一种适用于基因表达数据的特征加权 FCM 算法*

袁正午, 魏 荣, 叶明星
(重庆邮电大学 计算机科学与技术学院, 重庆 400065)

摘 要: 针对 FCM 算法应用于基因表达数据分析时存在的局限性,提出一种特征加权自适应 FCM 算法。该算法在 FCM 算法的基础上引入数据集预处理机制,可依据数据集的分布特征自适应地获取分类数目和初始聚类中心,并通过 ReliefF 算法实现特征权值的自动确定。同时,新算法考虑了不同属性对分类贡献的差异,在 FCM 算法中引入特征权重。将算法应用于真实基因表达数据集,实验结果表明,算法能够自适应地确定聚类数目、获得稳定性较好的聚类结果,而且具有较高的聚类精度。

关键词: 基因表达数据; 预处理算法; 类间熵; 加权模糊聚类
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2010)07-2483-03
doi:10.3969/j.issn.1001-3695.2010.07.022

Feature weight-based FCM clustering algorithm for gene expression data

YUAN Zheng-wu, WEI Rong, YE Ming-xing
(*Dept. of Computer Science & Technology, Chongqing University of Posts & Telecommunications, Chongqing* 400065, *China*)

Abstract: In view of fuzzy C-means algorithm having limitation for gene expression data analysis, this paper proposed a new feature weighted fuzzy C-means algorithm. The algorithm based on FCM algorithm could self-adaptive obtain the number of clusters and initial centerpoints according to the distribution characteristic of data set by pre-processing mechanism, and employed RefiefF algorithm to determine the weight for each feature. To consider the particular contributions of different feature, introduced feature weight in fuzzy clustering algorithm, applied the algorithm to gene expression data, experimental results illustrate that the proposed method can self-adaptive determine the number of clusters, obtain the clustering results with higher stability and clustering accuracy.

Key words: gene expression data; pre-processing algorithm; inter-cluster entropy; weighted fuzzy clustering

0 引言

基因芯片技术的发展可以同时监测成千上万个基因的表达情况[1],同时基因芯片技术的广泛应用也使基因表达数据(gene expression data)爆炸性增长。如何有效分析这些数据已成为当前生物信息学面临的主要问题之一。

目前,聚类方法已广泛地应用于基因表达数据分析。聚类分析的首要目标是将表达谱相似的基因归纳成类,然后重点关注那些可能参与某些生物过程的基因群,对这些基因群进行生物学注释,同时获得新的生物学知识。

常用于基因表达数据聚类分析的有层次聚类、自组织图(SOM)聚类、K-均值聚类、模糊 C 均值(FCM)聚类等。其中 FCM 聚类方法因其理论上可靠、算法简单,更符合生物学意义,而被广泛应用于基因表达数据分析[2],但 FCM 算法也存在参数依赖性强、对初始聚类中心和噪声敏感的缺陷。此外,FCM 算法隐含假定待分析样本的各维特征对分类的贡献均匀,不考虑各个特征对分类的不同影响。然而在基因表达数据分析中,由于构成数据对象的各维特征来自不同组织或不同观测条件,存在精确度、可靠性的不同,对分类的贡献也是有差别的。因此,FCM 聚类方法应用于基因表达数据聚类分析时,存在一定的局限性。

针对 FCM 算法的局限性, 本文在 FCM 算法的基础上引入数据集预处理机制。通过对数据集的预处理获得能反映数据集实际分布的分类数目、初始聚类中心和数据的特征权重;然后将预处理算法中获取的聚类数目、初始聚类中心和特征权重与 FCM 算法结合,提出一种适用于基因表达数据聚类分析的算法——特征加权自适应 FCM 算法。将算法应用于真实的基因表达数据集,实验表明,算法能够自适应地确定聚类数目,获得较稳定的聚类结果和较高的聚类精度。

1 数据集预处理算法

1.1 基本概念

设待聚类数据集为 p 维向量空间 X,$X \in R^p$,数据集 X 包含 n 个样本,每个样本可以表示为 p 维向量:$x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$,$1 \leqslant i \leqslant n$。样本 x_i 各维度上的权重用向量 W 表示:$W = (w_1, w_2, \cdots, w_p)$。$C_j (1 \leqslant j \leqslant c)$ 用于表示数据集划分的 c 个类别;N_j 表示划分到第 j 个类中的样本数目;$V = (v_1, v_2, \cdots, v_c)$,

收稿日期: 2009-12-31; **修回日期**: 2010-01-20 **基金项目**: 国家"863"计划资助项目(2007AA12Z226)

作者简介:袁正午(1968-),男,湖南人,教授,博士,CCF 高级会员,主要研究方向为空间定位、数据挖掘;魏荣(1986-),男,宁夏人,硕士研究生,主要研究方向为生物信息学、数据挖掘(weirongpt@yahoo.com.cn);叶明星(1984-),男,湖北人,硕士研究生,主要研究方向为生物信息学.

