

# 基于伪 F 统计 FAMC 算法的基因表达数据分析

刘文远 李建飞 王宝文 于家新

(燕山大学信息科学与工程学院 秦皇岛 066004)

**摘要** 基因芯片技术在给人类带来巨大机遇的同时也带来一些挑战。针对基因表达数据的海量性,以及基因类属的不确定性等问题,提出了一种基于伪 F 统计量(PFS)的模糊属性均值聚类 FAMC(fuzzy attribute c-means clustering)算法,并就模糊参数  $m$  的确定问题提出了有效的解决方法。最后将其在标准的基因表达数据上进行测试分析,取得了较优的聚类结果。

**关键词** 基因表达数据, FAMC 算法, 稳态函数, 伪 F-统计

## Gene Expression Data Analysis of FAMC Algorithm Based on Pseudo F-statistics

LIU Wen-yuan LI Jian-fei WANG Bao-wen YU Jia-xin

(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

**Abstract** Gene chip technology not only brings the huge opportunity to the humanity but also brings some challenges simultaneously. In view of gene expression data's magnanimous, uncertainty about the gene class and so on, this article proposed one kind fuzzy attribute c-means clustering algorithm based on Pseudo F-statistics, and proposed an effective solution for defining the fuzzy parameter  $m$ . Finally carried on it in the standard gene expression data for testing analysis, obtained superior cluster results.

**Keywords** Gene expression data, FAMC algorithm, Stable function, Pseudo F-statistics

## 1 引言

基因芯片(Gene Chip)是分子生物学实验技术的一个突破,利用该技术可以同时成千上万个基因的表达进行平行分析,产生了海量的有用数据,分析与整理这些数据成为利用这一技术的一个主要瓶颈问题。利用数据挖掘中的聚类分析技术分析基因芯片表达数据、预测基因功能,是对基因功能进行解读的一个有利工具。

模糊 C 均值聚类算法 FCM(Fuzzy C-means Clustering)是基于模糊数学理论的一种非监督学习方法,应用十分广泛。它在处理具有模糊性和海量性的数据方面十分有效,是进行基因表达数据分析的一个非常有效的手段。对于一般的 FCM 算法的研究, Karayiannis 提出了一种加权 FCM 算法<sup>[1]</sup>,程乾生基于稳态函数提出了属性聚类算法<sup>[2]</sup> AMC(Attribute Means Clustering),文献<sup>[3]</sup>探讨了该算法的稳健性能优于 FCM 算法。而 FAMC 算法正是在 ACM 算法的基础上推广得到的,其聚类效果、稳健性较之 AMC 算法又有了进一步提高。

FAMC 算法的性能虽然不错,但参数的初始化问题仍待解决。只有当算法在用适当的参数,在合适的数据上运行时才能产生满意的结果,这往往对于专业用户来说也难以对其做出正确的选择。为此,我们提出了基于伪 F 统计量的模糊属性均值聚类算法,并就模糊参数  $m$  的确定问题提出了有效

的解决方法;并使其成为基因表达数据分析的一个简便、有力的工具。

## 2 模糊属性均值聚类算法

假设模式空间为  $p$  维向量空间  $X, X \in R^p$ , 且  $X$  包含  $N$  个样本  $x_n = (x_{n1}, x_{n2}, \dots, x_{np})^T, (1 \leq n \leq N)$ , 则称  $X$  为对象空间。对于整数  $c, 2 \leq c \leq N$ ,  $X$  的模糊  $c$  划分就是要在  $X$  上产生  $c$  个类  $\{C_1, C_2, \dots, C_c\}$ , 则称其为属性空间。

设第  $k$  类  $C_k$  的类中心为  $v_k = (v_{k1}, v_{k2}, \dots, v_{kp})$ ,  $1 \leq k \leq c$ ; 令  $\mu_{kn}$  表示第  $n$  个样本属于第  $k$  类的属性测度。记  $U = (\mu_{kn})$ ,  $V = (v_1, v_2, \dots, v_c)$ 。

于是,模糊  $c$  划分空间可表示为:

$$M_{fc} = \{U \in R^{c \times N} \mid \mu_{kn} \in [0, 1], \forall n, k; \sum_{k=1}^c \mu_{kn} = 1, \forall n; 0 < \sum_{n=1}^N \mu_{kn} < n, \forall k\} \quad (1)$$

