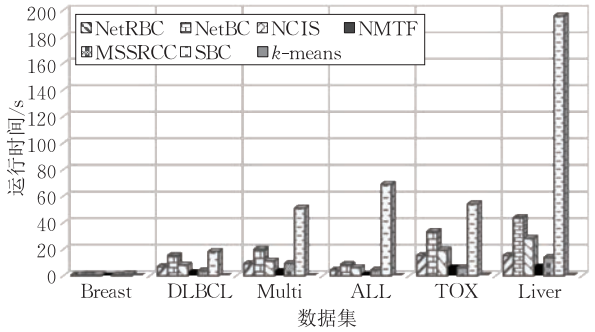

图 5　对比算法在 6 个已知癌症亚型数据集上的运行时间

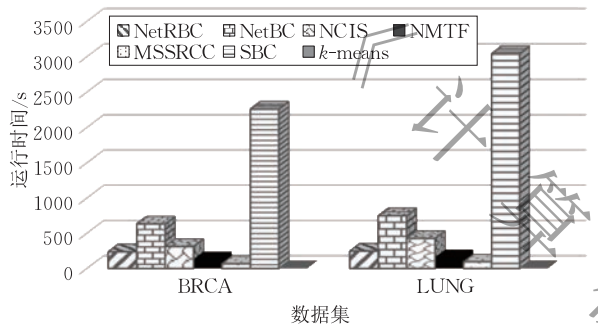

图 6　对比算法在 2 个 TCGA 数据集上的运行时间

5　总结与展望

癌症亚型分类是实现癌症个性化治疗的重要组成部分. 为准确区分癌症亚型，本文提出一种基于基因网络正则化的双聚类算法 NetRBC. 研究发现利用基因互作网络构建正则化项约束指导基因簇指示矩阵的分解，能够有效地利用基因间的间接互作和网络拓扑结构，降低基因互作网络中缺失互作和噪声互作的影响，进而有效提升癌症亚型聚类精度. 癌症的产生和发育过程是多层面因素共同作用的结果，仅利用基因表达谱数据很难全面揭示癌症的高度异质性，因此，如何整合多源数据包括 miRNA 以及 CNV 数据对癌症亚型进行更准确的分类是本文未来工作的重点.

参　考　文　献

[1] Perou C M, Sørlie T, Eisen M B, et al. Molecular portraits of human breast tumours. Nature, 2000, 406(6797): 747-752

[2] Xu Tao-Sheng. Research on Clustering Analysis of Cancer Subtypes Based on Genomics Data[Ph. D. dissertation]. University of Science and Technology of China, Hefei, 2016 (in Chinese)
(许桃胜. 基于基因组数据的癌症亚型发现聚类研究[博士学位论文]. 中国科学技术大学，合肥，2016)

[3] MacQueen J. Some methods for classification and analysis of multivariate observations//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. California, USA, 1967: 281-297

[4] Johnson S C. Hierarchical clustering schemes. Psychometrika, 1967, 32(3): 241-254

[5] Cheng Y, Church G M. Biclustering of expression data//Proceedings of the International Conference on Intelligent Systems for Molecular Biology. California, USA, 2000: 93-103

[6] Barabási A L, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. Nature Reviews Genetics, 2011, 12(1): 56-68

[7] Hwang T H, Atluri G, Xie M Q, et al. Co-clustering phenome—Genome for phenotype classification and disease gene discovery. Nucleic Acids Research, 2012, 40(19): e146

[8] Wang H, Nie F, Huang H, et al. Fast nonnegative matrix tri-factorization for large-scale data co-clustering//Proceedings of the International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 1553-1558

[9] Phillips H S, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell, 2006, 9(3): 157-173

[10] Monti S, Tamayo P, Mesirov J, et al. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning, 2003, 52(1): 91-118

[11] Eisen M B, Spellman P T, Brown P O, et al. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 1998, 95(25): 14863-14868

[12] Shai R, Shi T, Kremen T J, et al. Gene expression profiling identifies molecular subtypes of gliomas. Oncogene, 2003, 22(31): 4918-4923

[13] Steinbach M, Ertöz L, Kumar V. The challenges of clustering high dimensional data. New Directions in Statistical Physics. Springer Berlin Heidelberg, 2004: 273-309

[14] Vesanto J, Alhoniemi E. Clustering of the self-organizing map. IEEE Transactions on Neural Networks, 2000, 11(3): 586-600