表示 c 个聚类中心,第 i 类中心为 $v_i=(v_{i1},v_{i2},\cdots,v_{ip})$,$1\leqslant i\leqslant c$。

定义 1 Reny 熵。对于取值连续的样本空间 X,其概率密度函数为 $f_X(x)$,样本空间 X 的 Reny 熵[3]表示为

$$H_R(X)=\frac{1}{1-a}\lg\int f_X{}^a(x)\,\mathrm{d}x,a>0,a\neq 1$$

特别地,当 $a=2$ 时,称为二次 Reny 熵:

$$H_R(X)=-\lg\int f_X{}^2(x)\,\mathrm{d}x \tag{1}$$

当使用函数 $f_X(x)$ 估计聚簇 C_i 的概率密度时,式(1)就表示聚簇 C_i 的熵,记为

$$H(C_k)=-\lg V(C_k)$$

定义 2 类间熵。定义 1 中熵的计算是在类内进行的,称之为类内熵。可以依照文献[3]的式(12)定义两个类的类间熵:

$$H(C_i,C_j)=-\lg V(C_i,C_j)$$

参考文献[5],聚簇概率密度函数取

$$\hat{f}(x)=\frac{1}{N_k h^n}\sum_{i=1}^{N_k}\psi((x-x_i)/h) \tag{2}$$

核函数 $\psi(x)$ 选择辅对称的多变量高斯函数:

$$\psi(x)=\exp(-\|x\|^2/2)/(2\pi)^{n/2} \tag{3}$$

用聚簇概率密度的估计式(2)和(3)代替式(1)中的密度函数 $f_X(x)$,可以得到

$$V(C_i,C_j)=\frac{1}{(2\pi)^P N_i{}^2 N_j{}^2 h^{2p}}\sum_{i=1}^{n}\sum_{j=1}^{n}M(x_i,x_j)\exp\left(-\frac{(x_i-x_j)^{\mathrm{T}}(x_i-x_j)}{2h^2}\right)$$

$$M(x_i,x_j)=\begin{cases}1,x_i\in C_i,x_j\in C_j/x_i\in C_j,x_j\in C_i\\0,\text{其他}\end{cases}$$

这样就可以计算任意两个类的类间熵。类间熵描述了两个簇之间的关系,如果 C_i 和 C_j 分得比较开,则 $V(C_i,C_j)$ 比较小,从而 $H(C_i,C_j)$ 比较大;反之 $V(C_i,C_j)$ 比较大,而 $H(C_i,C_j)$ 比较小。

1.2 预处理算法

由于 FCM 算法采用随机的方法来选取初始聚类中心,且需要预先指定分类数目,算法存在迭代次数相对较多、聚类结果不稳定、对参数依赖性强的缺陷。同时 FCM 算法隐含假定待分析样本的各维特征对分类的贡献均匀,不考虑各个特征对分类的不同影响。针对 FCM 算法存在的上述局限性,本文结合初始聚类中心优化选取方法[4]、基于类间熵的簇合并方法[5]和 ReliefF 技术[6],提出一种数据集预处理算法。使用预处理算法处理数据集,可自适应获取 FCM 算法所需的初始聚类中心和分类数目,同时可获得加权 FCM 算法需要的数据特征权值。

数据集预处理算法的步骤如下:

a)从样本集中随机抽取 p 个数据对象作为初始聚类中心[4](p 的数目大于实际聚类数目),使用 K-means 算法聚类得到 p 个聚簇。

b)去除类成员数小于等于 $N_{\min}$(最小簇内样本数目,可取 1)的聚簇,得到 t 个聚簇(避免孤立点被选为聚类中心)。

c)通过基于类间熵的簇合并算法,合并剩余 t 个聚簇。根据类数与熵值的变化情况,取得能反映数据空间分布特征的聚类数目和作为 FCM 算法起始聚类中心的代表点。

d)基于步骤 c)的聚类结果,结合 ReliefF 技术对数据进行特征加权。

基于类间熵的簇合并算法步骤:

a)计算 t 个聚簇中不同类的类间熵,合并类间熵最小的两个类,类的总数减少一个,并重新标记每类的样本。重复这个步骤,直到剩两个类为止。

b)分析类数目与类间熵的变化情况,寻找陡变点。根据熵的物理意义,如果在某个时刻计算的两个类最小的类间熵比之前最小的类间熵显著改变,这实际上就是一种由无序到有序或由有序到无序的突变,这时各类分得最开,对应的类数就是所要聚的类数[5]。

c)分析类间熵陡变点对应的聚类结果,选择与各子类中距离聚类中心最近的数据对象作为代表点,用代表点作为 FCM 算法的起始聚类中心。

基于 ReliefF 算法进行特征加权的步骤如下:

