文章编号:1000－2367(2017)05－0093－08 DOI:10.16366/j.cnki.1000－2367.2017.05.016

# 一种基于粗糙均方残基的模糊双聚类方法

孙林[1,2], 刘弱南[1], 张霄雨[1], 孙印杰[1], 宋黎明[1]

(1.河南师范大学 计算机与信息工程学院,河南 新乡 453007;
2.计算智能与数据挖掘河南省高校工程技术研究中心,河南 新乡 453007)

**摘 要**:双聚类作为一种无监督的学习方法,其作用是对基因表达数据进行分析.为了获取较大容量的双聚类簇,弥补传统的双聚类方法在基因表达数据一致波动性方面的不足,引入粗糙集的上、下近似集概念,将粗糙集理论运用到模糊双聚类算法中,将粗糙上、下近似集与加权均方残差相结合,得到新的粗糙均方残基,进而提出一种基于粗糙均方残基的模糊双聚类算法.针对基因表达数据集,首先进行缺失值填补;其次,用非负矩阵分解算法对基因数据集进行降维;最后,计算数据矩阵的粗糙均方残基,结合综合评判度量函数与贴近度原则对矩阵的行列进行删除和添加,得到容量更大的双聚类结果.实验结果表明,该模糊双聚类算法是有效的.

**关键词**:粗糙集;粗糙均方残基;双聚类

**中图分类号**:TP181 **文献标志码**:A

随着基因信息量的不断增加,对基因数据进行处理从而得到有用信息的难度不断上升.基因数据集中通常包含大量的无关基因、冗余基因等,因此,如何从海量信息库中分析并获取有效的基因数据子集成为专家学者研究的重要课题.聚类分析是大数据挖掘的重要研究方向,其目的主要在于发现数据中隐含的类结构,将数据对象分成不同的簇或类,使得同一类内对象之间相似度较大,而不同类的对象之间相似度较小[1－3].作为一种主要的数据分析工具,聚类分析目前已经在数据挖掘、机器学习、模式识别和生物信息学等领域得到了广泛的应用研究[4,5].

传统的聚类分析包括样本聚类和基因聚类,是对行或列进行单一的聚类,将样本和基因分别聚类,得到的类之间没有交叉.传统聚类只能找到整个聚类结果的全部信息,而容易忽略聚类信息中隐含的局部信息;另外,传统聚类需要人为选定聚类中心,而且所选初始值往往会影响后续的聚类结果,容易产生误差[1].为了弥补传统聚类的不足,Cheng 和 Church 首先提出了双聚类算法(简称 CC 算法),是一种贪心迭代搜索策略[6].Getz 等在层次聚类的基础上研究了双向耦合聚类算法,该算法所用的聚类策略是迭代合并[7].Tang 等基于双向耦合聚类提出了关联双向聚类算法[8].Damelin 等将自适应模糊划分的理论融合到双聚类中,设计了一种自适应模糊划分算法[9].Yang 等在 CC 算法的基础上提出了基于概率学的灵活重叠双聚类算法[10].Seridi 等提出一种混合多目标的双聚类算法,并将其应用于生物基因表达谱数据[11].其他学者从评价函数出发对双聚类算法进行改进,Teng 等提出以平均相关系数作为评价函数,通过多次迭代选出基因共表达双聚类簇[12].Ayadi 等提出使用平均一致相似性指数作为约束标准来评价双聚类的连贯性[13].基于信息论中的互信息,张敏等研究基因间相似性的度量标准,可以检测基因中存在的非线性相关性[14].在矩阵数据的基础上,蒲国林等提出一种基于变分贝叶斯的半监督双聚类算法[15].对于双聚类的多目标优化问题,Brizuela 等提出了改进的多目标遗传双聚类算法,采用新颖的组信息编码方式来高效地编码双聚类,减少了局

**收稿日期**:2017-04-23;**修回日期**:2017-06-13.

**基金项目**:国家自然科学基金(61402153;61602158);中国博士后科学基金项目(2016M602247);河南省高等学校重点科研项目计划(14A520069).

**作者简介**:孙林(1979－),男,河南新乡人,河南师范大学副教授,博士,CAAI,CCF 会员,研究方向为粒计算、数据挖掘、生物信息学等.

**通信作者**:刘弱南,E-mail:liurn1202@126.com.

部搜索环节,提高了算法的计算效率[16].林勤等应用多目标人工蜂群算法来寻找双聚类,能够同时优化均方残差和尺寸等相互冲突的目标[17].考虑到各双聚类簇中不同条件会对基因组造成不同影响,刘文华等提出基于加权均方残差的改进双聚类算法[18].

