Journal of Computer Applications
计算机应用,2013,33(8):2188 -2193,2239
ISSN 1001-9081
CODEN JYIIDU
2013-08-01
http://www.joca.cn

文章编号:1001-9081(2013)08-2188-06
doi:10.11772/j.issn.1001-9081.2013.08.2188

相对行常量差异共表达双聚类挖掘算法

谢华博*,尚学群,王　森
(西北工业大学 计算机学院,西安 710129)
(*通信作者电子邮箱 lovetimil@sina.com)

摘　要:在生物信息学上,挖掘差异共表达双聚类有助于研究衰老、癌变类变化的生物过程。以往的差异共表达双聚类定义仅仅从一组基因的角度来衡量差异,导致包含了很多噪声。为了克服上述缺点提出新的差异共表达支持度 MiSupport,可以将一组基因的差异细化到基因级别;并由此定义提出 MiCluster 算法,可以在两个真实的基因芯片数据中挖掘最大的差异共表达双聚类。MiCluster 算法首先基于两个基因芯片数据构建差异共表达权值图,然后基于权值图,采用样本扩展和层次扩展,并利用精确的候选产生方法和高效的剪枝策略,挖掘出最大的差异共表达双聚类。实验结果证明,MiCluster 算法比现有的算法快速高效,而且通过均方误差(MSE)测试和基因本体(GO)评价,挖掘出来结果具有更大的统计意义和生物学意义。

关键词:基因芯片;基因共表达;双聚类;差异;行常量
中图分类号:TP311　　**文献标志码**:A

Differential co-expression relative constant row bicluster mining algorithm

XIE Huabo*, SHANG Xuequn, WANG Miao
(*School of Computer Science, Northwestern Polytechnical University, Xi'an Shaanxi* 710129, *China*)

Abstract: Bioinformaticly, it is useful to study the change process of biology, such as aging and canceration, by mining differential co-expression bicluster. The definition in the past only measured from the perspective of all set of genes, thus containing a lot of noise. Therefore, a new definition named MiSupport was put forward to measure the difference on gene level, and on the basis of MiSupport, an algorithm named MiCluster was proposed to mine the maximal differential co-expression bicluster in two real gene chips. Firstly, MiCluster constructed a differential weighted undirected sample-sample relational graph in two real-valued gene expression datasets. Secondly, the maximal differential biclusters was produced in the above differential weighted undirected sample-sample relational graph with efficiently pruning techniques and accurately generating candidates method by sample-growth and level-growth. The experimental results show that MiCluster is more efficient than the existing methods. Furthermore, the performance is evaluated by Mean Square Error (MSE) score and Gene Ontology (GO). The results show that this algorithm can find better statistical and biological significance.

Key words: gene chip; gene co-expression; bicluster; differential; constant row

0　引言

在生物信息学上认为,疾病常常是由维持细胞健康状态的基因网络及其衍生物的扰动所造成的,而基因芯片技术是大规模研究此类扰动和探究基因作用的最流行的技术之一。基因芯片技术中广泛使用的方法是双聚类。双聚类是在基因表达数据中,识别和一组实验条件相关的共表达的基因组。双聚类一般有以下类型[1]:固定值双聚类[2]、行是常量或者列是常量的双聚类[3]、行与列之间都紧密的双聚类[4]、行与列的变化紧密的双聚类[5]。这些不同类型的双聚类可以从真实数据挖掘出不同意义的重要知识。

差异共表达双聚类方法是基因芯片技术中另外一种流行的方法,它能识别有差异共表达的双聚类,即基因组在一组数据集中有很强的关联关系而在另外一组没有。差异共表达双聚类方法有助于发现和衰老、癌变类变化的生物过程相关的基因。例如,通过比较两个年龄段的基因表达数据可以发现一组和衰老相关的基因。在生物学上,差异共表达双聚类可以预示出错的调控网络[6]。

近年来有很多挖掘差异共表达双聚类的算法。Okada 等[7]采用了两步挖掘方法,先分别在两个基因芯片数据集中产生双聚类,然后把在两个数据集间有差异的双聚类保留下来。DeBi 算法[8]也采用类似的步骤,在单个数据集产生双聚类后采用 MAFIA 算法[9]挖掘具有差异的正调控或负调控模式。由于在每一个数据集都要产生双聚类或者基因模式,而这些双聚类在下一步差异挖掘过程中可能被剪枝,所以两步挖掘方法效率比较低。因此产生了直接从基因芯片数据中挖掘差异共表达的双聚类的方法。DiBiCLUS 算法[10]直接采用聚类的方式从两个数据集中挖掘满足差异共表达的双聚类。Fang 等[11]提出差异支持度概念,由此定义了子空间上的差异表达基因模式,并相应地提出 SDC 算法挖掘此类差异模式。DRCluster[12]算法提出了样本范围支持度,并在此基础上提出了新的行常量差异双聚类,此算法通过基于权值图的回溯扩展,效率较高。以上三个算法都存在不足。如 DiBiCLUS 算法可能会丢失部分信息:一组基因在不同实验条件中可能同时

收稿日期:2013-03-05;**修回日期**:2013-05-02。
基金项目:国家 973 计划项目(2012CB316203);国家自然科学基金资助项目(61272121)。
作者简介:谢华博(1987 -),男,江西于都人,硕士研究生,主要研究方向:生物数据挖掘、差异共表达;　尚学群(1973 -),女,陕西西安人,教授,博士,主要研究方向:数据库、数据挖掘、生物信息学;　王森(1981 -),男,河南义马人,博士,主要研究方向:数据挖掘、生物信息学。

存在正共表达和负共表达,DiBiCLUS 算法只保存样本数最多的那个表达方式;而且 DiBiCLUS 算法对原始数据进行了离散化,也会造成信息的缺失。SDC 算法所挖掘的子空间差异表达模式使用范围支持度来衡量,而不是用基因之间的共表达关系,可能会丢失信息;同时 SDC 算法采用类 Apriori 结构,所以 SDC 算法的不足和 Apriori 类似,效率不高,需要保留候选集,这些缺点都不适用于大规模的基因芯片数据集。DRCluster 算法中差异共表达双聚类定义有点弱,挖掘出的结果差异效果不大明显。

