doi:10.3969/j.issn.1672-5565.2014.02.05

ILLUMINA Golden Gate DNA 甲基化芯片的 KL-FCM 聚类分析

张 林,石 玥,汪 菲,李 琪,万苏磊,王雪松*
(中国矿业大学信息与电气工程学院,江苏 徐州 221116)

摘 要:DNA 甲基化作为一种重要的表观遗传修饰,其甲基化水平被发现与疾病的发生发展密切相关,对其进行聚类分析有希望发现新的疾病亚型并建立有效的疾病预测预后方法。传统的聚类分析方法之一模糊 C-均值(FCM:Fuzzy C-means)适用于特征空间呈球形或椭球形分布的场景,缺乏普适性。而 Illumina Golden Gate 平台通过计算基因的各甲基化位点的甲基化百分比描述其甲基化程度,其值位于(0,1)之间,服从混合贝塔分布,不能直接采用 FCM 进行聚类分析。鉴于此,本文提出基于 KL 特征测度的 KL-FCM 聚类算法,采用各样本间的 K-L 距离作为样本划分时的度量准则。最后,本文基于 KL-FCM 算法实现 IRIS 测试数据集和基因的 DNA 甲基化水平数据的聚类分析。实验结果表明该方法可以以更低的计算负荷获得优于 k-均值(k-means)和传统 FCM 的分类效果。

关键词:模糊 C 均值;ILLUMINA DNA 甲基化芯片;K-L 距离
中图分类号:TP181 **文献标志码**:A **文章编号**:1672-5565(2014)-02-106-04

KL-FCM clustering analysis inIllumina golden gate DNA methylation microarrray

ZHANG Lin,SHI Yue,WANG Fei,LI Qi,WAN Sulei,Wang Xuesong*
(*School of Information and Electrical Engineering China University of Mining and Technology*,*Xuzhou Jiangsu* 221116,*China*)

Abstract:DNA methylation is an important epigenetic modification, which has been found to be closely related to the occurrence and development of disease. Clustering analysis of DNA methylation is expected to find novel subtype of disease or novel method of prediction and prognosis. Fuzzy C-means (FCM) is one of the common clustering methods. However it is more suitable in the condition that the feature space follows spherical or elliptical distribution, which makes it lack in universality. Illumina Golden Gate platform describes the methylation level based on the methylation percentage of each locus in each gene, and it is in (0,1), which follows beta mixture distribution. Thus we can not adopt FCM for clustering directly. This paper introduces the KL-FCM clustering method, which calculates the K-L distance of samples as partition measure. The KL-FCM is used to cluster the IRIS test dataset and some DNA methylation profile data. The validation results show that KL-FCM, with less computational load, can get better clustering performance than k-means and traditional FCM clustering methods.

Keywords:Fuzzy C-means;DNA methylation expressionmicroarray;K-L distance

聚类分析即将具有相似特性的对象归为一类,作为一种预处理的方法,在模式识别、图像处理、医疗诊断、生物信息等领域都是非常重要的一项技术。传统的模糊 C-均值(FCM)即为一种应用广泛的聚类算法。FCM 是一种基于划分的聚类算法,其思想使得被划分到同一簇的对象之间相似度最大,而不同簇之间的相似度最小[1]。目前,FCM 及其变例算法大多以欧几里德距离度量各样本点与 C-聚类中心值之间的差异,因而该算法对于样本数据在特征空间呈球形或椭球形分布的情形可以保证良好的聚类效果,但缺乏普遍适用性。

DNA 甲基化作为一种重要的表观遗传修饰,不同甲基化位点的甲基化水平可能与疾病甚至癌症的发生发展密切相关[2]。Illumina Golden Gate DNA 甲基化芯片可同时检测 800 多个基因、超过1 500 个 CpG 位点的甲基化水平,基于该数据,可开发各种聚

收稿日期:2014-03-14;**修回日期**:2014-03-27.
基金项目:中国博士后基金面上项目(2012M511336、2012M511335);江苏省大学生创新创业训练计划;霍英东教育基金会青年教师基金(121066)资助。
作者简介:张林,女,副教授,研究方向:生物信息处理、机器学习;E-mail:lin.zhang@cumt.edu.cn.
*通信作者:王雪松,女,教授,研究方向:机器学习、生物信息处理;E-mail:wangxuesongcumt@163.com.