针对 CC 算法存在不能准确发现重叠的双聚类簇,在基因表达数据一致波动性方面的效果较差等问题,文献[19]利用粗糙集上、下近似和边界域的相关理论计算双聚类簇的粗糙比率残基,从而得到更多可能重叠的聚类簇.文献[18]考虑到不同的条件属性对基因组的影响程度不同,提出了基于加权的均方残差双聚类算法,能够得到共表达水平不同的双聚类簇.在上述两种算法思想的基础上,将粗糙集上、下近似集和加权均方残差相结合,构建新的粗糙均方残基,进而提出基于粗糙均方残基的模糊双聚类算法.采用非负矩阵分解算法对基因数据集进行维度约简,再利用基于粗糙均方残基的模糊双聚类算法对降维后的数据集进行处理,获取容量较高、质量较好的双聚类结果.

## 1 相关知识

### 1.1 经典双聚类算法

双聚类算法是一种基于均方残差函数测量双聚簇均一性的方法,其基本思想为:采用贪婪迭代搜索策略,产生 $T$ 个双聚簇,该策略通过增删行列来实现;和原始设定的阈值比较,将均方残差较大的行列删除[6];再根据一定准则对行和列进行添加,得到更大容量的聚类簇;将该算法进行迭代,得到 $T$ 个双聚簇.

平均平方残基作为一种评判标准,对矩阵中数据的一致性进行度量.针对一个数据矩阵 $\boldsymbol{A}=(a_{ij})^{I\times J}$,其矩阵的平均平方残基表示如下:

$$H(I,J)=\frac{\sum\limits_{i<I,j<J}(a_{ij}-a_{Ij}-a_{iJ}-a_{IJ})^2}{|I||J|}=\frac{\sum\limits_{i<I,j<J}F_{ij}}{|I||J|},$$

其中,$|I|$ 是矩阵 $\boldsymbol{A}$ 的行数,$|J|$ 是矩阵 $\boldsymbol{A}$ 的列数,$1\leq i\leq I,1\leq j\leq J$,$a_{iJ}=\frac{\sum\limits_{j<J}a_{ij}}{|J|}$ 是矩阵的行平均值,$a_{Ij}=\frac{\sum\limits_{i<I}a_{ij}}{|I|}$ 是矩阵的列平均值,$a_{Ij}=\frac{\sum\limits_{i<I,j<J}a_{ij}}{|I||J|}=\frac{\sum\limits_{i<I}a_{iJ}}{|I|}=\frac{\sum\limits_{j<J}a_{Ij}}{|J|}$ 是矩阵的平均值,$F_{ij}=(a_{ij}-a_{Ij}-a_{iJ}+a_{IJ})^2$ 是矩阵元素的平方残基.为便于后面计算最大最小平均平方残基,令 $R_i=\frac{\sum\limits_{i<I,j<J}F_{ij}}{|J|}$ 是上述矩阵的单行平均平方残基,$C_j=\frac{\sum\limits_{i<I,j<J}F_{ij}}{|I|}$ 是上述矩阵的单列平均平方残基.

经典双聚类算法的目的是使所得双向聚类的均方残差不大于初始设定的阈值,且容量尽可能大.采用贪心迭代搜索使算法更容易实现,时间复杂度低.但该算法对缺失值的处理方法是用随机值来填充,具有很大的不确定性;而且每次迭代只能产生一个双聚类,若需要得到多个双聚类则需要再次选取一个随机值来替换.

### 1.2 基于模糊集的双聚类算法

设 $U=\{u_1,u_2,\cdots,u_n\}$ 为相关因素集,$n$ 为相关因素个数.$H=\{h_1,h_2,\cdots,h_m\}$ 为相应评判集,$m$ 为评判标准的个数.对集合 $U$ 中的所有元素进行综合评判,建立如下隶属矩阵:

$$\boldsymbol{R}=\begin{pmatrix}r_{11}&r_{12}&\cdots&r_{1m}\\r_{21}&r_{22}&\cdots&r_{2m}\\\vdots&\vdots&&\vdots\\r_{n1}&r_{n2}&\cdots&r_{nm}\end{pmatrix},$$

然后,计算相关因素集的权重向量 $\boldsymbol{O}=(o_1,o_2,\cdots,o_n)$,将评判集 $H=\{h_1,h_2,\cdots,h_m\}$ 进行标准化处理,令评判目标函数 $D=\boldsymbol{O}\circ\boldsymbol{R}$,其中,"$\circ$"表示合成算子;最后,根据最大隶属原则或择近原则的贴近度来选择评判结

果.本文采用择近原则的贴近度来处理评判结果.