[15] Sørlie T, Perou C M, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences, 2001, 98(19): 10869-10874

[16] Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics, 2002, 18(S1): S136-S144

[17] Roy S, Bhattacharyya D K, Kalita J K. CoBi: Attern based co-regulated biclustering of gene expression data. Pattern Recognition Letters, 2013, 34(14): 1669-1678

[18] Kluger Y, Basri R, Chang J T, et al. Spectral biclustering of microarray data: Coclustering genes and conditions. Genome Research, 2003, 13(4): 703-716

[19] Liu B, Wan C, Wang L. An efficient semi-unsupervised gene selection method via spectral biclustering. IEEE Transactions on Nanobioscience, 2006, 5(2): 110-114

[20] Cho H, Dhillon I S. Coclustering of human cancer microarrays using minimum sum-squared residue coclustering. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2008, 5(3): 385-400

[21] Angiulli F, Cesario E, Pizzuti C. Random walk biclustering for microarray data. Information Sciences, 2008, 178(6): 1479-1497

[22] Lee M, Shen H, Huang J Z, et al. Biclustering via sparse singular value decomposition. Biometrics, 2010, 66(4): 1087-1095

[23] Min W, Liu J, Zhang S. L0-norm sparse graph-regularized SVD for biclustering. arXiv preprint, 2016, arXiv: 1603.06035

[24] Pontes B, Giráldez R, Aguilar-Ruiz J S. Biclustering on expression data: A review. Journal of Biomedical Informatics, 2015, 57(C): 163-180

[25] Zhang Y, Xie J, Yang J, et al. QUBIC: A bioconductor package for qualitative biclustering analysis of gene co-expression data. Bioinformatics, 2016, 33(3): 450-452

[26] Bunte K, Leppäaho E, Saarinen I, et al. Sparse group factor analysis for biclustering of multiple data sources. Bioinformatics, 2016, 32(16): 2457-2463

[27] Hanisch D, Zien A, Zimmer R, et al. Co-clustering of biological networks and gene expression data. Bioinformatics, 2002, 18(S1): S145-S154

[28] Nguyen T, Tagett R, Diaz D, et al. A novel approach for data integration and disease subtyping. Genome Research, 2017, 27(12): 2025-2039

[29] Wang B, Mezlini A M, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods, 2014, 11(3): 333-337

[30] Mo Q, Wang S, Seshan V E, Olshen A B, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences, 2013, 110(11): 4245-4250

[31] Liu Y, Gu Q, Hou J P, et al. A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. BMC Bioinformatics, 2014, 15(1): 37

[32] Morrison J L, Breitling R, Higham D J, et al. GeneRank: Using search engine technology for the analysis of microarray experiments. BMC Bioinformatics, 2005, 6(1): 233

[33] Yu G, Yu X, Wang J. Network-aided bi-clustering for discovering cancer subtypes. Scientific Reports, 2017, 7(1): 1046

[34] Li T, Wernersson R, Hansen R B, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nature Methods, 2017, 14(1): 61-64

[35] Chua H N, Sung W K, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein—protein interactions. Bioinformatics, 2006, 22(13): 1623-1630

[36] Gu Q, Zhou J. Co-clustering on manifolds//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 359-368

[37] Chung F R K. Spectral Graph Theory. Washington, USA: American Mathematical Society, 1997

[38] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological), 1977: 1-38

[39] Boyd S, Vandenberghe L. Convex Optimization. Cambridge, UK: Cambridge University Press, 2004

[40] Stark C, Breitkreutz B J, Reguly T, et al. BioGRID: A general repository for interaction datasets. Nucleic Acids Research, 2006, 34(suppl_1): D535-D539

[41] Van't Veer L J, Dai H, Van De Vijver M J, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002, 415(6871): 530-536

[42] Rosenwald A, Wright G, Chan W C, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. New England Journal of Medicine, 2002, 346(25): 1937-1947

[43] Tamayo P, Scanfeld D, Ebert B L, et al. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. Proceedings of the National Academy of Sciences, 2007, 104(14): 5959-5964

[44] Wang S, Huang A. Penalized nonnegative matrix tri-factorization for co-clustering. Expert Systems with Applications, 2017, 78(C): 64-73