为了更加有效地挖掘,本文提出了一个新的差异共表达双聚类定义 MiSupport 以及一种基于差异权值图的扩展算法 MiCluster,以从两个真实的基因芯片数据集中挖掘出差异共表达双聚类。首先从两个数据集中产生满足定义的差异权值图,这个权值图中包含了在每两个实验条件满足 MiSupport 定义的基因组;然后在差异权值图的基础上,采用样本(实验条件)扩展的方式并使用相应的剪枝策略来挖掘最大的差异共表达双聚类。本文工作主要包括:

1)提出相对样本范围支持度,用以在基因表达值的基础上产生具有相关性样本(实验条件)的集合和衡量基因间的共表达关系。

2)在基因共表达关系的基础上提出新的差异共表达双聚类定义 MiSupport。该定义满足反单调性,采用该定义的算法可以利用 Apriori 性质剪枝。从实验结果中可以得出采用此定义的算法能挖掘出更好的结果。

3)MiCluster 算法在差异权值图的基础上扩展,由于保留了中间结果,大大提高了算法效率。

4)MiCluster 算法采用了有效的剪枝策略,可以有效地一次性挖掘满足定义的最大差异双聚类。

1　问题描述及相关定义

基因芯片数据也被称为基因表达数据,通常被定义为一个一个矩阵,$\boldsymbol{D} = G \times C$,其中:行 G 代表基因;列 C 表示实验条件,也被命名为样本;矩阵中每一个元素 D_{ij} 是基因 i 在实验条件 j 下的真实表达水平值。表 1 和表 2 是两个基因表达矩阵的例子。

表 1　一个真实基因表达数据集 *A*

行	列			
	S_1	S_2	S_3	S_4
G_1	2.11	2.12	0.56	2.13
G_2	-1.68	-4.56	5.63	5.51
G_3	3.78	3.76	5.64	3.75
G_4	2.77	2.79	-4.33	-4.35

表 2　一个真实基因表达数据集 *B*

行	列			
	S_1	S_2	S_3	S_4
G_1	1.90	1.91	3.78	1.92
G_2	0.01	0.78	7.31	7.56
G_3	1.37	1.42	1.91	1.39
G_4	3.34	4.46	-4.99	-4.87

差异共表达双聚类通常表示为 $\boldsymbol{DiffCluster} = G \times C$ 的矩阵,其中:G 代表基因集合,记作 $\boldsymbol{DiffCluster}.Gene$;$C$ 表示样本集合,记作 $\boldsymbol{DiffCluster}.Sample$。$\boldsymbol{DiffCluster}.Gene$ 表现为两部分:一部分是在基因表达数据集 $\boldsymbol{A}$ 中共表达同时在某种定义上有较高的表达值,另一部分在数据集 $\boldsymbol{B}$ 中共表达且有较低表达值;或者一部分在基因表达数据集 $\boldsymbol{B}$ 中共表达同时在某种定义上有较高表达值,而另一部分在数据集 $\boldsymbol{A}$ 中共表达且有较低表达值。为了描述方便,定义一个属性函数 $\Phi(g) = PGene$ 或 $\Phi(g) = NGene$,分别表示 $g(g \in G)$ 来自前一部分和来自后一部分。

在介绍差异共表达定义 MiSupport 之前,首先介绍怎么由真实的基因表达数据来确定基因之间的关系。本文基于 Wang 等[12]提出的样本范围支持度提出相对样本范围支持度。定义如下:

定义 1　样本范围支持度。$\boldsymbol{D}$ 是一个基因表达数据矩阵,α 是用户定义用来衡量样本在基因下关联度的参数。一个基因组 $G = \{g_1, g_2, \cdots, g_k\}$ 的相对样本范围支持度定义为 $RelativeSampleRangeSupport(G) = \sum_{g \in G} srs(g, C)$,其中 $srs(g, C)$ 称为单行常量值,定义如下:

$$srs(g,C) = \begin{cases} \min\limits_{c \in C} |D_{g,c}|, & \forall c \in C, \left(\max\limits_{c \in C} D_{g,c} - \min\limits_{c \in C} D_{g,c}\right) \leqslant \alpha\left(\min\limits_{c \in C} |D_{g,c}|\right) \text{且} \Phi(g) = PGene \\ \max\limits_{c \in C} |D_{g,c}|, & \forall c \in C, \left(\max\limits_{c \in C} D_{g,c} - \min\limits_{c \in C} D_{g,c}\right) \leqslant \alpha\left(\min\limits_{c \in C} |D_{g,c}|\right) \text{且} \Phi(g) = NGene \\ 0, & \text{其他} \end{cases}$$

相对样本范围支持度与样本范围支持度的不同在于根据基因所在的数据集不同,分别取不同的值。当基因属性为 $PGene$ 时,如果满足限制的条件,则取基因表达值的绝对值中的最小值,否则为零;当基因属性为 $NGene$ 时,如果满足限制的条件,则取基因表达值的绝对值中的最大值,否则为零。相对于不同数据集取不同值的意义将在下文中结合差异共表达支持度再介绍。为了表述方便,将某个基因满足定义 1 的实验条件称为该基因的关联样本,将某一组基因满足定义 1 的实验条件称为共表达样本。

采用样本范围支持度,可以在同一组实验条件下产生一组具有相关性的基因。另外还需要对这一组基因的相关性进行定义。这里使用 Wang 等[12]的定义,来确定两个基因之间的共表达关系。其定义如下:

定义 2　两个基因间的关系。g_1 和 g_2 是基因表达数据矩阵 $\boldsymbol{D}$ 中任意两个基因,C_1 和 C_2 是 $\boldsymbol{D}$ 中的任意两个实验条件,g_1 和 g_2 在 C_1 和 C_2 下的共表达关系有如下三种情况:

1)如果$\left[\forall g \in (g_1, g_2) \mid \left(\max\limits_{c \in (C_1, C_2)} D_{g,c} - \min\limits_{c \in (C_1, C_2)} D_{g,c}\right) \leqslant \alpha\left(\min\limits_{c \in (C_1, C_2)} |D_{g,c}|\right)\right]$且$\left[\forall c \in (C_1, C_2) \mid (D_{g_1,c} \times D_{g_2,c} > 0)\right]$,那么 g_1 和 g_2 是正共表达关系;