择近原则是指一个模糊集对已确定的标准集的识别,它的本质是求两个集合的贴近度.此贴近度用 $N(A,D)$ 表示,它和两个集合的接近程度成正比.给定 $P=\{p_1,p_2,\cdots,p_n\}\sim[0,1]$,有 $A,D\in P$,那么欧几里得贴近度可表示为

$$N(A,D)=1-\frac{1}{\sqrt{n}}\left\{\sum_{i-1}^{n}[A(p_i)-D(p_i)]^2\right\}^{\frac{1}{2}},\tag{1}$$

将欧几里得贴近度归一化处理后,可得

$$N(A,D)=1-\frac{1}{\sqrt{n}}\left\{\sum_{i-1}^{n}\left[\frac{A(p_i)}{\sum_{j-1}^{I}A(p_j)}-\frac{D(p_i)}{\sum_{j-1}^{I}D(p_j)}\right]^2\right\}^{\frac{1}{2}}.$$

## 2　基于粗糙均方残基的模糊双聚类方法

粗糙集理论是 1982 年由 Pawlak 提出的.它是一种非常有效的处理不确定性的工具[19,20].在近年来的数据挖掘方法中,粗糙集越来越受到重视,且已经在许多科学与工程领域中得到了应用.它是当前数据挖掘与人工智能领域的研究热点之一.粗糙集以等价关系为基础,用来处理不确定性数据问题.本文在粗糙集上下近似集概念的基础上,将粗糙集理论运用到模糊双聚类算法中,将上下近似集与加权均方残差相结合,提出一种基于粗糙均方残基的模糊双聚类算法.

### 2.1　粗糙均方残基

在聚类过程中,将聚类簇的位置和形状用簇的核心来表达,该核心由簇的下近似集得到.聚类簇的边界域有可能相互重叠,该区域由上下近似集的差集得到.因此,可以通过粗糙集上下近似算法,获得基因表达数据中可重叠的簇.文献[19]利用粗糙集上、下近似和边界域的相关理论计算双聚类簇的粗糙比率残基,从而得到更多可能重叠的聚类簇.文献[18]考虑到不同的条件属性对基因组的影响程度不同,提出了基于加权的均方残差双聚类算法,能够得到共表达水平不同的双聚类簇.基于上述两种算法的思想,将文献[19]中粗糙集上下近似理论和文献[18]中的加权均方残差概念相结合,提出一种新的粗糙均方残基,并将其应用于模糊双聚类算法中,得到基于粗糙均方残基的模糊双聚类算法.

**定义 1**[19]　设 $U$ 为论域,$R\subseteq U\times U$ 是论域上的一个等价关系,二元对 $S=(U,R)$ 称为论域上的近似空间,$X$ 为论域 $U$ 的任意子集,集合 $X$ 的上近似集 $\overline{R}(X)$、下近似集 $\underline{R}(X)$ 和边界域 $G_R(X)$ 分别表示如下

$$\overline{R}(X)=\bigcup\{E_i\mid E_i\in R\wedge E_i\cap X\neq\varnothing\},$$
$$\underline{R}(X)=\bigcup\{E_i\mid E_i\in R\wedge E_i\subseteq X\},$$
$$G_R(X)=\overline{R}(X)-\underline{R}(X),$$

其中,$E_i$ 为等价集,$\overline{R}(X)$ 表示它的元素不一定属于集合 $X$,$\underline{R}(X)$ 表示它的所有元素都属于 $X$.

基于上述粗糙集上、下近似集给出聚类簇的上、下近似和边界域定义.

**定义 2**　设 $U$ 为论域,$R\subseteq U\times U$ 是论域上的一个等价关系,二元对 $S=(U,R)$ 称为论域上的近似空间;假定二元对 $S$ 与数据矩阵 $\boldsymbol{A}$ 存在一一对应关系,数据矩阵 $\boldsymbol{A}$ 中有 $T$ 个双聚类簇,第 $k$ 个簇为 $B_k=(I_k,J_k)$,其中 $1\leq k\leq T$,则有

(1) $B_k$ 的上近似集 $\overline{B}_K$、下近似集 $\underline{B}_k$ 和边界域 $G_{B_k}$ 定义如下:

$$\overline{B}_k=\bigcup\{a_{ij}\mid a_{ij}\in A\wedge a_{ij}\cap B_k\neq\varnothing\},$$
$$\underline{B}_k=\bigcup\{a_{ij}\mid a_{ij}\in A\wedge a_{ij}\subseteq B_k\},$$
$$G_{B_k}=\overline{B}_k-\underline{B}_k;$$

(2) $I_k$ 的上近似集 $\overline{I}_k$、下近似集 $\underline{I}_k$ 和边界域 $G_{I_k}$ 定义如下:

$$\overline{I}_k=\bigcup\{i\mid i\in A\wedge i\cap I_k\neq\varnothing\},$$
$$\underline{I}_k=\bigcup\{i\mid i\in A\wedge i\subseteq I_k\},$$
$$G_{I_k}=\overline{I}_k-\underline{I}_k;$$