[45] Jolly R A, Goldstein K M, Wei T, et al. Pooling samples within microarray studies: A comparative analysis of rat liver transcription response to prototypical toxicants. Physiological Genomics, 2005, 22(3): 346-355

[46] Rand W M. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 1971, 66(336): 846-850

[47] Van Rijsbergen C J. A non-classical logic for information retrieval. The Computer Journal, 1986, 29(6): 481-485

[48] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 2002, 3(11): 583-617

[49] Farris J S. On the cophenetic correlation coefficient. Systematic Zoology, 1969, 18(3): 279-285

[50] Chuang H Y, Lee E, Liu Y T, et al. Network-based classification of breast cancer metastasis. Molecular Systems Biology, 2007, 3(1): 140

[51] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 1995, Series B (Methodological): 289-300

附录 1.　收敛性分析.

参考文献[36]，本文通过辅助函数的方法对 NetRBC 的目标方程式(10)进行收敛性分析，具体如下：

定义 1.　设 $Z(h,h')$ 为 $R(h)$ 的辅助函数，且满足

$$Z(h,h')\geq R(h),Z(h,h)=R(h).$$

引理 1.　假设 Z 是 R 的一个辅助函数，那么 R 在下面的更新公式中是非增的：

$$h^{(t+1)}=\arg\min_{h} Z(h,h^{(t)}).$$

证明.　$R(h^{(t+1)})\leq Z(h^{(t+1)},h^{(t)})\leq Z(h^{(t)},h^{(t)})=R(h^{(t)})$.

引理 2.　对于任意非负矩阵 $\boldsymbol{A}\in\mathbb{R}^{n\times n},\boldsymbol{B}\in\mathbb{R}^{k\times k},\boldsymbol{S}\in\mathbb{R}^{n\times k},\boldsymbol{S}'\in\mathbb{R}^{n\times k},\boldsymbol{A}\in\mathbb{R}^{n\times n}$，并且 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 是对称的，那么有下列不等式：

$$\sum_{i=1}^{n}\sum_{p=1}^{k}\frac{(\boldsymbol{AS}'\boldsymbol{B})_{ip}\boldsymbol{S}_{ip}^{2}}{\boldsymbol{S}_{ip}}\geq \mathrm{tr}(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{ASB}).$$

接下来将证明 4 个定理来保证算法的收敛性.

定理 1.　令

$$\Phi(\boldsymbol{R})=\mathrm{tr}(\lambda\boldsymbol{R}^{\mathrm{T}}\boldsymbol{LR}-2\boldsymbol{AR}^{\mathrm{T}}+\boldsymbol{RBR}^{\mathrm{T}}) \tag{23}$$

则以下函数：

$$Z(\boldsymbol{R},\boldsymbol{R}')=$$

$$\lambda\sum_{ij}\frac{(\boldsymbol{L}^{+}\boldsymbol{R}')_{ij}\boldsymbol{R}_{ij}^{2}}{\boldsymbol{R}'_{ij}}-\lambda\sum_{ijk}(\boldsymbol{L}^{-})_{jk}\boldsymbol{R}'_{ji}\boldsymbol{R}'_{ki}\left(1+\log\frac{\boldsymbol{R}_{ji}\boldsymbol{R}_{ki}}{\boldsymbol{R}'_{ji}\boldsymbol{R}'_{ki}}\right)+$$

$$2\sum_{ij}\boldsymbol{A}_{ij}^{+}\boldsymbol{R}'_{ij}\left(1+\log\frac{\boldsymbol{R}_{ij}}{\boldsymbol{R}'_{ij}}\right)+2\sum_{ij}\boldsymbol{A}_{ij}^{-}\frac{\boldsymbol{R}_{ij}^{2}+\boldsymbol{R}_{ij}'^{2}}{2\boldsymbol{R}'_{ij}}+$$

$$\sum_{ij}\frac{(\boldsymbol{R}'\boldsymbol{B}^{+})_{ij}\boldsymbol{R}_{ij}^{2}}{\boldsymbol{R}'_{ij}}-\sum_{ijk}\boldsymbol{B}_{jk}^{-}\boldsymbol{R}'_{ij}\boldsymbol{R}'_{ik}\left(1+\log\frac{\boldsymbol{R}_{ij}\boldsymbol{R}_{ik}}{\boldsymbol{R}'_{ij}\boldsymbol{R}'_{ik}}\right)$$