2)如果$\left[\forall g \in (g_1, g_2) \mid \left(\max\limits_{c \in (C_1, C_2)} D_{g,c} - \min\limits_{c \in (C_1, C_2)} D_{g,c}\right) \leqslant \alpha\left(\min\limits_{c \in (C_1, C_2)} |D_{g,c}|\right)\right]$且$\left[\forall c \in (C_1, C_2) \mid (D_{g_1,c} \times D_{g_2,c} < 0)\right]$,那么个 g_1 和 g_2 是负共表达;

3)不满足上述两种情况时,g_1 和 g_2 不共表达。

有了真实基因表达数据下基因表达之间的关系,可以定义新的差异共表达支持度 MiSupport,定义如下:

定义 3　MiSupport。G 是一组基因,C 是一组样本,G 的差异共表达支持度被定义为:

$$MiSupport(P,C) = \min_{g\in p}\left(srs_{\Phi(g)=PGene}(g,C) - srs_{\Phi(g)=NGene}(g,C) \right)$$

同时在样本集合 C 下，基因组 P 分别在两个数据集中共表达。其中 $srs_{\Phi(g)=PGene}(g,C)$，$srs_{\Phi(g)=NGene}(g,C)$ 分别是基因 g 在两个基因表达数据集中前一个数据中单行常量值和在后一个数据集中的单行常量值。

在定义 MiSupport 后，可以说明相对样本范围支持度比样本范围支持度更有优势。如图 1、2 所示，MiSupport(new) 表示相对样本范围支持度定义的 MiSupport，而 MiSupport(old) 表示采用样本范围支持度定义的 MiSupport。从两个图中可以看出，两个 MiSupport 的区别是新的定义用一组共表达基因的最小值和最大值比较，而旧定义用一组共表达基因的最小值和最小值比较。在 α 较小时，MiSupport(new) 和 MiSupport(old) 都可以正确识别；但是在 α 较大时，本来没有差异的两个双聚类，差异度衡量 MiSupport(old) 表现为正，被挖掘出来，而 MiSupport(new) 表现为负，被摒弃掉。

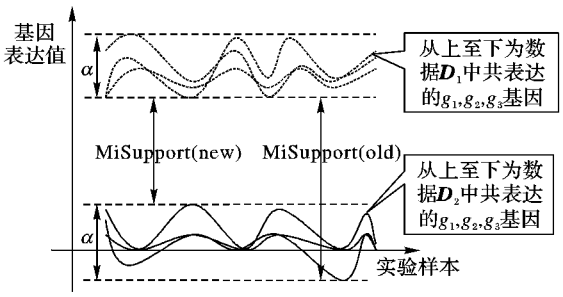


图 1 当 α 较小时采用两种定义的 MiSupport

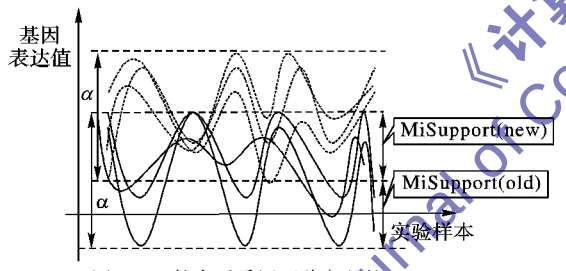


图 2 α 较大时采用两种定义的 MiSupport

根据图 3 可以阐述 MiSupport 和 SDC 定义的区别。由图可知，当 α 较大时，采用范围支持度定义计算 SDC，本来没有差异的两组基因，由于后者计算的是一组基因和之间的差异，基因组的表达值被一个异常高的基因表达值拔高了和值，导致基因组 g_1,g_2,g_3 在数据集 D_1 中表达水平比 D_2 明显高而误挖掘。MiSupport 定义差异细化到基因级别，用每一个基因差异的最小值来衡量整个基因组的差异，所以图 3 中基因组的差异明显小于零，因而未被识别出来。

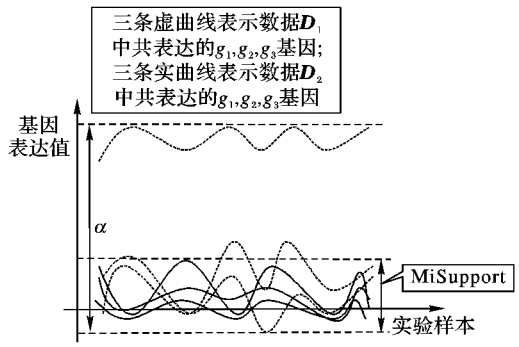


图 3 当 α 较大时 MiSupport 和 SDC 的比较

2 MiCluster 算法

2.1 差异权值图

为了从两个真实的基因表达数据中快速有效地挖掘出最大差异双聚类，MiCluster 算法使用基于样本的差异权值图(Differential Weight Graph，DWG)提高效率，权值图定义如下：

定义 4 差异样本权值图。差异样本权值图 DWG 可用集合 $DWG = \{E,S,W\}$ 来表示，其中每个顶点 S_i 表示为一个样本，如果顶点 S_i 和 S_j 之间存在满足定义 3 的差异共表达基因集合，那么这两个顶点之间就有一条边 E_{ij}，该条边上的权值 W_{ij} 是 S_i 和 S_j 之间的差异共表达基因集合。

算法采用差异权值图后，效率大大提高，主要由以下几点体现：

1)不保留候选集，而采取有结构的差异权值图，这相当于保存了中间结果，而且接下来的扩展可以直接从权值图开始。这个举措省去了每次重新计算的时间。

2)由于基因表达数据的特点，基因个数大大超过样本个数，所以采取构建基于样本的差异权值图，这将意味在回溯扩展时，也是采取样本扩展。由于扩展的节点少，大大提高效率。

3)差异权值图的信息含量丰富，权值中包含从 A 到 B 中求差异，也包含从 B 到 A 中求差异，这样可以避免二次挖掘。为了表述方便，在某组共表达样本下，基因组 G 在数据集 A 的表达水平比数据集 B 高，把这一组基因表示为 $a(G)$；类似地，如果基因组 G 在数据集 B 的表达水平比数据集 A 高，把这一组基因表示为 $b(G)$。由此一个 MiCluster 可以表示为 $S_i\cdots S_{j-1}S_j(a(G_S\cdots G_t),b(G_x\cdots G_y))$。图 4 是一个由表 1 和表 2 构建的 DWG，其中 $\alpha = 0.1$，MiSupport 最小阈值为 0.15。