模糊聚类分析的目的就是在该空间上求出属性测度  $\mu_{kn}$ , 然后根据  $\arg \max_{1 \leq k \leq c} \mu_{kn}$  判定样本  $x_n$  所属的类。

### 2.1 AMC 算法

基于稳态函数和加权均方逼近,文献<sup>[2]</sup>提出了属性均值聚类算法。

**定义** 设  $\rho(t)$  为  $[0, +\infty)$  上的正的可微函数( $\rho(0)$  可以为 0)。令

$$\omega(t) = \rho'(t)/2t \quad (2)$$

到稿日期:2008-04-08 本文受国家自然科学基金(60671025, 60474065)资助。

刘文远(1968—),男,博导,教授,研究方向为数据挖掘、生物信息处理;李建飞(1981—),男,硕士研究生,研究方向为生物信息数据挖掘, E-mail: lijianfei-518@163.com;王宝文(1957—),男,副教授,硕士生导师,研究方向为模式识别、软计算。

若  $\omega(t)$  为  $[0, +\infty)$  上的正的不增函数, 就称  $\rho(t)$  为稳态函数, 称  $\omega(t)$  为权函数。由式(2)知, 可以取  $\rho(t)$  为

$$\rho(t) = 2 \int_0^t s \cdot \omega(s) ds \quad (3)$$

然而, 属性聚类算法 ACM 的目标函数为

$$O(U, V) = \sum_{k=1}^c \sum_{n=1}^N \rho(\| \mu_{kn} (x_n - v_k) \|) \quad (4)$$

最小化目标函数可以转化为迭代加权最小平方问题:

$$Q^{(i)}(U, V) = \sum_{k=1}^c \sum_{n=1}^N \omega(\| \mu_{kn}^{(i)} (x_n - v_k^{(i)}) \|) \cdot \| \mu_{kn} (x_n - v_k) \|^2 \quad (5)$$

其中,  $V, U$  可分别由  $\min Q^{(i)}(U^{(i)}, V)$ ,  $\min Q^{(i)}(U, V^{(i+1)})$  得到, 此即 AMC 算法。

## 2.2 FAMC 算法

FAMC 算法正是在 AMC 算法的基础上推广而来, 所以将式(4)和式(5)改写为:

$$O(U, V) = \sum_{k=1}^c \sum_{n=1}^N \rho(\mu_{kn} \| x_n - v_k \|) \quad (6)$$

$$Q^{(i)}(U, V) = \sum_{k=1}^c \sum_{n=1}^N \omega(\mu_{kn}^{(i)} \| x_n - v_k^{(i)} \|) \cdot ((\mu_{kn})^2 \| x_n - v_k \|^2) \quad (7)$$

进一步将式(6)和式(7)推广为:

$$O(U, V) = \sum_{k=1}^c \sum_{n=1}^N \rho((\mu_{kn})^{\frac{m}{2}} \| x_n - v_k \|) \quad (8)$$

$$Q^{(i)}(U, V) = \sum_{k=1}^c \sum_{n=1}^N \omega((\mu_{kn}^{(i)})^{\frac{m}{2}} \| x_n - v_k^{(i)} \|) \cdot ((\mu_{kn})^m \| x_n - v_k \|^2) \quad (9)$$

其中  $m > 1$ , 称由式(8)和式(9)定义的即为模糊属性均值聚类算法 FAMC。同样,  $V, U$  可分别由  $\min Q^{(i)}(U^{(i)}, V)$ ,  $\min Q^{(i)}(U, V^{(i+1)})$  得到。当  $m = 2$  时, FAMC 即为 AMC 算法, 当  $\rho(t) = t^2$ ,  $\omega(t) = 1$  时, FAMC 即为一般 FCM 算法。所以一般 FCM 和 AMC 均为 FAMC 的特例。

## 2.3 模糊度 $m$ 对算法的影响

对于不同的数据集, 就会有不同的  $m$  值; 对于不同的  $m$  值, 就会有不同的模糊  $c$ -划分, 进而产生不一样的聚类效果。 $m$  值过大或过小都是不合适的, 故必须考虑  $m$  的优选问题<sup>[4,5]</sup>。通过查阅文献, 提出如下的解决思路: 首先, 我们通过分析计算为  $m$  确定一个极值(因为找到了上限值, 再来确定  $m$  的近似值就相对容易了); 其次, 就是利用一些规则推导出一个理想的、不依赖于类数  $c$  的  $m$  值。