类算法鉴别可用于疾病诊断和预后的甲基化位点。但 Illumina Golden Gate DNA 甲基化芯片测量的是各基因的甲基化位点的甲基化百分比,其取值为(0,1),服从混合贝塔分布[3],不利于直接使用 FCM 算法。基于 FCM 算法对 DNA 甲基化水平数据开展聚类分析,需要在其特征空间中构造普适能力较强的衡量方法来度量两个样本矢量之间的差异。本文以 K-L 特征测度取代通常采用的欧几里德距离重构 FCM 算法,对基因的 DNA 甲基化水平数据开展聚类分析,有效地改善了 FCM 算法适用于球形或椭球形分布的样本数据的限制。

2　FCM 聚类算法

2.1　算法原理

FCM 聚类算法,即众所周知的模糊 ISODATA,采用隶属度确定每个数据点属于某个聚类的程度。它是普通 C 均值算法的改进。普通 C 均值算法对于数据的划分是硬性的,1974 年,Dunn 从硬性普通 C-均值中导出的一种对数据的柔性的模糊划分,利用类内加权平均误差和函数定义了 FCM 算法的目标函数[4],后来 Bezdek 引入模糊加权指数 m,对 FCM 算法加以归纳推广完善[5]。总结 FCM 的中心思想,FCM 是通过目标函数的迭代优化来实现集合的划分[6]。

设基因的 DNA 甲基化水平数据中包括 n 个样本,分为 k 个模糊类,提取了 s 个用于分类的特征向量。则 FCM 算法可以求出一个隶属矩阵,表示第 j 个样本隶属于第 i 类的概率,并且,一个样本的隶属度的和总等于 1:

$$\sum_{i=1}^{k} u_{ij} = 1, \forall j = 1,\cdots,n。 \tag{1}$$

FCM 的目标函数可表示为:

$$J(U,c_1,\cdots,c_k) = \sum_{i=1}^{k}\sum_{j}^{n} u_{ij}^m d_{ij}^2。 \tag{2}$$

这里 c_i 表示第 i 个模糊类的聚类中心,$d_{ij} = \| c_i - x_j \|$,$\| \cdot \|$ 是 R^s 上任一内积导出的范数,为加权指数。

FCM 算法即是求出使目标函数 $J(U,c)$ 达到最小值的隶属矩阵 U 和聚类中心 $c = \{c_1,c_2,\cdots,c_k\}$,其中 $\forall k \in \{k \mid 1 \le k \le n\}$,$\forall i \in \{i \mid 1 \le i \le k\}$。为此,以式(1)作为约束条件构造目标函数,如式(3)所示:

$$\bar{J}(U,c_1,c_2,\cdots,c_k,\lambda_1,\lambda_2,\cdots,\lambda_n) = J(U,c_1,c_2,\cdots,c_k) + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{k} u_{ij} - 1\right) = \sum_{i=1}^{k}\sum_{j}^{n} u_{ij}^m d_{ij}^2 + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{k} u_{ij} - 1\right)。 \tag{3}$$

这里,$\lambda_j, j=1,\cdots,n$ 是与式(1)的 n 个约束式相关的拉格朗日算子,利用拉格朗日乘子法,求得使式(2)达到极小值的必要条件,如式(4)所示:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, \qquad u_{ij} = \frac{1}{\sum_{h=1}^{k}\left(\frac{d_{ij}}{d_{hj}}\right)^{\frac{2}{m-1}}}。 \tag{4}$$

由上述 FCM 算法的原理可知,FCM 通过一个简单的迭代过程实现聚类分析。因此可按如下步骤实现:

①初始化隶属矩阵 U,使其满足式(1)所列约束条件;

②根据式(4)计算 k 个聚类中心 $c_i, i=1,2,\cdots,k$;