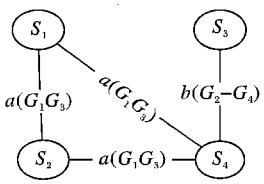


图 4 由表 1 和表 2 构建的 DWG

2.2 挖掘最大的差异共表达双聚类

本节介绍如何从真实的数据集中挖掘出最大的 MiCluster。由于基因表达数据中实验条件的个数远少于基因的个数，所以将在差异权值图的基础上采用样本扩展和层次扩展方式。在阐述 MiCluster 算法之前，先介绍 MiSupport 的反单调性。

定理 1 MiSupport 支持度是满足反单调性的。

证明 设 C 是一组样本集合，C' 是另外一组样本集合，其中 $C' = C \cup c, c \notin C$。对于一组基因 G，g 是 G 中任意一基因，$MRSupport(P,C) = \min_{g\in p}\left(srs_{\Phi(g)=PGene}(g,C) - srs_{\Phi(g)=NGene}(g,C) \right)$，其中 $srs_{\Phi(g)=PGene}(g,C)$ 分为两种情况：

1) 当 $\left(\max_{c\in C} D_{g,c} - \min_{c\in C} D_{g,c} \right) > \alpha \left(\min_{c\in C} |D_{g,c}| \right)$ 时，$srs_{\Phi(g)=PGene}(g,C') = 0$，并且 $srs_{\Phi(g)=PGene}(g,C) = 0$，因此 $srs_{\Phi(g)=PGene}(g,C') \leqslant srs_{\Phi(g)=PGene}(g,C)$；

2) 当 $\left(\max_{c\in C} D_{g,c} - \min_{c\in C} D_{g,c} \right) \leqslant \alpha \left(\min_{c\in C} |D_{g,c}| \right)$ 时，$srs_{\Phi(g)=PGene}(g,C') = \min_{c\in C'} |D_{g,c}| \leqslant \min_{c\in C} |D_{g,c}| = srs_{\Phi(g)=PGene}(g,C)$。

由上述两种情况可得$srs_{\Phi(g)=PGene}(g,C') \leqslant srs_{\Phi(g)=PGene}(g,C)$;同理可证$srs_{\Phi(g)=PGene}(g,C') \geqslant srs_{\Phi(g)=PGene}(g,C)$。

综上可得$\min_{g\in p}\left(srs_{\Phi(g)=PGene}(g,C)-srs_{\Phi(g)=NGene}(g,C)\right) \leqslant \min_{g\in p}\left(srs_{\Phi(g)=PGene}(g,C')-srs_{\Phi(g)=NGene}(g,C')\right)$,即 $MRSupport(P,C) \leqslant MRSupport(P,C')$,定理 1 得证。

上述定理可以保证 MiCluster 算法在采用样本扩展方式挖掘最大的差异双聚类时满足反单调性。这样算法可以使用 Apriori 性质对算法进行剪枝设计以提高效率。为了进一步提高效率,MiCluster 算法规定了候选产生方法,方法定义如下:

定义 5　候选产生方法。设差异共表达双聚类为 $S_i\cdots S_{j-1}S_j(a(G_S\cdots G_t),b(G_x\cdots G_y))$,$MinGene$ 是双聚类中最小基因个数限制,S 是数据集所有样本集合,样本 $S_c \in S$ 是其候选,S_c 必须满足 $|S_i\cdots S_{j-1}S_c.Genes \cap S_i\cdots S_{j-1}S_j.Genes| \geqslant MinGene$。

上述定义能较为精确地产生候选,大大地减少回溯扩展的分支以提高效率。同时,为了高效地获得候选,MiCluster 算法采用层次扩展,这样候选直接从上一层的节点中计算得到,而且层次扩展不和样本扩展冲突。除了上述处理之外,MiCluster 算法还采用了常规的剪枝策略,定理如下:

定理 2　假设 P 是差异共表达双聚类,M 是 P 的候选样本集合,N 是 P 的前驱候选样本集合。若候选样本 $M_i(M_i \in M)$,存在一个前驱候选样本 $N_j(N_j \in N)$ 满足 $PM_i.Gene \subseteq PN_i.Gene$,那么 $PM_iN_i.Gene = PM_i.Gene$,节点 M_i 可以剪枝。

证明　由 MiSupport 的定义可得:$PN_iM_i.Gene = PN_i.Gene \cap PM_i.Gene \cap M_iN_i.Gene$,因为 $PM_i.Gene \subseteq PN_i.Gene$,可得 $PN_iM_i.Gene = PM_i.Gene \cap M_iN_i.Gene$。从集合的观点出发,由 $PM_i.Gene \subseteq PN_i.Gene$ 可得 $M_i.Gene \subseteq N_i.Gene$,所以 $M_iN_i.Gene = M_i.Gene$,最终可得 $PN_iM_i.Gene = PM_i.Gene$,上述定理得证。

显然,上述剪枝定理对于一个差异共表达双聚类 $S_i\cdots S_{j-1}S_j(a(G_S\cdots G_t),b(G_x\cdots G_y))$ 中的基因集合 $a(G_S\cdots G_t)$,$b(G_x\cdots G_y)$ 都适用,为了节省篇幅,不分开阐述。同时把上述剪枝统称为 Pruning。在解决了扩展方式、候选产生方法和剪枝策略后就能给出 MiCluster 算法了。

2.3　MiCluster 算法流程

输入　两个真实的基因芯片数据 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$,差异双聚类中基因最小个数 $MinGene$,差异双聚类中最小的样本个数 $MinSample$,相对行常量共表达阈值 α,最小差异阈值 $MinDiff$,差异表达样本权值图 DWG,当前扩展的差异共表达双聚类 P;