(3) $J_k$ 的上近似集$\overline{J}_k$、下近似集$\underline{J}_k$ 和边界域$G_{J_k}$ 定义如下:

$$\overline{J}_k = \bigcup \{j \mid j \in A \wedge j \cap J_k \neq \varnothing\},$$

$$\underline{J}_k = \bigcup \{j \mid j \in A \wedge j \subseteq J_k\},$$

$$G_{J_k} = \overline{J}_k - \underline{J}_k.$$

**定义 3** 设数据矩阵 $\boldsymbol{A}$ 中有 $T$ 个双聚类簇,第 $k$ 个簇 $B_k = (I_k, J_k)$ 的粗糙均方残基 $R'(I_k, J_k)$、行粗糙均方残基 $R'_i$、列粗糙均方残基 $C'_j$ 和粗糙平方残基 $F'_{ij}$ 分别定义如下:

$$R'(I_k, J_k) = \frac{\sum\limits_{i<I_k, j<J_k} w_k (a_{ij} - a'_{I_k j} - a'_{i J_k} + a'_{I_k J_k})^2}{|I_k||J_k|}, \tag{2}$$

$$R'_j = \frac{\sum\limits_{i<I_k, j<J_k} w_k (a_{ij} - a'_{I_k j} - a'_{i J_k} + a'_{I_k J_k})^2}{|J_k|}, \tag{3}$$

$$C'_j = \frac{\sum\limits_{i<I_k, j<J_k} w_k (a_{ij} - a'_{I_k j} - a'_{i J_k} - a'_{I_k J_k})^2}{|I_k|}, \tag{4}$$

$$F'_{ij} = (a_{ij} - a'_{I_k j} - a'_{i J_k} + a'_{I_k J_k})^2, \tag{5}$$

其中

$$a'_{I_k j} = \begin{cases} w_{\text{low}} \times \dfrac{\sum\limits_{a_{ij} \in \underline{B}_k} a_{ij}}{|\underline{I}_k|} + w_{\text{up}} \times \dfrac{\sum\limits_{a_{ij} \in G_{B_k}} a_{ij}}{|G_{I_k}|}, & \text{若 } j \in G_{J_k}, \\ \dfrac{\sum\limits_{a_{ij} \in \underline{B}_k} a_{ij}}{|\overline{I}_k|}, & \text{若 } j \in \underline{J}_k, \end{cases}$$

$$a'_{i J_k} = \begin{cases} w_{\text{low}} \times \dfrac{\sum\limits_{a_{ij} \in \underline{B}_k} a_{ij}}{|\underline{J}_k|} + w_{\text{up}} \times \dfrac{\sum\limits_{a_{ij} \in G_{B_k}} a_{ij}}{|G_{J_k}|}, & \text{若 } i \in G_{I_k}, \\ \dfrac{\sum\limits_{a_{ij} \in \underline{B}_k} a_{ij}}{|\overline{J}_k|}, & \text{若 } j \in \underline{I}_k, \end{cases}$$

$$a'_{I_k J_k} = \begin{cases} w_{\text{low}} \times \dfrac{\sum\limits_{a_{ij} \in \underline{B}_k} a_{ij}}{|\underline{B}_k|} + w_{\text{up}} \times \dfrac{\sum\limits_{a_{ij} \in G_{B_k}} a_{ij}}{|G_{B_k}|}, & \text{若 } G_{B_k} \neq \varnothing, \\ \dfrac{\sum\limits_{a_{ij} \in \underline{B}_k} a_{ij}}{|\overline{B}_k|}, & \text{否则}, \end{cases}$$

$1 \leq k \leq T$, $w_k$ 为第 $k$ 个簇的属性权重, $\sum\limits_{k-1}^{T} w_k = 1$, $w_{\text{low}}$ 为下界权值, $w_{\text{up}}$ 为上界权值,且 $w_{\text{low}} + w_{\text{up}} = 1$.

## 2.2 非负矩阵分解降维

非负矩阵分解通过寻找低秩,达到分解数据矩阵的效果[21]. 非负矩阵分解通过"乘性"迭代规则来保证经过每次迭代后数据矩阵的元素为非负的,给定非负矩阵 $\boldsymbol{V}_{m\times n}$,寻找两个非负子矩阵 $\boldsymbol{W}_{m\times k}$ 和 $\boldsymbol{H}_{k\times n}$,使得: $\boldsymbol{V}_{m\times n} = \boldsymbol{W}_{m\times k}\boldsymbol{H}_{k\times n}$,通常 $k < m$ 且 $k < n$. 本文用"乘性"迭代规则求矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{H}$, 其计算公式如下:

$$\boldsymbol{W}_{ik} \leftarrow \boldsymbol{W}_{ik} \frac{(\boldsymbol{V}\boldsymbol{H}^T)_{ik}}{(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T)_{ik}}, \tag{6}$$

$$\boldsymbol{H}_{kj} \leftarrow \boldsymbol{H}_{kj} \frac{(\boldsymbol{W}^T\boldsymbol{V})_{kj}}{(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H})_{kj}}. \tag{7}$$

非负矩阵分解算法中求解分解矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 的方法是一种基于"乘性"迭代的方法,它能够使得左右非负矩阵存储空间变小,使得高维的数据矩阵维数大大降低,进而适合处理大规模数据[21]. 又因为文献[22-23]

指出马氏距离法可以根据整个空间上的特征分布情况对数据进行判别分析,不用考虑各参数特征的量纲.因此,本文先采用马氏距离法对基因表达数据集中的缺失数据进行填补,再运用非负矩阵分解算法进行降维,其算法详细步骤如下.