是 $\Phi(\boldsymbol{R})$ 的辅助函数，且 $Z(\boldsymbol{R},\boldsymbol{R}')$ 是一个收敛的函数，它的全局最优解是：

$$\boldsymbol{R}_{ij}=\boldsymbol{R}_{ij}\sqrt{\frac{[\lambda\boldsymbol{L}^{-}\boldsymbol{R}+\boldsymbol{A}^{+}+\boldsymbol{RB}^{-}]_{ij}}{[\lambda\boldsymbol{L}^{+}\boldsymbol{R}+\boldsymbol{A}^{-}+\boldsymbol{RB}^{+}]_{ij}}}.$$

证明.　将等式(23)改写成如下形式：

$$L(\boldsymbol{R})=\mathrm{tr}(\lambda\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}^{+}\boldsymbol{R}-\lambda\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}^{-}\boldsymbol{R}-2\boldsymbol{R}^{\mathrm{T}}\boldsymbol{A}^{+}+2\boldsymbol{R}^{\mathrm{T}}\boldsymbol{A}^{-}+\boldsymbol{RB}^{+}\boldsymbol{R}^{\mathrm{T}}-\boldsymbol{RB}^{-}\boldsymbol{R}^{\mathrm{T}}) \tag{24}$$

由引理 2，可以得到

$$\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{L}^{+}\boldsymbol{R})\leq\sum_{ij}\frac{(\boldsymbol{L}^{+}\boldsymbol{R}')_{ij}\boldsymbol{R}_{ij}^{2}}{\boldsymbol{R}'_{ij}},$$

$$\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}^{+}\boldsymbol{R})\leq\sum_{ij}\frac{(\boldsymbol{R}'\boldsymbol{B}^{+})_{ij}\boldsymbol{R}_{ij}^{2}}{\boldsymbol{R}'_{ij}}.$$

通过应用不等式 $a\leq\frac{(a^{2}+b^{2})}{2b},\forall a,b>0$，可以得到

$$\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{A}^{-})=\sum_{ij}\boldsymbol{A}_{ij}^{-}\boldsymbol{R}_{ij}\leq\sum_{ij}\boldsymbol{A}_{ij}^{-}\frac{\boldsymbol{R}_{ij}^{2}+\boldsymbol{R}_{ij}'^{2}}{2\boldsymbol{R}'_{ij}}.$$

为求取其它项的下界，应用不等式 $z\geq1+\log z,\forall z>0$，可得：

$$\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{A}^{+})\geq\sum_{ij}\boldsymbol{A}_{ij}^{+}\boldsymbol{R}'_{ij}\left(1+\log\frac{\boldsymbol{R}_{ij}}{\boldsymbol{R}'_{ij}}\right),$$

$$\mathrm{tr}(\boldsymbol{R}^{\mathrm{T}}\boldsymbol{LR})\geq\sum_{ijk}(\boldsymbol{L}^{-})_{jk}\boldsymbol{R}'_{ji}\boldsymbol{R}'_{ki}\left(1+\log\frac{\boldsymbol{R}_{ji}\boldsymbol{R}_{ki}}{\boldsymbol{R}'_{ji}\boldsymbol{R}'_{ki}}\right),$$

$$\mathrm{tr}(\boldsymbol{RB}^{-}\boldsymbol{R}^{\mathrm{T}})\geq\sum_{ijk}\boldsymbol{B}_{jk}^{-}\boldsymbol{R}'_{ij}\boldsymbol{R}'_{ik}\left(1+\log\frac{\boldsymbol{R}_{ij}\boldsymbol{R}_{ik}}{\boldsymbol{R}'_{ij}\boldsymbol{R}'_{ik}}\right).$$

通过对等式右边求和，可得 $Z(\boldsymbol{R},\boldsymbol{R}')$，该函数显然满足：

$$Z(\boldsymbol{R},\boldsymbol{R}')\geq\Phi(\boldsymbol{R}),Z(\boldsymbol{R},\boldsymbol{R})=\Phi(\boldsymbol{R}).$$