在文献[4]中, 明确提到: 当  $m$  趋于其极值时,  $\mu_{kn}$  的值趋于  $1/c$ 。因此, 对于一个给定的数据集, 对于  $m$  来说一定存在一个极大值  $m_u$ , 此时令  $\mu_{kn}$  的值等于  $1/c$ 。当  $m$  趋于 1 时, 聚类算法退化为硬划分。在此, 我们主张  $m$  的取值小于等于 2, 这有利于得到较高的隶属度, 使类内基因更加紧凑。即: 令  $m = 1 + m_0$ , 如果  $m_u \geq 10$ , 则  $m_0 = 1$ ; 如果  $m_u < 10$ , 则  $m_0 = m_u / 10$ 。

## 3 多维 PFS 判别函数<sup>[6,7]</sup>

PFS(伪 F-统计, Pseudo F-statistics)是来自 ANOVA 的一个统计量, 常用于 F-检验。对于  $P(P \geq 1)$  维变量时的样本, 定义一个“伪 F 统计比率”如下:

$$PFS = \frac{tr(S_B^p)(N-c)}{tr(S_W^p)(c-1)} \quad (10)$$

其中  $tr(S_B^p)$  是矩阵  $S_B^p$  之迹,  $S_B^p$  和  $S_W^p$  分别是  $P$  维变量样本的类间和类内散布矩阵,  $c$  为类数,  $N$  为样本数(如, 基因个数)。

$$S_B^p = \sum_{k=1}^c \sum_{n=1}^N \mu_{kn} v_k v_k^T \quad (11)$$

$$S_W^p = \sum_{k=1}^c \sum_{n=1}^N \mu_{kn} (x_n - v_k)(x_n - v_k)^T \quad (12)$$

$$S_T^p = \sum_{k=1}^c \sum_{n=1}^N \mu_{kn} x_n x_n^T \quad (13)$$

其中  $S_T^p$  是  $P$  维向量的混合散布矩阵。 $v_k (k=1, 2, \dots, c)$  是第  $k$  类  $C_k$  的中心;  $x_n (n=1, 2, \dots, N)$  是第  $n$  个样本的值。两者均为  $P$  维向量(如, 一个基因在不同时相点的表达数据)。

$$v_k = \{v_{k1}, v_{k2}, \dots, v_{kp}\}^T \quad (14)$$

$$x_n = \{x_{n1}, x_{n2}, \dots, x_{np}\}^T \quad (15)$$

$\mu_{kn}$  可以表示如下:

$$\mu_{kn} = \begin{cases} 0 & x_n \in C_k \\ 1 & x_n \notin C_k \end{cases} \quad (16)$$

就式(10)而言, PFS 有一些非常好的特性: 当聚类个数  $c$  值固定时, 若要求 PFS 值为最大, 必须  $tr(S_W^p)$  为最小, 所以使类内散布矩阵之迹  $tr(S_W^p)$  达到最小的聚类结果是最佳的; 另外, 随着  $c$  值的增大,  $tr(S_W^p)$  趋于下降, 并且  $tr(S_B^p)$  趋于上升, PFS 值是先上升而后不断下降, 即 PFS 可能在某一个  $c$  值处达到最大值, 而这正是最佳聚类个数。即: 这使得我们寻找最佳的分类数  $c$  等同于寻找最大的 PFS 值。

## 4 算法描述

PFS 模糊属性均值聚类法的具体步骤如下:

- (1) 输入基因表达数据  $X = \{x_1, x_2, \dots, x_n\} \subset R^P, (1 \leq n \leq N)$ ;
- (2) 为  $R^P$  选择距离矩阵, 限定  $m(m \geq 1)$  的值, 迭代标准  $\epsilon > 0$ , 初始化  $U^{(0)} \in M_{fc}$ , 然后在第  $l$  步进行,  $l=0, 1, 2, \dots$ ;
- (3) 计算聚类中心  $\{v_k^{(l)}\}, (k=1, 2, \dots, c)$ ;
- (4) 利用  $\{v_k^{(l)}\}$  和  $\{\mu_{kn}^{(l)}\}$  修正  $U^{(l)}$ ;
- (5) 用一个矩阵范数  $\| \cdot \|$  比较  $U^{(l)}$  和  $U^{(l+1)}$ , 如果  $\|U^{(l+1)} - U^{(l)}\| \leq \epsilon$ , 停止并转入(6), 否则转至(3);
- (6) 计算 PFS 聚类标准, 如果  $c \geq c_{\max}$ , 那么转至(7), 否则  $c = c + 1$ , 并运行(2);
- (7) 寻找最大 PFS 和最佳类数, 结束。