算法 1

输入　原始数据集 data

输出　降维后的数据矩阵 $\boldsymbol{A}$

步骤 1　使用马氏距离公式 $d^2(X,\omega)=(X-\overline{X^{(\omega_i)}})^{\mathrm{T}}S^{-1}(X-\overline{X^{(\omega_i)}})$ 计算缺失数据 $X=\{x_1,x_2,x_3,\cdots,x_n\}$ 与数据集中每个类中心 $\omega_i$ 的距离;其中,$\overline{X^{(\omega_i)}}$ 为第 $i$ 个类的类中心,$S$ 为全体样本的协方差.选出缺失数据的最近邻基因计算其加权平均值,并将其作为估计值.

步骤 2　利用非负矩阵分解算法对填补后的数据矩阵 $\boldsymbol{V}$ 进行降维.初始化随机选取两个非负矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{H}$,利用公式(6)和(7)分别对 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 进行迭代更新,使得 $\boldsymbol{V}=\boldsymbol{WH}$,得到降维后的数据矩阵.

### 2.3　基于粗糙均方残基的模糊双聚类算法

首先利用公式(2)计算基因表达数据矩阵的粗糙均方残基、公式(3)计算基因表达数据矩阵的行粗糙均方残基、公式(4)计算基因表达数据矩阵的列粗糙均方残基,然后求出矩阵中的最大最小粗糙均方残基,其计算公式表示如下:

$$M_k=\min(C'_1,C'_2,C'_3,\cdots,C'_J,R'_1,R'_2,R'_3,\cdots,R'_I),\tag{8}$$

$$M_l=\max(C'_1,C'_2,C'_3,\cdots,C'_J,R'_1,R'_2,R'_3,\cdots,R'_I).\tag{9}$$

根据最大最小粗糙均方残基值,可以计算其方差选取阈值为 $\delta=\dfrac{(M_1-M_k)^2}{12}$.

由公式(5)计算矩阵元素的粗糙平方残基 $F'_{ij}$,把得到的值作为评判条件,并依此建立平方残基矩阵 $\boldsymbol{B}$ 作为矩阵 $\boldsymbol{A}$ 的评判矩阵,其中评判矩阵 $\boldsymbol{B}$ 表示为

$$\boldsymbol{B}=\begin{Bmatrix}F'_{11} & F'_{12} & \cdots & F'_{1J}\\ F'_{21} & F'_{22} & \cdots & F'_{2J}\\ \vdots & \vdots & & \vdots\\ F'_{I1} & F'_{I2} & \cdots & F'_{IJ}\end{Bmatrix}.\tag{10}$$

各列对应的评判函数为 $D_r=\boldsymbol{B}\circ\boldsymbol{A}$,其中 $\boldsymbol{A}_r=(a_{r1},a_{r2},\cdots,a_{rI})$ 由矩阵转化为列向量可得,$\boldsymbol{A}_r$ 表示各列的权数;各行对应的评判函数为 $D_c=\boldsymbol{B}\circ\boldsymbol{A}$,其中 $\boldsymbol{A}_c=(a_{c1},a_{c2},\cdots,a_{cJ})$ 由矩阵转化为行向量可得,$\boldsymbol{A}_c$ 表示各行的权数.

对公式(1)的贴近度函数进行如下归一化处理:

$$N(A,D)=1-\frac{1}{\sqrt{n}}\left\{\sum_{i-1}^{n}\left[\frac{A(p_i)}{\sum_{j-1}^{I}A(p_i)}-\frac{D(p_i)}{\sum_{j-1}^{I}D(p_j)}\right]^2\right\}^{\frac{1}{2}}.\tag{11}$$

这里设计的模糊双聚类算法是通过多次迭代得到聚类结果.经过一次迭代后,为了避免对下次迭代产生影响,需要对上次迭代后的聚类结果进行屏蔽;若某个元素存在于子矩阵,则将其在原始数据矩阵中相对应的粗糙平方残基值增大,从而屏蔽掉上次迭代的聚类结果,尽可能避免对最终双聚类产生影响.

随机给定 $\boldsymbol{V}=(v_1,v_2,\cdots,v_I)\in R_+$,$\sum_{x-1}^{I}v_x=I$,使得

$$S_{ij}=\max_{V}\{v_x(a_{ij}-a'_{I_kj}-a'_{iJ_k}+a'_{I_kJ_k})^2\}.\tag{12}$$

利用公式(12)选取最大的粗糙平方残基值,然后用该值替换原始数据矩阵中的粗糙平方残基值.由此,可以设计基于粗糙均方残基的模糊双聚类算法,其详细步骤如下:

算法 2

输入　降维后的数据矩阵 $\boldsymbol{A}$,上、下界权重 $w_{\mathrm{up}}$ 和 $w_{\mathrm{low}}$,终止条件阈值 $\delta$,评判次数 $g$.