为了找到 $Z(\boldsymbol{R},\boldsymbol{R}')$ 的最小值，对它求一阶导数：

$$\frac{\partial Z(\boldsymbol{R},\boldsymbol{R}')}{\partial\boldsymbol{R}_{ij}}=2\lambda\frac{(\boldsymbol{L}^{+}\boldsymbol{R}')_{ij}\boldsymbol{R}_{ij}}{\boldsymbol{R}'_{ij}}-2\lambda\frac{(\boldsymbol{L}^{-}\boldsymbol{R}')_{ij}\boldsymbol{R}'_{ij}}{\boldsymbol{R}_{ij}}-2\boldsymbol{A}_{ij}^{+}\frac{\boldsymbol{R}'_{ij}}{\boldsymbol{R}_{ij}}+2\boldsymbol{A}_{ij}^{-}\frac{\boldsymbol{R}_{ij}}{\boldsymbol{R}'_{ij}}+2\frac{(\boldsymbol{R}'\boldsymbol{B}^{+})_{ij}\boldsymbol{R}_{ij}}{\boldsymbol{R}'_{ij}}-2\frac{(\boldsymbol{R}'\boldsymbol{B}^{-})_{ij}\boldsymbol{R}'_{ij}}{\boldsymbol{R}_{ij}}$$

以及 $Z(\boldsymbol{R},\boldsymbol{R}')$ 的 Hessian 矩阵：

$$\frac{\partial^{2}Z(\boldsymbol{R},\boldsymbol{R}')}{\partial\boldsymbol{R}_{ij}\partial\boldsymbol{R}_{kl}}=\delta_{ik}\delta_{jl}\left(2\lambda\frac{(\boldsymbol{L}^{+}\boldsymbol{R}')_{ij}}{\boldsymbol{R}'_{ij}}+2\lambda\frac{(\boldsymbol{L}^{-}\boldsymbol{R}')_{ij}\boldsymbol{R}'_{ij}}{\boldsymbol{R}_{ij}^{2}}+2\boldsymbol{A}_{ij}^{+}\frac{\boldsymbol{R}'_{ij}}{\boldsymbol{R}_{ij}^{2}}+2\frac{\boldsymbol{A}_{ij}^{-}}{\boldsymbol{R}'_{ij}}+2\frac{(\boldsymbol{R}'\boldsymbol{B}^{+})_{ij}}{\boldsymbol{R}'_{ij}}-2\frac{(\boldsymbol{R}'\boldsymbol{B}^{-})_{ij}\boldsymbol{R}'_{ij}}{\boldsymbol{R}_{ij}^{2}}\right).$$

得到的 Hessian 矩阵是一个对角矩阵，且对角元素非负. 因此 $Z(\boldsymbol{R},\boldsymbol{R}')$ 是一个凸函数，通过设置 $\frac{\partial Z(\boldsymbol{R},\boldsymbol{R}')}{\partial\boldsymbol{R}_{ij}}=0$，解得

$$\boldsymbol{R}_{ij}=\boldsymbol{R}_{ij}\sqrt{\frac{[\lambda\boldsymbol{L}^{-}\boldsymbol{R}+\boldsymbol{A}^{+}+\boldsymbol{RB}^{-}]_{ij}}{[\lambda\boldsymbol{L}^{+}\boldsymbol{R}+\boldsymbol{A}^{-}+\boldsymbol{RB}^{+}]_{ij}}}.$$

定理 2.　使用式(16)更新 $\boldsymbol{R}$，目标函数(10)单调递减，因此，式(10)是收敛的.

证明.　通过引理 1 和定理 1，可得

$$\Phi(\boldsymbol{R}^{0})=Z(\boldsymbol{R}^{0},\boldsymbol{R}^{0})\geq Z(\boldsymbol{R}^{1},\boldsymbol{R}^{0})\geq\Phi(\boldsymbol{R}^{1})\geq\cdots$$

所以 $\Phi(\boldsymbol{R})$ 是单调递减的，且有下界，收敛性得证.

定理 3.　令

$$\Phi(\boldsymbol{C})=\mathrm{tr}(-2\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}+\boldsymbol{CFC}^{\mathrm{T}}) \tag{25}$$