## 5 实验

为了验证 PFS 模糊属性均值聚类法中参数初始化的有效性及其算法的识别性能, 本文采用 MATLAB7.0 进行仿真实验, 并为此选取了一个标准的数据集: 血清刺激后成纤维细胞数据<sup>[8]</sup>, 进行测试分析。

血清数据中共有 517 个基因, 这些基因的 mRNA 水平随血清刺激而变化。在进行数据分析时, 人们往往依据经验或是表达谱颜色的形似性, 主观地把基因分为 10 类。图 1 是基于 PFS 判别函数的结果, 该图显示当类数是 8 时 PFS 值达到最大, 所以这 517 个基因应分成 8 类。

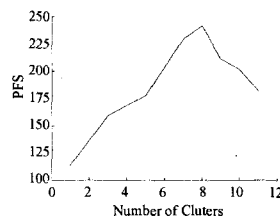


图 1 最大的 PFS 值对应于最佳的分类数

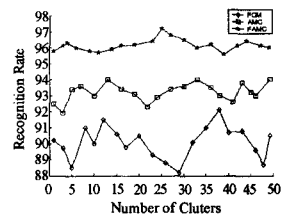


图 2 FCM, AMC, FAMC 三者的识别率

进一步,为了比较 FAMC 与 AMC 和 FCM 的识别性能,取实验参数: $c=8$ ;模糊度  $m(1 < m \leq 2)$ 。为了避免伪随机数、误差等因素对实验结果的影响,针对每个聚类参数和模糊度,3 种算法各计算 50 次,其实验结果如图 2 所示。

**结束语** 本文研究了基于伪 F 统计的 FAMC 算法,实验表明:该算法不仅有效地解决了参数的初始化问题;而且与 ACM 及 FCM 算法相比,FAMC 算法在稳健性和识别率上都有明显的改善。因此,本文算法对于处理基因芯片数据这类典型的高维小样本对象来说,是一种较为有效的分析工具。但是,目前关于基因表达数据分析方法的研究仍处于起步阶段,并且聚类结果的正确性在很大程度上仍旧依赖于基因表达数据的预处理。因此,今后我们应着重研究基因表达数据的标准化,使得标准化后的基因数据能够在同一个较小的范围内,从而可以在一定程度上提高聚类结果的正确性。

## 参考文献

[1] Karayiannis N B. Weighted fuzzy learning vector quantization

and weighted generalized fuzzy C-means algorithms. IEEE International Conference on Fuzzy Systems, 1996, 2: 773-779

[2] 程乾生. 属性均值聚类[J]. 系统工程理论与实践, 1998, 18(9): 124-126  
[3] 张媛祥. 一种稳健的聚类方法[J]. 数学的实践与认识, 2003, 33(8): 8-10  
[4] Dembele D, Kastner P. Fuzzy C-means method for clustering microarray data. Bioinformatics, 2003: 973-980  
[5] 高新波, 裴继红, 谢维信. 模糊 C-均值聚类算法中加权指数  $m$  的研究[J]. 电子学报, 2000, 28(4): 80-83  
[6] Vogel M A, Wong A C. PFS clustering method [J]. IEEE PAMI (S0001-0782), 1979, 3: 237-245  
[7] Culley T M, Wallace L E. Calculating F-Statistics [EB/OL]. (2001) [2004]. [Http://ib. Berkeley. edu/courses/ibl60/hl3a. html](http://ib.Berkeley.edu/courses/ibl60/hl3a.html)  
[8] Iyer, et al. [Http://www. sciencemag. org/feature/data/984559. shl](http://www.sciencemag.org/feature/data/984559.shl). 1999

(上接第 149 页)

|                        |     |              |             |   |    |
|------------------------|-----|--------------|-------------|---|----|
| Breast cancer          | 699 | 467(161,306) | 232(80,152) | 2 | 10 |
| Pima Indians Diabetics | 768 | 513(179,334) | 255(89,166) | 2 | 31 |
| Glass2v                | 214 | 143(52,91)   | 71(25,46)   | 6 | 10 |
| Ionosphere             | 351 | 234(84,150)  | 117(42,75)  | 2 | 34 |
| Iris3v12               | 150 | 100(33,67)   | 50(17,33)   | 3 | 4  |
| Bupa liver disorder    | 345 | 231(97,134)  | 114(48,66)  | 2 | 6  |