输出　满足定义 MiSupport 的最大差异共表达双聚类集合。

```
BEGIN
  P = NULL; DWG = NULL;
  MiCluster(D1,D2,MinGene,MinSample,α,MinDiff, DWG,P)
  If DWG = NULL then
    构造差异表达样本权值图 DWG;
  end if
  扫描 DWG 生成 P 的候选样本集合 S;
                              //输出符合要求的差异双聚类
  if S = NULL 和 P.Sample 满足个数不小于 MinSample 和
    P.Gene 满足个数不小于 MinGene
    输出 P
  end if                      //对每一个候选,进行挖掘
  for S 集合中每一个 Si
    if PSi 符合剪枝策略 Pruning 和 PSi 的 MiSupport 不小于
      MinDiff, then
      P.Sample = P.Sample ∩ Si, P.Gene = P.Gene ∩ PSi.Gene;
      MiCluster(D1,D2,MinGene,MinSample,α,MinDiff,DWG,P)
    end if
  end for
END
```

为了节省篇幅,算法流程并没有把 $P.Gene$ 区分为 $P(a(G),b(G))$ 来阐述,但在算法具体实现时这一点是必须的。为了让描述更清楚,使用此算法在图 4 上的 DWG 来展示挖掘过程,如图 5 所示($MinGene = 2$,$MinSample = 2$,$\alpha = 0.1$,$MinDiff = 0.15$)。

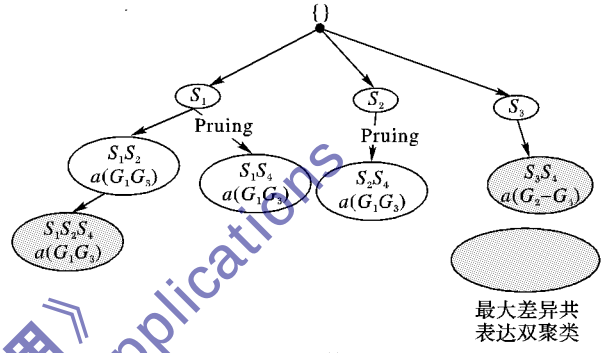


图 5　MiCluster 扩展过程示

2.4　复杂度分析

为了描述方便,令 F 表示最大基因个数,H 表示最大的样本个数。算法主要分两个部分:一部分是构建差异权值图,另一部分是在权值图上回溯搜索。构建权值图部分,由于每两个样本构建一条边,则可知边数最大为 C_H^2,每条边的内容为符合 MiSupport 定义基因组,则综合可知在构建权值图的最大时间复杂度为 $O(F*C_H^2)$。权值图最多有 C_H^2 条边,每条边所占的空间和基因的个数成正比,可得空间复杂度也为 $O(F*C_H^2)$。在权值图上回溯搜索这一部分,整个回溯树的节点个数乘以每个节点时间复杂度为此部分的时间复杂度。回溯树的最大节点数为 2^H,回溯树中节点可以从权值图查找后计算得出,查找复杂度最大为 H,容易得最大时间复杂度为 $O(H*2^H)$。空间复杂度为整个回溯树最大深度乘以基因个数,即为 $O(H*F)$。综合整个算法,最大时间复杂度 $O(F*C_H^2+H*2^H))$,最大空间复杂度 $O(F*C_H^2+H*F)$。由于基因表达数据的特点实验样本个数(一般不会超过 20)远小于基因个数,同时采取有效的 Apriori 和 Pruing 剪枝策略和精确的候选产生方法,使回溯树的节点数远远少于最大节点数,算法效率较好。若不基于权值图扩展,算法耗时则只体现为回溯搜索部分:时间复杂度为整个回溯树的节点数乘以每一个节点时间复杂度,节点数和基于权值图扩展算法一样为 2^H,但是每一个节点的都需要对所有基因进行计算,时间复杂度为 $O(F)$,由此可得整个算法时间复杂度为 $O(F*2^H))$。由于基因表达数据特点,基因个数远大于样本个数,算法效率低。从下文的实验结果也可以看出,采用权值图扩展的算法 MiCluster 运行时间比其他算法少 1 ~ 3 个数量级。

3　实验分析

本章将 MiCluster 算法与 SDC 算法和 DRCluster 算法进行比较。其中 SDC 算法采取原始的 Apriori 算法来挖掘差异共表达模式,它首先用范围支持度来产生符合行常量的双聚类,

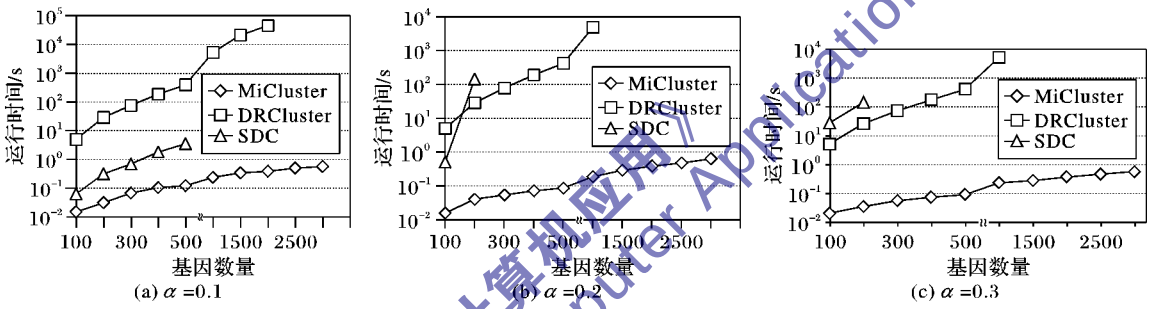
然后采用类 Apriori 算法利用基因扩展产生符合差异范围支持度的差异共表达双聚类，最后筛选出其中最大的差异共表达双聚类。DRcluster 算法首先构建权值图，然后采用样本扩展的方式挖掘出最大 DRcluster。

实验数据采用的是 AGEMAP 基因表达数据库，该数据库是关于研究与衰老有关的基因芯片数据库。AGEMAP 数据库中包含来自 16 个组织中的 8932 个基因。每个组织分别来自年龄为 1、6、16、24 个月的 5 只小鼠。由于数据缺失和噪声，选取了小鼠在 6 个月和 16 个月两个年龄段的基因表达数据。为了挖掘潜在和衰老相关的共表达基因，选取了在 AGEMAP 中标号为 c 的一只雄性老鼠和一只雌性老鼠作为实验数据。实验的硬件环境是 Pentium E5800、内存 2 GB 的电脑，软件运行环境为 Windows 7，算法实现和编译软件为 VC++6.0。