③根据式(2)计算目标函数的值,若求出的值已达到收敛,则算法停止。

④根据式(4)计算新的隶属矩阵 U,返回(2)。

2.2　模糊加权指数 m

在 FCM 算法中,模糊加权指数 m 也是一个重要的参数。一方面,它影响着目标函数的凹凸性,另一方面,它又控制着聚类的模糊性。文献[5]指出"参数 m 控制着模糊类间的分享程度",显然,m 的取值对模糊聚类的性能存在着重要的影响,要实现 FCM 算法就必须选择一个合适的 m 值。关于 m 值对 FCM 算法的影响,有如下定理[7]:

定理 1 对于 $m \in (1,\infty]$ 的 FCM 算法有:

当 $m=1$ 时,FCM 算法变成硬 c 均值(HCM)聚类算法;

当 $m \to 1^+$ 时,FCM 算法以概率 1 退化为 HCM 算法;

当 $m \to \infty$ 时,FCM 算法的聚类结果最模糊,丧失划分功能,隶属矩阵 $U=[\frac{1}{s}]$。

有关如何选择加权指数 m 的问题,Bezdek 在研究中发现,取 $m=2$ 时,FCM 算法具有明确的物理意义,但对不同的应用背景,选择相同 m 的值并不合适[7-9]。

本文参考文献[10]中所述方法进行参数 m 的在线优化,定义模糊目标函数如式(5)所示:

$$\mu_{C_i} = 1 - e^{-Kn_i}。 \tag{5}$$

其中,K 是常数;n_i 是聚类 i 所包含的数据个数。

同时,定义模糊约束函数如式(6)所示:

$$\mu_{C_i} = \left(1 + c\sum_{1}^{n_i} d_{ik}^2\right) - 1。 \tag{6}$$

其中,d_{ik} 为第 k 个数据点到第 i 个聚类中心的距离;

c 则为一个常数。

根据不同的 m 取值,由式(5)和式(6)计算相应的 μ_{D_i},如式(7)所示:

$$\mu_{D_i}=\min[\mu_{G_i}(m),\mu_{C_i}(m)]。\quad(7)$$

则 $\mu'_{D_i}=\max[\mu_{D_i}(m)]$ 所对应的 m 值为优选值,即第 i 次聚类时,所应选取的 m 值。

3 KL-FCM 聚类算法

3.1 对于传统 FCM 中距离测度的改进

传统 FCM 算法的特点是采用类的均值作为一个聚类的代表点,这只有在数据的自然分布呈现球形或者近似球形情况下,才可能达到理想的聚类效果。但 DNA 甲基化微阵列数据处于(0,1)中,并呈现复杂的混合贝塔分布[11-12],如图 1 所示为基于 Illumina GoldenGate 微阵列技术测量的某正常膀胱样本的 DNA 甲基化表达水平的直方图及分布拟合。从图中可以看出,样本在各个甲基化位点的表达水平近似服从贝塔分布,只是不同的样本,贝塔分布的参数选择不同。本文针对传统 FCM 算法的局限性,采用 K-L 特征测度衡量各样本间 DNA 甲基化差异程度。

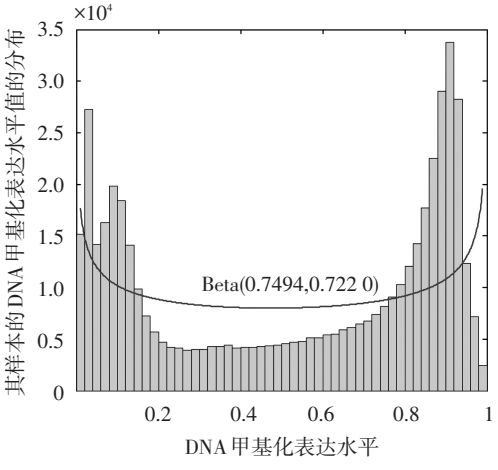

图 1 DNA 甲基化表达数据分布拟合

Fig.1 Curve:fitting of DNA methylation level distribution

K-L 特征测度,也称 K-L 距离,是统计独立性的最佳测度,常用来度量两个分布之间的差异程度。两个分布 $p_1(x)$ 和 $p_2(x)$ 之间的 K-L 距离定义为:

$$D(p_1(x)\parallel p_2(x))=\int p_1(x)\log\frac{p_1(x)}{p_2(x)}\mathrm{d}x。\quad(8)$$