输出　双聚类簇 $\boldsymbol{C}$.

步骤 1　使用公式(8)计算矩阵元素的最大粗糙平方残基 $M_l$、公式(9)计算矩阵元素的最小粗糙平方残

基 $M_k$,由 $\delta=\frac{(M_l-M_k)^2}{12}$ 计算终止条件阈值 $\delta$;

步骤 2 使用公式(3)和公式(4)计算矩阵和各行列的粗糙均方残基值,选出并删除最大粗糙均方残基所在的行或列,更新删除后的矩阵,直到均方残基值小于步骤 1 中所得阈值 $\delta$;

步骤 3 使用公式(3)和公式(4)计算矩阵和各行列的粗糙均方残基值,并将结果和公式(2)所得的粗糙均方残基 $R'(I_k,J_k)$ 进行比较,将小于粗糙均方残基 $R'(I_k,J_k)$ 且不在 $I$ 行或 $J$ 列的粗糙均方残基值添加到步骤 2 所得的矩阵中,更新矩阵 $\boldsymbol{A}$,直到其粗糙均方残基值小于步骤 1 中所得阈值 $\delta$;

步骤 4 使用公式(10)计算矩阵 $\boldsymbol{A}$ 对应的粗糙平方残基矩阵 $\boldsymbol{B}$,作为评判矩阵;

步骤 5 利用 $D=\{D_{r1},D_{r2},\cdots,D_{rI},D_{c1},D_{c2},\cdots,D_{cJ}\}$ 计算矩阵各行列的评判向量;

步骤 6:使用公式(8)计算最小粗糙均方残基 $M_k$ 所在的行或列 $\boldsymbol{A}_k$,再使用公式(11)计算各行列的评判向量对于 $\boldsymbol{A}_k$ 的贴近度 $N=\{N_1,N_2,\cdots,N_{I+J}\}$;

步骤 7 若 $\delta<\frac{R'(I_k,J_k)}{N_x}$,则删除对应的行或列;

步骤 8 更新子矩阵 $\boldsymbol{A}$,如果没有删减行列,则把 $\boldsymbol{A}_k$ 对应的粗糙均方残基从公式(8)中删除;令 $g=g-1$,返回步骤 4 执行循环操作,当 $g=0$ 时停止,可得到双聚类簇结果 $\boldsymbol{C}$.

## 3 实验结果及分析

本文算法运行的实验环境:32 位 Windows7 操作系统,Intel(R)Core(TM)i3-2328MCPU@2.20GHz 处理器,2.00GB 内存,采用 MATLAB2012b 工具箱进行编码.下载 2 种基因表达谱数据集(http://bioinformatics.rutgers.edu/Static/Supplements/CompCancer/datasets.htm),分别进行实验.数据集详见表 1.

**表 1 基因表达数据集**

| 序号 | 数据集 | 样本 | 属性 |
|---|---|---|---|
| 1 | Laiho-2007 | 37 | 2202 |
| 2 | Armstrong-2002-v2 | 72 | 2194 |
| 3 | GLIOMA | 50 | 4434 |
| 4 | Carcinom | 174 | 9182 |
| 5 | Prostate_GE | 102 | 5966 |

首先,对原始的待处理基因表达谱数据集进行处理,用算法 1 对缺失值进行填补,并对填补后的数据降维.算法 1 对数据集 Laiho-2007,Armstrong-2002-v2,GLIOMA,Carcinom 和 Prostate_GE 的分解结果如图 1~图 5 所示.

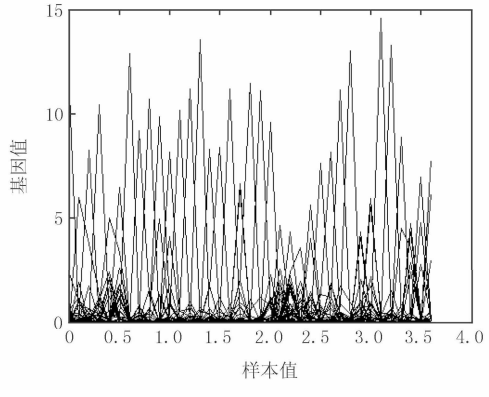


图1 Laiho-2007非负矩阵降维结果

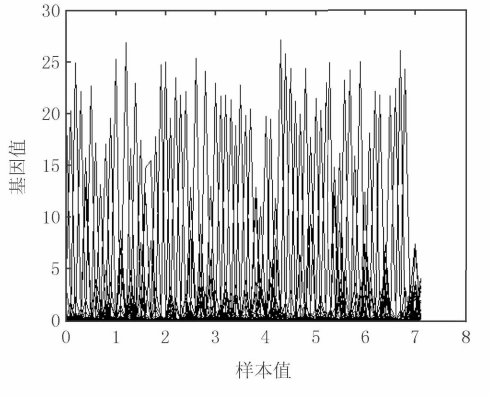


图2 Armstrong-2002-v2非负矩阵降维结果

由图 1~图 5 可知,只有少部分的数据具有显著表达特性,因此选出这些数据得到降维后的数据子集,对降维后的数据子集进行双聚类.根据专家经验设置参数如下:权重 $w_{\text{low}}=0.75$,$w_{\text{up}}=0.25$,评判次数 $g=$