3.1 效率比较

为了让实验结果更有说服力，从 AGEMAP 中选取了不同规模的数据集，选取的方式是依据基因在 AGEMAP 中出现的次序。SDC算法中差异范围支持度设置越高，运行时间越短，结果越好，因此将 SDC 差异范围支持度设置为 1；DRCluster 算法不需要设置差异支持度；本文的 MiSupport 差异支持度阈值设置为 0.15。三个算法都有使用范围支持度或此范围支持度的变种，实验中这一参数将设置成一致。三个算法的最少基因个数阈值和最少样本个数阈值都设置为 2。

从图 6 中容易看到，MiCluster 算法比其他两个算法快 1～3个数量级。从图中可以看出，SDC 算法是基于 Apriori 框架来扩展的，其运行时间和内存占用都随基因个数增加而呈指数级增长。当 $\alpha = 0.1$ 时，SDC 算法在基因规模为 1000 时内存耗尽而崩溃，当 α 为 0.1 和 0.2 时，SDC 算法只能运行前 2 个较小的数据集，后面都因为内存耗尽导致程序无法运行。DRCluster 算法采用回溯扩展，对内存要求不高，但 DRCluster 算法输出结果较多，效率并不是很高。DRCluster 算法在运行基因规模为 2500 和 3000 的数据时由于运行时间超过 12 h 而放弃运行。MiCluster 算法效率较高，除了上文分析的原因外，也是因为 MiSupport 定义过于严格，导致算法可回溯扩展的节点少，从而使算法运行较快。



(a) α =0.1　(b) α =0.2　(c) α =0.3

图 6 算法效率比较

3.2 统计学意义评价

本节将比较实验结果的质量。一般使用均方误差（Mean Squared Error，MSE）[13] 来衡量双聚类的差异度。均方误差是用于衡量一组基因在某一组时间点下的关联度，均方误差得分越小，说明基因表达差异度越小，即具有高关联度。假设 I 和 J 分别是数据集中基因和时间点的集合，$\boldsymbol{D}_{ij}$ 是第 i 个基因在第 j 个时间点下的真实表达值。那么这组基因在所有时间点下的均方误差得分可由下面公式进行计算：

$$M(I,J) = \frac{1}{|I||J|}\sum_{i\in I,j\in J}(\boldsymbol{D}_{ij} - \boldsymbol{D}_{iJ} - \boldsymbol{D}_{Ij} + \boldsymbol{D}_{IJ})^2$$

其中：$\boldsymbol{D}_{iJ} = \frac{1}{|J|}\sum_{j\in J}\boldsymbol{D}_{ij}$ 和 $\boldsymbol{D}_{Ij} = \frac{1}{|I|}\sum_{i\in I}\boldsymbol{D}_{ij}$ 分别是第 i 行和第 j 列的均值，$\boldsymbol{D}_{IJ} = \frac{1}{|I||J|}\sum_{i\in I,j\in J}\boldsymbol{D}_{ij}$ 是这组基因在所有时间点下表达值的均值。

本文实验的目的是识别在 6 个月与 16 个月之间和衰老相关的差异双聚类，但并不是所有在 AGEMAP 中的 cDNA 都和衰老相关。在这里采用了文献[14]收集的在多个小鼠组织中和衰老相关的 305 个基因。在接下来的实验中，所有算法都将在这 305 个基因数据上进行。为了产生的结果个数最为接近，将 SDC 差异范围支持度设置为 1，最小基因个数阈值为 3，最小样本个数阈值为 4；DRCluster 算法不需要设置差异支持度，最小基因个数阈值为 7，最小样本个数阈值为 7；MiSupport 差异支持度设置为 0.15，最小基因个数阈值为 2，最小样本个数阈值为 2。

从图 7～9 容易看出，除了极个别结果外，DRCluster 算法结果 MSE 最大，SDC 算法结果 MSE 介于两者之间，MiCluster 算法结果的 MSE 最小。这说明 MiCluster 算法产生结果最紧凑。就同一算法而言，挖掘结果在 6 个月数据中的 MSE 值比在 16 个月数据中的 MSE 值要大。总的分析，由于三个算法都基于范围支持度来产生双聚类，所以三个算法的 MSE 值都极小，最大不超过 0.08；MiCluster 算法结果 MSE 值最小，和算法结果的规模较少也有关。

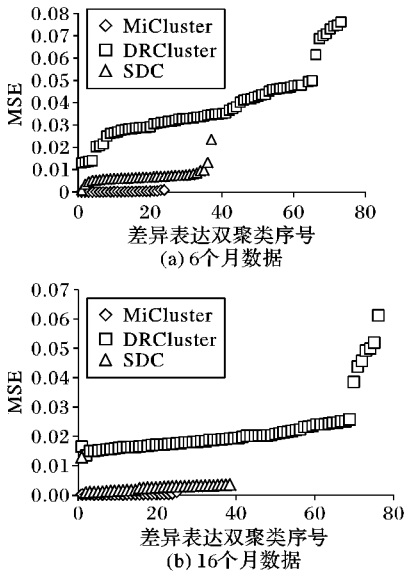
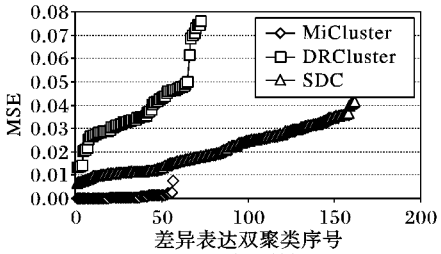


图 7 α = 0.1 时差异共表达的双聚类 MSE 值分布

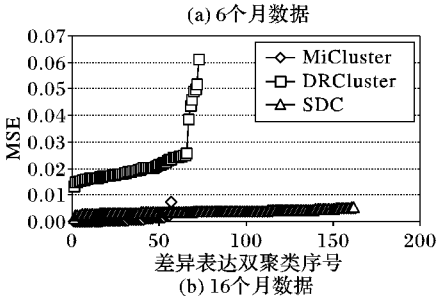
3.3 生物学意义评价

为了评价算法挖掘结果的生物学意义，使用 GO（Gene

Ontology)[15]来评价实验结果。在 GO 数据库中,对每一个生命载体,例如基因、蛋白质或者 cDNA 等,都有一个或一组 GO 类别。在这里使用 GO 识别比值来评价算法的挖掘结果,即一组基因模式被 GO 中已知功能的某个 GO 类别识别的比例(基因同源率)大于某个阈值的比值。同样,为了达到最好的实验效果,三个算法都采用上文中和衰老相关的 305 个基因,三个算法参数设置和上节的统计意义评价一致。