则以下函数：

$$Z(\boldsymbol{C},\boldsymbol{C}')=-2\sum_{ij}\boldsymbol{D}_{ij}^{+}\boldsymbol{C}'_{ij}\left(1+\log\frac{\boldsymbol{C}_{ij}}{\boldsymbol{C}'_{ij}}\right)+2\sum_{ij}\boldsymbol{D}_{ij}^{-}\frac{\boldsymbol{C}_{ij}^{2}+\boldsymbol{C}_{ij}'^{2}}{2\boldsymbol{C}'_{ij}}+\sum_{ij}\frac{(\boldsymbol{C}'\boldsymbol{F}^{+})_{ij}\boldsymbol{C}_{ij}^{2}}{\boldsymbol{C}'_{ij}}-\sum_{ijk}\boldsymbol{F}_{jk}^{-}\boldsymbol{C}'_{ij}\boldsymbol{C}'_{ik}\left(1+\log\frac{\boldsymbol{C}_{ij}\boldsymbol{C}_{ik}}{\boldsymbol{C}'_{ij}\boldsymbol{C}'_{ik}}\right)$$

是 $\Phi(\boldsymbol{C})$ 的辅助函数，且 $Z(\boldsymbol{C},\boldsymbol{C}')$ 是一个收敛的函数，它的全局最优解是：

$$\boldsymbol{C}_{ij}=\boldsymbol{C}_{ij}\sqrt{\frac{[\boldsymbol{D}^{+}+\boldsymbol{CF}^{-}]_{ij}}{[\boldsymbol{D}^{-}+\boldsymbol{CF}^{+}]_{ij}}}.$$

证明.　将等式(24)改写成如下形式：

$$L(\boldsymbol{C})=\mathrm{tr}(-2\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}^{+}+2\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}^{-}+\boldsymbol{CF}^{+}\boldsymbol{C}^{\mathrm{T}}-\boldsymbol{CF}^{-}\boldsymbol{C}^{\mathrm{T}}) \tag{26}$$

由引理 2，可以得到：

$$\mathrm{tr}(\boldsymbol{CF}^{+}\boldsymbol{C}^{\mathrm{T}})\leq\sum_{ij}\frac{(\boldsymbol{C}'\boldsymbol{F}^{+})_{ij}\boldsymbol{C}_{ij}^{2}}{\boldsymbol{C}'_{ij}}.$$

通过应用不等式 $a\leq\frac{(a^{2}+b^{2})}{2b},\forall a,b>0$，得到：

$$\mathrm{tr}(2\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}^{-})=\sum_{ij}\boldsymbol{D}_{ij}^{-}\boldsymbol{C}_{ij}\leq\sum_{ij}\boldsymbol{D}_{ij}^{-}\frac{\boldsymbol{C}_{ij}^{2}+\boldsymbol{C}_{ij}'^{2}}{2\boldsymbol{C}'_{ij}}.$$

为了得到其它项的下界，应用不等式 $z\geq1+\log z,\forall z>0$ 可得：

$$\mathrm{tr}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}^{+})\geq\sum_{ij}\boldsymbol{D}_{ij}^{+}\boldsymbol{C}'_{ij}\left(1+\log\frac{\boldsymbol{C}_{ij}}{\boldsymbol{C}'_{ij}}\right),$$

$$\mathrm{tr}(\boldsymbol{CF}^{-}\boldsymbol{C}^{\mathrm{T}})\geq\sum_{ijk}\boldsymbol{F}_{jk}^{-}\boldsymbol{C}'_{ij}\boldsymbol{C}'_{ik}\left(1+\log\frac{\boldsymbol{C}_{ij}\boldsymbol{C}_{ik}}{\boldsymbol{C}'_{ij}\boldsymbol{C}'_{ik}}\right).$$

通过对等式右边求和，可得 $Z(\boldsymbol{C},\boldsymbol{C}')$，该函数显然满足：$Z(\boldsymbol{C},\boldsymbol{C}')\geq\Phi(\boldsymbol{C})$，$Z(\boldsymbol{C},\boldsymbol{C})=\Phi(\boldsymbol{C})$.

为了找到 $Z(\boldsymbol{C},\boldsymbol{C}')$的最小值，对它求一阶导数：

$$\frac{\partial Z(\boldsymbol{C},\boldsymbol{C}')}{\partial \boldsymbol{C}_{ij}}=-2\boldsymbol{D}_{ij}^{+}\frac{\boldsymbol{C}'_{ij}}{\boldsymbol{C}_{ij}}+2\boldsymbol{D}_{ij}^{-}\frac{\boldsymbol{C}_{ij}}{\boldsymbol{C}'_{ij}}+2\frac{(\boldsymbol{C}'\boldsymbol{F}^{+})_{ij}\boldsymbol{C}_{ij}}{\boldsymbol{C}'_{ij}}-2\frac{(\boldsymbol{C}'\boldsymbol{F}^{-})_{ij}\boldsymbol{C}'_{ij}}{\boldsymbol{C}_{ij}}$$