表 2 SVM 与 PDM 试验结果比较

| 问题                     | C   | C-SVM( $b=0$ ) |          |          | $b_1$   | PDM      |          |          |
|------------------------|-----|----------------|----------|----------|---------|----------|----------|----------|
|                        |     | Perr (%)       | Nerr (%) | Merr (%) |         | Perr (%) | Nerr (%) | Merr (%) |
| Breast cancer          | 10  | 12.5           | 4.61     | 8.56     | 0.1672  | 5        | 7.24     | 6.12     |
| Pima Indians Diabetics | 100 | 39.33          | 21.08    | 30.21    | -0.4634 | 24.72    | 26.51    | 25.62    |
| Glass2v                | 10  | 0              | 0        | 0        | 0.0253  | 0        | 0        | 0        |
| Ionosphere             | 100 | 45.24          | 2.67     | 23.91    | 8.056   | 0        | 20       | 10       |
| Iris3v12               | 100 | 0              | 0        | 0        | -0.711  | 0        | 0        | 0        |
| Bupa liver disorder    | 100 | 35.42          | 27.27    | 31.35    | 0.0389  | 33.33    | 27.27    | 30.3     |

## 5 结果及分析

表 2 中,参数  $C$  是在 1, 10, 100, 1000, 10000 中利用验证取得最有利的数据。由表 2 可知,6 个试验数据中有两个数据 Glass 和 Iris 使用 C-SVM 方法与 PDM 方法具有相同的效果,因为它们的错分率都是 0;其他 4 个数据 PDM 方法不仅平衡了正类和负类的错分率,而且降低了错分率。这就说明提出的方法不仅能保持 C-SVM 好的分类效果,而且能够平衡和减少两类的错分率。此结果并不是偶然的,因为毕竟考虑了密度因素,所以我们相信对于大样本不平衡该效果会更好。

**结束语** 本文提出的调整方法有如下特点:

1) 不需要修改原始样本信息,仅根据原始样本点投影特征分布密度的估计,纠正了由 C-SVM 训练的偏置  $b$ 。试验结果表明,本方法不仅能保持标准支持向量机的良好性能,而且能平衡甚至减少错分率。

2) 巧妙地把高维样本点的分类问题转化为一维来处理,利用 Parzen 窗估计法增强了理论根据,减少了以往方法的盲目性。

3) 本方法适合于任何形式的不平衡数据,包括样本容量

差异和分散程度差异。

本方法不足之处是本文仅在少数几个数据上验证了提出方法的有效性,没能应用到大量样本的不平衡数据。今后的研究应考虑把本方法推广到多类分类问题以及实际应用上。

## 参考文献

[1] Nello C, John S T. An introduction to support vector machines and other kernel based learning methods. Cambridge: Cambridge University Press, 2000  
[2] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent Data Analysis, 2002, 6(5): 429-449  
[3] Chew H G, Bonger R E, Lim C C. Dnal nu-support vector machine with error rate and training size biasing // Proceedings of the 25th IEEE International Conference Acoustics, Speech, and Signal. Piscataway, USA, IEEE, 2001: 1269-1272  
[4] Chawla N V, Bowyer K W, Hall L O, et al. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(3): 321-357  
[5] Ling C, Li C. Data mining for direct marketing problems and solutions // Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998: 73-79  
[6] Kubat M, Matwin S. Addressing the curse of imbalanced datasets // One-sided Sampling Proceedings of the Fourteenth International Conference on Machine Learning. Nashville: Tennessee, 1997: 178-186  
[7] Rehan A, Stephen K, Nathalie J. Applying support vector machines to imbalanced datasets // Fifteenth European Conference on Machines Learning. Berlin: Springer-Verlag, 2004: 39-50  
[8] 郑恩辉, 李平, 宋执环. 不平衡数据挖掘: 类分布对支持向量机的影响. 信息与控制, 2005, 34(6): 703-708  
[9] Lin Y, Lee Y, Wahba G. Support vector machines for classification in nonstandard situations. Machine Learning, 2002, 46(2): 191-202  
[10] 贾银山, 贾传炎. 一种加权支持向量机分类算法. 计算机工程, 2005, 31(12): 23-25  
[11] 边肇祺, 张学工. 模式识别. 北京: 清华大学出版社, 2000