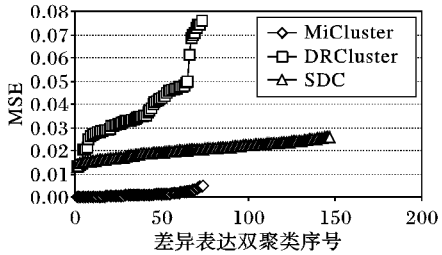


(a) 6个月数据

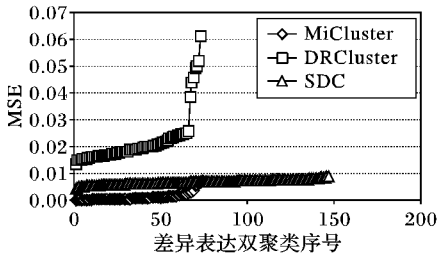


(b) 16个月数据

图 8　α = 0.2 时差异共表达的双聚类 MSE 值分布



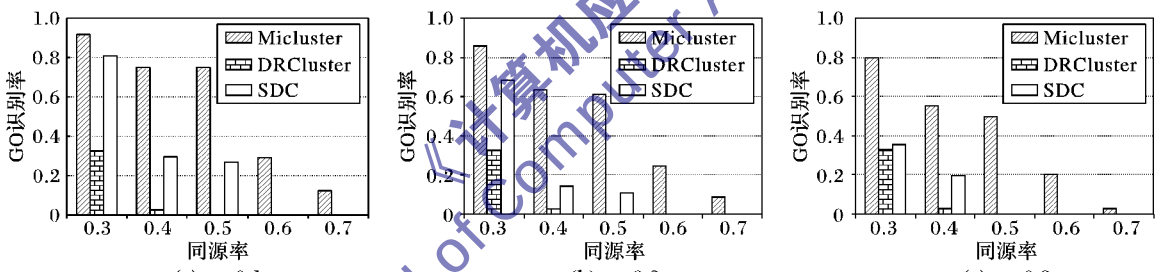
(a) 6个月数据



(b) 16个月数据

图 9　α = 0.3 时差异共表达的双聚类 MSE 值分布

从图 10 可以看出,在 α 不同的情况下,MiCluster 算法结果在不同 GO 同源率情况下的识别率最高,SDC 算法次之,DRCluster 最低。在 GO 同源率阈值为 0.6 和 0.7 时,SDC 算法和 DRCluster 算法结果不被识别。上述说明 MiCluster 算法产生结果中的基因参与了更多的生物过程,有更高的生物学意义评价。



(a) α=0.1　(b) α=0.2　(c) α=0.3

图 10　α 取不同值时的 GO 识别率比较

4　结语

本文提出了在两个真实的基因芯片数据中挖掘最大的差异共表达双聚类算法 MiCluster 算法。该算法使用相对样本范围支持度来产生共表达双聚类,使用 MiSupport 定义来生成差异共表达双聚类。MiCluster 算法首先基于两个基因芯片数据构建差异共表达权值图,然后基于权值图,采用样本扩展和层次扩展,并利用精确的候选产生方法和高效的剪枝策略,快速高效地挖掘出最大的差异共表达双聚类。但是,从定义和结果中可以发现,由于 MiSupport 定义的严格,使挖掘结果挖掘的结果很少并且结果碎片化,同时差异双聚类整体的 GO 识别率并不高。下一步的研究想从生物模型中定义数据模型,可以将碎片化的结果整合为一个较大并且更具有生物学意义的结果。

参考文献:

[1] MADEIRA S C, OLIVEIRA A L. Biclustering algorithms for biological data analysis: a survey [C]// IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2004, 1(1):24 – 45.

[2] HARTIGAN J A. Direct clustering of a data matrix [J]. Journal of the American Statistical Association, 1972, 67(337):123 – 129.

[3] GETZ G, LEVINE E, DOMANY E. Coupled two-way clustering analysis of gene microarray data [J]. Proceedings of the Natural Academy of Sciences of United States of America, 2000, 97(22): 12079 – 12084.

[4] CORMEN T H, ELEISERSON C E, RIVEST R L, *et al*. Introduction to algorithms [M]. 2nd ed. Cambridge: MIT Press, 2001.

[5] LAZZERONI L, OWEN A. Plaid models for gene expression data [J]. Statistica Sinica, 2002, 12: 61 – 86.

[6] KOSTKA D, SPANG R. Finding disease specific alterations in the coexpression of genes [J]. Bioinformatics, 2004, 20(Suppl. 1): i194 – i199.

[7] OKADA Y, INOUE Y. Identification of differentially expressed gene modules between two-class DNA microarray data [J]. Bioinformation, 2009, 4(4): 134 – 137.

[8] SERIN A, VINGRON M. DeBi: discovering differentially expressed biclusters using a frequent itemset approach [J]. Algorithms for Molecular Biology, 2011, 6(1):18.

[9] BURDICK D, CALIMLIM M, GEHRKE J. MAFIA: a maximal frequent itemset algorithm for transactional databases [C]// Proceedings of the 17th International Conference on Data Engineering. Piscataway: IEEE, 2001: 443 – 452.