a)对于 $\forall x_i$,首先从步骤 c)中陡变点对应的聚类结果中找出 R 个与 x_i 同类的最近邻样本 $h_j(j=1,2,\cdots,R)$,然后在每个与 x_i 不同类的子集中找出 R 个最近邻的样本 $m_{lj}(j=1,2,\cdots,R,l\neq\mathrm{class}(x_i))$。

b)设 $\boldsymbol{D}_h$ 为 $p\times 1$ 矩阵,表示 h_j 与 x_i 在特征上的差异:

$$\boldsymbol{D}_h=\sum_{j=1}^{R}\frac{|x_i-h_i|}{\max(X)-\min(X)}$$

设 $\boldsymbol{D}_m$ 为 $p\times 1$ 矩阵,表示 m_{lj} 与 x_i 在特征上的差异:

$$\boldsymbol{D}_m=\sum_{l\neq\mathrm{class}(x_i)}\frac{P(l)}{1-P(\mathrm{class}(x_i))}\sum_{j=1}^{R}\frac{|x_i-m_{lj}|}{\max(X)-\min(X)}$$

其中:$P(l)$ 为第 l 类出现的概率。

c)ReliefF 算法中特征权重 W 由下式更新:

$$W=W-\boldsymbol{D}_h/R+\boldsymbol{D}_m/R$$

如此重复若干次,收敛后即可得到特征集中每一维特征的权重[7,8]。

2 加权 FCM 算法

2.1 FCM 算法

给定数据集 X 和分类数目 c,模糊 C 均值聚类问题可以表示为如下的目标函数求极值的问题:

$$\min\{J_1(\boldsymbol{U},V)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}{}^m\sum_{k=1}^{p}|x_{jk}-v_{ik}|^2\}$$

$$\text{s.t. }\boldsymbol{U}\in M_{fc}$$

其中,$\boldsymbol{U}=[\mu_{ij}]_{c\times n}$ 是模糊划分矩阵,有

$$M_{fc}=\{\boldsymbol{U}\in R^{c\times n}\mid\mu_{ij}\in[0,1];\sum_{i=1}^{c}\mu_{ij}=1,\forall j;0<\sum_{j=1}^{n}\mu_{ij}<n,\forall i\}$$

为样本集 X 的模糊 C 划分空间;m 为模糊加权指数,控制聚类分析的模糊程度。

2.2 加权 FCM 算法

显然,FCM 算法没有考虑样本不同属性对分类贡献的差异,认为每一维属性特征对分类的贡献是均匀的。事实上在基因表达数据分析中,由于构成数据对象的各维特征来自不同组织或者不同的观测条件,存在精确度、可靠性的不同,对分类的贡献也是不同的。本文考虑属性特征对聚类贡献的差异,提出基于特征加权的 FCM 算法。

对于给定的数据集 X 和分类数目 c,将样本的属性特征引入目标函数。基于特征加权的模糊 C 均值聚类问题可以表示

成如下的目标函数求极值问题:

$$\min\{J_2(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\sum_{k=1}^{p}w_k|x_{jk}-v_{ik}|^2\}$$
$$\text{s.t. } U\in M_{fc}$$

其中:w_k 表示样本第 k 维特征的权重,m 为模糊加权指数。

2.3　适用于基因表达数据分析的特征加权自适应 FCM 算法

基于以上的分析,给出适用于基因表达数据分析的特征加权自适应 FCM 算法:

a)数据集预处理。通过预处理算法处理待分类数据集,依据数据集实际的分布特征,获取 FCM 算法所需的分类数目、初始聚类中心,同时获得数据的特征权重。

b)将预处理算法中获取的聚类数目、初始聚类中心和特征权重与 FCM 算法结合,设计出适用于基因表达数据分析的特征加权自适应 FCM 算法。

新算法在 FCM 算法基础上作出以下两点改进:

a)引入数据集预处理机制。通过对数据集预聚类,可自适应获取 FCM 算法中需要人为指定的分类数目,避免了 FCM 算法对参数的依赖性。同时使用预聚类结果中的代表点代替 FCM 算法中随机选取的初始聚类中心,不仅可降低随机选取初始聚类中心引起的聚类结果的不稳定性,而且可减小噪声对聚类精度的影响。

b)考虑不同属性对聚类贡献的差异,在 FCM 算法的聚类准则中引入权重。

3　实验与结果

本文选择 Yeast 细胞周期数据[9]作为检测算法性能的数据集。该数据集包含酵母细胞在两个细胞周期 17 个时间点的基因表达数据,对应细胞周期的五个不同阶段,可根据基因达到峰点所在细胞周期的不同将该子集划分为五个聚类。本文选用文献[10]中提取的一个注释完全的子集,称为 Cho384,它是 384 个在不同时间点达到峰值的基因,在 17 个时间点的表达数据。