10. 为了验证本文提出算法 2 的有效性，选用文献[6]中 CC 算法、文献[24]中模糊双聚类算法与本文算法 2 进行仿真实验. 针对 Laiho-2007 数据集，CC 算法和本文算法 2 得到的双聚类簇的平均样本数相同，但是模糊双聚类算法的平均样本数较大；3 种算法获得的平均基因数和平均容量依次增加；另外，本文算法 2 计算的均方残基与 CC 算法相比较小，和模糊双聚类算法相比，本文算法 2 所得的均方残基较大. 因此，本文算法能够筛选出容量更大的双聚类结果，但在选取均方残差值方面尚不够完善. 总之，基于粗糙上、下近似集的模糊双聚类在获取容量较大的双聚类结果方面效果较好，能够使数据呈现出很好的波动一致性.

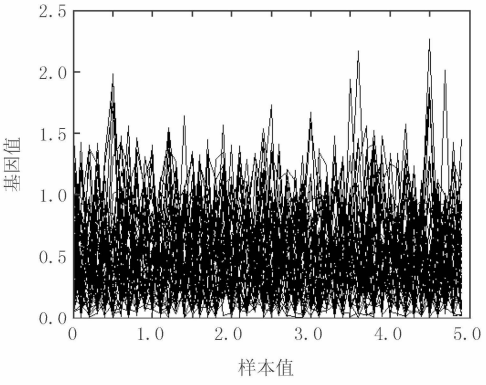


图3　GLIOMA非负矩阵降维结果

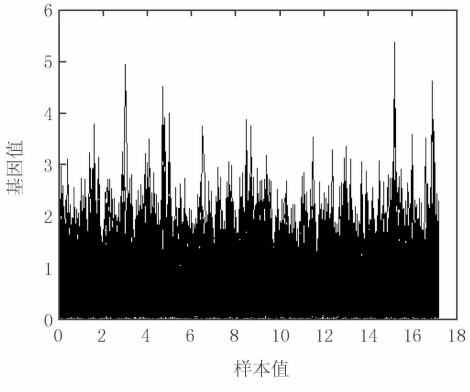


图4　Carcinom非负矩阵降维结果

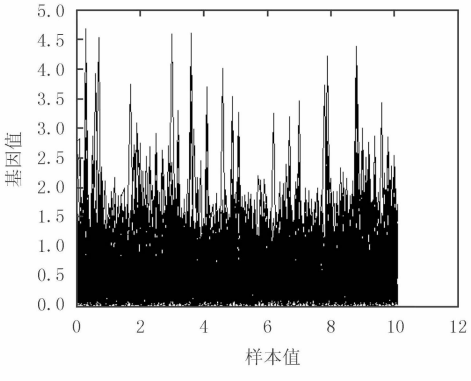


图5　Prostate_GE非负矩阵降维结果

## 4　结　论

本文改进了传统的模糊双聚类算法，对基因表达谱数据集进行聚类分析. 首先对原始基因数据集进行预处理，用马氏距离算法对缺失值进行填补，将填补后的数据集采用非负矩阵分解法降维，对降维后的数据子集进行双聚类. 结合粗糙上、下近似集和加权均方残差，给出一种新的粗糙均方残基，提出一种基于粗糙均方残基的模糊双聚类算法. 最后，将该算法应用于基因表达数据集上并进行仿真实验，结果表明所提算法能够获取容量较高、质量较好的双聚类结果，且能够使基因表达数据呈现更好的一致波动性.

## 参　考　文　献

[1] Sun L, Xu J C, Yin J J. An effective fuzzy kernel clustering analysis approach for gene expression data[J]. Bio-Medical Materials and Engineering, 2015, 26(S1): S1863-S1869.

[2] 赵兴旺，梁吉业. 一种基于信息熵的混合数据属性加权聚类算法[J]. 计算机研究与发展，2016，53(5)：1018-1028.

[3] 杨大勇，葛琪，董永超，等. 基于 K 均值聚类的光伏电站运行状态模式识别研究[J]. 电力系统保护与控制，2016，44(14)：25-30.

[4] Ray S S, Ganivada A, Pal S K. A granular self-organizing map for clustering and gene selection in microarray data[J]. IEEE transactions on neural networks and learning systems, 2016, 27(9): 1890-1906.

[5] 高洁，李群湛，汪佳，等. 基于模糊聚类的 NExT-ERA 低频振荡类噪声辨识[J]. 电力系统保护与控制，2016，44(22)：40-49.