(下转第 2239 页)

以上实验结果不仅对推理出的参量估算公式进行了验证,而且也验证了水印长度、量化步长和 PSNR 之间的关系,从图 2 ~4 以及表 1 可以看出:在水印长度一定时,PSNR 随着量化步长取值增加而减小;在量化步长取值一定时,PSNR 随着水印序列长度增加而减小;在 PSNR 一定时,量化步长的取值随着水印序列长度增加而减小。

表 1　不同长度水印嵌在不同小波系数时 PSNR 实验测得值(EV)和理论估算值(TV)　　dB

量化步长	db1(L=2048)		sym1(L=4096)		bior1.1(L=8192)		rbio1.1(L=16384)	
	EV	TV	EV	TV	EV	TV	EV	TV
20	47.96	47.95	45.00	44.94	41.96	41.93	38.93	38.92
30	44.43	44.43	41.40	41.42	38.37	38.41	35.40	35.40
40	42.01	41.93	38.96	38.92	35.95	35.91	32.94	32.90
50	40.25	39.99	37.05	36.98	34.01	33.97	30.98	30.96
60	38.24	38.41	35.32	35.40	32.34	32.39	29.34	29.38
70	37.13	37.07	34.01	34.06	31.01	31.05	28.04	28.04
80	35.98	35.91	33.02	32.90	29.97	29.89	26.95	26.88
90	34.96	34.89	31.96	31.88	29.00	28.87	26.01	25.86
100	34.07	33.97	31.02	30.96	27.96	27.95	24.90	24.94

4　结语

量化调制方法因其嵌入信息多、计算复杂度低,在数字水印和信息隐藏领域得到了广泛应用。但是无论是隐藏信息还是嵌入水印都要保证嵌入后载体的视觉质量,量化调制方法中影响载体视觉质量的参数有量化步长、水印数据量和量化系数,水印数据量和量化系数往往是事先确定的,而量化步长对应于嵌入深度,是可以调节的。目前确定量化步长值的方法是通过反复实验,在实际应用中进行大批量嵌入时效率很低,而且很难找到平衡鲁棒性和不可感知性的最佳步长值。针对这个问题本文以最常用的抖动量化调制方法为研究对象,用小波系数作为水印载体,给出了含水印图像的 PSNR、量化步长和水印数据量之间的定量关系式。在水印数据量固定时,根据嵌入后载体的 PSNR 要求,可以直接计算出最佳量化步长值,无需反复实验。虽然仅以奇偶量化调制方法为例进行研究,对于其他量化调制方法,只要明确量化误差的分布区间,可以很容易推导出基于 PSNR 的小波域量化调制方法参数定量关系公式。推导的量化参数之间的定量估算公式仅适用于小波域系数作为量化系数的情况,实际上量化调制算法应用范围很广,离散余弦域系数、离散傅里叶域系数、空域像素值常常用作量化系数,然而推导的定量关系公式不适用于这些情况,下一步的工作是针对这些量化系数对量化步长与视觉质量之间的定量关系进行研究。

参考文献:

[1] 凌洁, 刘琚, 孙建德, 等. 基于视觉模型的迭代 AQIM 水印算法[J]. 电子学报, 2010, 38(1): 151 - 155.

[2] 邓艺, 赵险峰, 冯登国. 基于非均匀 DCT 的量化索引调制隐写[J]. 电子与信息学报, 2010, 32(2): 323 - 328.

[3] 苗锡奎, 孙劲光, 张语涵. 分形与伪 Zernike 矩的鲁棒水印算法研究[J]. 计算机应用, 2010, 30(4): 1038 - 1041.

[4] CHEN B, WORNELL G W. Provably robust digital watermarking [C]// Proceedings of SPIE: Multimedia Systems and Applications II, SPIE 3845. Bellingham: SPIE, 1999: 43 - 54.

[5] CHEN B, WORNELL G W. Quantization index modulation: a class of provably good methods for digital watermarking and information embedding [J]. IEEE Transactions on Information Theory, 2001, 47(4): 1423 - 1443.

[6] CHEN B, WORNELL G W. Quantization index modulation methods for digital watermarking and information embedding of multimedia [J]. Journal of VLSI Signal Processing Systems, 2001, 27(1): 7 - 33.

[7] 肖俊, 王颖. 扩展变换抖动调制水印算法中投影向量的研究[J]. 中国图象图形学报, 2006, 11(12): 1799 - 1805.

[8] 肖俊, 王颖, 李象霖. 带失真补偿的抖动调制水印算法中的补偿因子研究[J]. 电子学报, 2007,35(4): 786 - 790.

[9] 李雷达, 郭宝龙, 表金峰. 基于奇偶量化的空域抗几何攻击图像水印算法[J]. 电子与信息学报, 2009,31(1):134 - 138.

[10] 王宏霞, 何晨, 丁科. 基于混沌映射的鲁棒性公开水印[J]. 软件学报, 2004, 18(8): 1245 - 1251.

[11] 叶天语. DWT-SVD 域全盲自嵌入鲁棒量化水印算法[J]. 中国图象图形学报, 2012, 17(6): 644 - 650.

[12] KIM H D, LEE J W, OH T W, *et al*. Robust DT-CWT watermarking for DIBR 3D images [J]. IEEE Transactions on Broadcasting, 2012, 58(4): 533 - 543.

[13] 肖筱南. 新编概率论与数理统计[M]. 北京:北京大学出版社, 2008: 141 - 142.

[14] TSAI M J, YU K Y, CHEN Y Z. Joint wavelet and spatial transformation for digital watermarking [J]. IEEE Transactions on Consumer Electronics, 2000, 46(1): 241 - 245

[15] 飞思科技产品研发中心. 小波分析理论与 Matlab 7 实现[M]. 北京: 电子工业出版社, 2005: 33 - 37.

(上接第 2193 页)

[10] ODIBAT O, REDDY C K, GIROUX C N. Differential biclustering for gene expression analysis [C]// Proceedings of the ACM Conference on Bioinformatics and Computational Biology. New York: ACM, 2010: 275 - 284.

[11] FANG G, KUANG R, PANDEY G, *et al*. Subspace differential coexpression analysis: problem definition and a general approach [C]// Proceedings of the 15th Pacific Symposium on Biocomputing. Singapore: World Scientific Publishing, 2010: 145 - 156.

[12] WANG M, SHANG X Q, LI X Y, *et al*. Efficient mining differential co-expression constant row bicluster in real-valued gene expression datasets [J]. Applied Mathematics & Information Sciences, 2013, 7(2): 587 - 598.

[13] CHENG Y, CHURCH G M. Biclustering of expression data [C]// Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology. [S.l.]: AAAI, 2000: 93 - 103.

[14] ZAHN J M, POOSALA S, OWEN A B, *et al*. AGEMAP: a gene expression database for aging in mice [J]. PLoS Genetics, 2007, 3(11): e201.

[15] The Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource [J]. Nucleic Acids Research, 2004, 32(1): D258 - D261.