实验 1　验证算法检测数据集分类数目的能力。设置抽样数 p 为 10 和 15,通过 K-means 算法聚类,分别得到 10、15 个簇;然后基于类间熵合并簇(窗宽 h 取 0.5)。图 1、2 显示了聚类数目与类间熵的变化关系,可以看出,当聚成 5 类时,两种抽样数的类间熵变化异常,出现陡变点,而在其他类数时变化相对平稳,因此可以确定最终聚类的数目应为 5 类,与实际分类数相符。实验说明,用本文提出的预聚类算法能正确地确定聚类数。

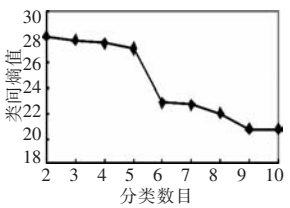

图 1　p=10 时聚类数目与类间熵的关系

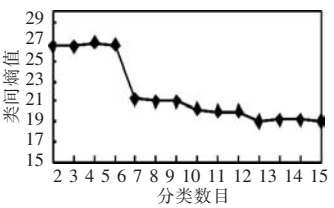

图 2　p=15 时聚类数目与类间熵的关系

实验 2　比较 FCM 算法与特征加权 FCM 算法聚类结果的稳定性和精度。分别运行 10 次 FCM 算法和特征加权 FCM 算法(模糊参数 m 取 2),比较两种算法的调整 rand 指数,如图 3 所示。

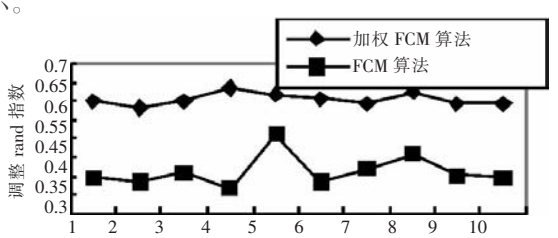

图 3　Cho384 数据集 10 次聚类的调整 rand 指数

由于传统的 FCM 算法初始中心是随机选取的,导致聚类结果存在较大的不稳定性,而基于预处理算法的特征加权 FCM 算法不但可以自适应地确定分类数目,而且能够选取有代表性的样本作为初始聚类中心,使得算法的聚类结果较为稳定。图 3 验证了这个结论。

另外,由于考虑了不同属性对聚类贡献的差异,降低噪声对聚类的影响,使特征加权 FCM 算法的聚类精度较 FCM 算法有较大提高。从图 3 中可看出,加权 FCM 算法对应的调整 rand 指数基本上都在 0.6 左右,而 FCM 算法大部分聚类结果 rand 指数位于 0.5 以下。

4　结束语

本文在分析基因表达数据集特征和 FCM 算法的基础上,提出一种用于基因表达数据聚类分析的特征加权自适应 FCM 算法。算法不仅能够自适应地确定聚类数,获得较稳定的聚类结果,而且可以显著地提高聚类的正确率。在真实的基因表达数据集上的测试证明了上述结论。该算法可作为基因表达数据实际分析中方便、有效的工具。

参考文献:

[1] SCHENA M,SHALON D,DAVIS R W,*et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray [J]. Science,1995,270(5235):467-470.

[2] BEZEDEK J C,HATHAWAY R J,SABIN M J,*et al*. Convergence theory for fuzzy C-means: counter-examples and repairs [J]. IEEE Trans on System Man and Cybernetics-Parts B: Cybernetics,1987,17(5):873-877.

[3] GOKCA E,PRINCIPE I C. Information theoretic clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2002,24(2):158-171.

[4] 毛韶阳,李肯立. K-means 初始聚类中心优化算法研究[J]. 重庆邮电大学学报:自然科学版,2007,19(4):422-425.

[5] 王洪春,彭宏. 一种基于熵的聚类算法[J]. 计算机科学,2007,34(11):178-179.

[6] KONONENKO I. Estimating attributes: analysis and extensions of relief[C]//Proc of the 7th European Conference on Machine Learning. Berlin: Springer,1994.

[7] 李洁,高新波,焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报,2006,34(1):89-92.

[8] 罗建军,管涛,冯博琴. 一种基于加权特征的可能模糊聚类方法[J]. 计算机应用研究,2006,23(6):52-54.

[9] CHO R J,CAMPBELL M J,WINZELER E A,*et al*. A genome-wide transcriptional analysis of the mitotic cell cycle[J]. Mol Cell,1998,2(1): 65-73.

[10] YEUNG K Y,FRALEY C,MURUA A,*et al*. Model-based clustering and data transformations for gene expression data [J]. Bioinformatics,2001,17(10):977-987.