设两个样本 $x_i=(x_{i1},x_{i2},\cdots,x_{is})$,$x_j=(x_{j1},x_{j2},\cdots,x_{js})\in R^s$,K-L 距离矢量定义如式(9)所示。

$$\mathrm{d}(x_i,x_j)=\sum_{h=1}^{s}\mathrm{d}(x_{ih},x_{jh})。\quad(9)$$

式中各分量 $\mathrm{d}(x_{ih},x_{jh})=x_{ih}\ln\frac{x_{ih}}{x_{jh}}+x_{jh}\ln\frac{x_{jh}}{x_{ih}}$,$h=1,2,\cdots,s$,即为两个样本中相应特征上的 K-L 距离。

3.2 基于 KL 距离的 FCM 聚类算法(KL-FCM)

传统的 FCM 采用欧几里德距离计算目标函数,这对样本数据在特征空间呈球形或椭球形分布的场合有效。然而,基因的 DNA 甲基化水平数据中各样本在各位点的表达值呈现贝塔分布。因此,本文对传统 FCM 算法加以重构,采用了 K-L 距离计算各样本间的差异度。

同样通过简单的迭代过程实现基于 K-L 距离的聚类分析:

①初始化隶属矩阵 U,使其满足式(1)所列约束条件;

②根据式(4)计算 k 个聚类中心 c_i,$i=1,2,\cdots,k$;

③根据式(7)选取优化的 m 值;

④根据式(2)计算目标函数的值,式中 d_{ij} 采用式(9)计算。若求出的值已达到收敛,则算法停止。

⑤根据式(4)计算新的隶属矩阵 U,返回 2)。

4 实验结果分析

4.1 IRIS 测试数据集

为了验证本文提出算法的有效性,首先选取国际通用 UCI 数据库中的 IRIS 测试数据集[13]。IRIS 数据集是从加拿大加斯帕半岛上的鸢尾属花朵中提取的地理变异数据,其中包含 150 个样本,分属于鸢尾属下的三个亚属,分别是山鸢尾(Iris setosa)、变色鸢尾(Iris versicolor)和维吉尼亚鸢尾(Iris virginica),每个亚属 50 个样本,每个样本包含 4 个属性,分别是萼片长度、萼片宽度、花瓣长度和花瓣宽度,单位均为厘米。

表 1 中给出基于 k-means、传统 FCM 和 KL-FCM 算法在 IRIS 测试数据集中根据各个样本的 4 个属性对各鸢尾花朵的亚属进行聚类分析的结果,运行环境为:Pentium Dual-Core CPU E5400@ 2.70 GHz、2G 内存、Matlab R2009b。可以看出,基于 K-L 距离的 FCM 聚类算法能够获得比 k-means 以及传统 FCM 聚类算法更低的错误率。

表 1 kmeans、传统 FCM 和基于 K-L 距离的 FCM 算法实验结果比较

Table 1 Comparison results of FCM and KL-FCM

算法	聚错样本个数	运行时间(秒)	平均错误率	达到收敛循环次数
k-means	14	0.236 745	9.3%	
传统 FCM	12	0.187 306	8.0%	49
KL-FCM	6	0.132 924	4.0%	68

4.2　基因的 DNA 甲基化水平数据

本文选取 Illumina GoldenGate 微阵列产生的基因的 DNA 甲基化水平数据[14],该数据中包括 217 个样本上的 1 473 个甲基化位点的甲基化水平,包括 5 个取自膀胱的样本,85 个取自血液的样本,12 个取自大脑的样本,3 个取自子宫颈的样本,11 个取自头部和颈部的样本,6 个取自肾的样本,53 个取自肺部的样本,19 个取自胎盘的样本,18 个取自胸膜的样本以及 5 个取自小肠的样本。本文根据各个样本在 1 473 个甲基化位点上的甲基化水平对其来源进行了聚类分析。