[6] Cheng Y, Church G M. Biclustering of expression data[C]//Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, 2000, 93-103.

[7] Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data[J]. Proceedings of the National Academy of Sciences, 2000, 97(22): 12079-12084.

[8] Tang C, Zhang L, Ramanathan M, et al. Interrelated two-way clustering: an unsupervised approach for gene expression data analysis[C]//

Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering. Piscataway: IEEE Press, 2001: 41-48.

[9] Damelin S B, Gu Y, Wunsch D C, et al. Fuzzy adaptive resonance theory, diffusion maps and their applications to clustering and biclustering[J]. Mathematical Modelling of Natural Phenomena, 2015, 10(3): 206-211.

[10] Yang J, Wang W, Wang H X, et al. δ-clusters: Capturing subspace correlation in a large data set[C]//Proceedings of the 18th IEEE International Conference on Data Engineering. Piscataway: IEEE Press, 2002: 517-528.

[11] Seridi K, Jourdan L, Talbi E G. Parallel hybrid meta heuristic for multi-objective biclustering in microarray data[C]//IEEE International Parallel and Distributed Processing Symposium Workshops. Piscataway: IEEE Press, 2012: 625-633.

[12] Teng L, Chan L. Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data[J]. Journal of Signal Processing Systems, 2008, 50(3): 267-280.

[13] Ayadi W, Elloumi M, Hao J K. BicFinder: a biclustering algorithm for microarray data analysis[J]. Knowledge and Information Systems, 2012, 30(2): 341-358.

[14] 张敏, 戈文航. 基于概率计算的重叠双聚类算法[J]. 计算机工程与设计, 2012, 33(9): 3579-3583.

[15] 蒲国林, 邱玉辉. 一种基于变分贝叶斯的半监督双聚类算法[J]. 计算机应用研究, 2015, 32(8): 2299-2301.

[16] Brizuela C A, Luna-Taylor J E, Martinez-Perez I, et al. Improving an evolutionary multi-objective algorithm for the biclustering of gene expression data[C]//IEEE Congress on Evolutionary Computation. Piscataway: IEEE Press, 2013: 221-228.

[17] 林勤, 薛云, 林斯达, 等. 多目标人工蜂群双聚类算法在基因表达数据中的应用研究[J]. 华南师范大学学报(自然科学版), 2016, 48(2): 116-123.

[18] 刘文华, 梁永全, 冯政. 基于加权均方残差的改进双聚类算法[J]. 模式识别与人工智能, 2016, 29(6): 519-526.

[19] 李刚, 苗夺谦, 王睿智. 一种基于粗糙遗传算法的缩放模式双聚类分析方法[J]. 计算机科学, 2010, 37(1): 225-228.

[20] Pawlak Z. Rough sets[J]. International Journal of Parallel Programming, 1982, 11(5): 341-356.

[21] Li L, Zhang Y J. FastNMF: highly efficient monotonic fixed-point nonnegative matrix factorization algorithm with good applicability[J]. Journal of Electronic Imaging, 2009, 18(3): 033004.

[22] 杨涛, 骆嘉伟, 王艳, 等. 基于马氏距离的缺失值填充算法[J]. 计算机应用, 2005, 25(12): 2868-2871.

[23] 郝胜轩, 宋宏, 周晓锋. 一种基于双聚类的缺失数据填补方法[J]. 计算机应用研究, 2015, 32(3): 674-678.

[24] 邱杰, 喻昕, 罗海琼, 等. 一种基于模糊集理论的双聚类算法[J]. 平顶山学院学报, 2013, 28(5): 8-11.

# A Fuzzy Biclustering Approach Based on Rough Average Square Residue

Sun Lin[1,2], Liu Ruonan[1], Zhang Xiaoyu[1], Sun Yinjie[1], Song Liming[1]

(1. Collage Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China;
2. Engineering Technology Research Center for Computing Intelligence & Data Mining of Henan Province, Xinxiang 453007, China)

**Abstract**: Biclustering as an unsupervised learning method can analyze gene expression data. However, some traditional biclustering methods have the shortcoming of consistent volatility for gene expression data. To solve this problem, and obtain large capacity clusters of biclustering, the upper and lower approximation of rough set was introduced in this paper, and the rough set theory was applied into fuzzy biclustering algorithm. By combining upper and lower approximation with weighted mean square residual, a novel rough mean square residue was defined. Then an improved fuzzy biclustering algorithm based on rough mean square residue was proposed. For gene expression dataset, the missing values were filled up firstly. Λ factorization algorithm of non-negative matrix was used to reduce dimension of gene dataset. And the rough mean square residue of data matrix was calculated. Finally, through integrating a comprehensive evaluation measure function and nearness degree, the rows and columns of matrixes were deleted or added in order to obtain a larger of biclustering results. Experimental results show that the proposed fuzzy biclustering algorithm is efficient.

**Keywords**: rough set; rough average square residue; biclustering

[责任编辑 杨浦]