以及 $Z(\boldsymbol{C},\boldsymbol{C}')$的 Hessian 矩阵：

$$\frac{\partial^{2} Z(\boldsymbol{C},\boldsymbol{C}')}{\partial \boldsymbol{C}_{ij}\partial \boldsymbol{C}_{kl}}=\delta_{ik}\delta_{jl}\left(2\boldsymbol{D}_{ij}^{+}\frac{\boldsymbol{C}'_{ij}}{\boldsymbol{C}_{ij}^{2}}+2\frac{\boldsymbol{D}_{ij}^{-}}{\boldsymbol{C}'_{ij}}+2\frac{(\boldsymbol{C}'\boldsymbol{F}^{+})_{ij}}{\boldsymbol{C}'_{ij}}-2\frac{(\boldsymbol{C}'\boldsymbol{F}^{-})_{ij}\boldsymbol{C}'_{ij}}{\boldsymbol{C}_{ij}^{2}}\right).$$

得到的 Hessian 矩阵是一个对角矩阵，且对角元素非负，因此 $Z(\boldsymbol{C},\boldsymbol{C}')$是一个凸函数，通过设置$\frac{\partial Z(\boldsymbol{C},\boldsymbol{C}')}{\partial \boldsymbol{C}_{ij}}=0$，解得：

$$\boldsymbol{C}_{ij}=\boldsymbol{C}_{ij}\sqrt{\frac{[\boldsymbol{D}^{+}+\boldsymbol{C}\boldsymbol{F}^{-}]_{ij}}{[\boldsymbol{D}^{-}+\boldsymbol{C}\boldsymbol{F}^{+}]_{ij}}}.$$

定理 4. 使用式(22)更新 $\boldsymbol{C}$，将会使得目标函数式(10)单调递减，因此，式(10)是收敛的.

证明. 通过引理 1 和定理 3，我们可以得到：

$$\Phi(\boldsymbol{C}^{0})=Z(\boldsymbol{C}^{0},\boldsymbol{C}^{0})\geq Z(\boldsymbol{C}^{1},\boldsymbol{C}^{0})\geq\Phi(\boldsymbol{C}^{1})\geq\cdots$$

所以 $\Phi(\boldsymbol{C})$是单调递减的，且有下界，收敛性得证.

由定理 1～4 可以保证 NetRBC 算法的收敛，但是不能保证收敛到全局最优.

WANG Xing, M. S. candidate. His main research interests include machine learning and bioinformatics.

WANG Jun, Ph. D., associate professor. Her main research interests include data mining and bioinformatics.

YU Guo-Xian, Ph. D., professor. His main research interests include machine learning, data mining and bioinformatics.

GUO Mao-Zu, Ph. D., professor. His main research interests include machine learning, data mining and bioinformatics.

Background

Clustering gene expression data is an interdisciplinary study of machine learning and bioinformatics. Gene expression data contain two dimensions (gene and sample). Traditional clustering techniques can only group gene expression data from gene-wise or sample-wise, they can only discover the global pattern of gene expression data. In real world scenarios, similar genes may exhibit similar behaviors only over a subset of samples, or relevant samples exhibit similar expression profiles over a subset of genes. Therefore, bi-clustering, which can simultaneously group genes and samples along two dimensions, is proposed for analyzing gene expression data. For gene clusters, pathway analysis with protein-protein interactions network and gene set enrichment analysis with Gene Ontology can be explored. For sample clusters, the subtype of cancers can be discovered.

In this paper, motivated by the fact that interacting genes are more likely to share similar biological functions and associated with same cancer subtypes, we propose a network regularized bi-clustering method (NetRBC) based on nonnegative matrix factorization. NetRBC defines a regularization term based on the gene interaction networks and uses this regularization to guide the gene-cluster and sample-cluster indicator matrix factorization. Experimental results on various gene expression datasets show that NetRBC surpasses the existing methods and can serve as an effective tool for categorizing cancer subtypes.

The work described in this study is supported by the National Natural Science Foundation of China under Grant (Nos. 61741217, 61402378, 61571163, 61532014, 61873214, 61872300, 61871020, and 61671189), and the Natural Science Foundation of CQ CSTC (Nos. cstc2016jcyjA0351 and cstc2018jcyjA0228).