表 2 中同样给出基于 k-means、传统 FCM 和 KL-FCM 算法在上述基因的 DNA 甲基化水平数据中的聚类结果,运行环境与 IRIS 测试数据集的相同。从实验结果可以看出,基于 K-L 距离的 FCM 聚类算法在基因的 DNA 甲基化水平数据的聚类分析中同样能够获得比 k-means 以及传统 FCM 算法更低的错误率,并且计算负荷更低。

表 2　kmeans、传统 FCM 和基于 K-L 距离的 FCM 算法实验结果比较

Table 2　Comparison results of FCM and KL-FCM

算法	聚错样本个数	运行时间(秒)	平均错误率	达到收敛循环次数
k-means	34	0.193 252	15.7%	
传统 FCM	26	0.050 297	12.0%	16
KL-FCM	18	0.030 084	8.3%	13

5　结　论

基于 K-L 特征测度的 FCM 聚类算法,改善了传统的 FCM 聚类算法适用于球形或椭球形特征空间的应用限制。IRIS 测试数据集和基因的 DNA 甲基化水平数据的聚类实验结果表明该方法的有效性,其在聚类分析的正确率、计算负荷等方面表现均优于 k-means 算法和传统的 FCM 聚类算法。

参考文献(References)

[1] MOHAN A, MOORTHY K. Early detection of diabetic retinopathy edema using FCM[J]. International Journal of Science and Research (IJSR), India, 2013,2(5):115-118.

[2] WEISENBERGER D. Characterizing DNA methylation alterations from the cancer genome atlas[J]. The Journal of clinical investigation, 2014, 124(1): 17-23.

[3] KUAN P, WANG S, ZHOU X, et al. A statistical framework for Illumina DNA methylation arrays [J]. Bioinformatics, 2010, 26(22): 2849-2855.

[4] DUNN J. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters [J]. Cybernetics and Systems, 1973, 3(3): 32-57.

[5] BEZDEKJ. Pattern recognition with fuzzy objective function algorithms[M]. New York:PLENUM Press,1981.

[6] 陈佳妮, 段文英,丁徽. 模糊 C-均值聚类分析在基因表达数据分析中的应用[J]. 森林工程, 2010, 26(002): 54-57.
CHEN Jiani, DUAN Wenying, DING Hui. Application of fuzzy C-means clustering analysis in gene expression data analysis[J]. Forest Engineering, 2010, 26(002): 54-57.

[7] 高新波,李洁, 谢维信. 模糊 C 均值聚类算法中参数 m 的优选[J]. 模式识别与人工智能, 2000, 13(1): 7-11.
GAO Xinbo, LI Jie, XIE Weixin. Optimal choice of weighting exponent in a fuzzy C-means clustering algorithm[J]. Pattern Recognition & Artificial Intelligence, 2000, 13(1): 7-11.

[8] 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.
GAO Xinbo, PEI Jihong, XIE Weixin. A study of weighting exponent m in a fuzzy C-means algorithm[J]. ACTA Electronica Sinica, 2000, 28(4):80-83.

[9] 于剑,程乾生. 关于 FCM 算法中的权重指数 m 的一点注记[J]. 电子学报, 2003, 31(3): 478-480.
YU Jian, CHENG Qiansheng. A note on the weighting exponeng m in FCM algorithm [J]. ACTA Electronica Sinica, 2003, 31(3): 478-480.

[10] 刘宜平,沈毅. 一种参数 m 在线优化的 FCM 改进算法[J]. 电机与控制学报, 2000, 4(3): 164-167.
LIU Yiping, SHEN Yi. An advanced FCM algorithm with parameter m optimized on-line[J]. Electric Machines and Control, 2000, 4(3): 164-167.

[11] JI Yuan, WU Chunlei, LIU Ping, et al. Applications of beta-mixture models in bioinformatics[J]. Bioinformatics, 2005, 21(9): 2118-2122.

[12] ZHANG Lin, MENG Jia, LIU Hui, et al. A nonparametric bayesian approach for clustering bisulfate-based DNA methylation profiles[J]. BMC Genomics, 2012, 13(Suppl 6): S20.

[13] RICHARD O, HART P. Pattern recognition and scene analysis[M]. New York:Wiley,1973.

[14] HOUSEMAN E, CHRISTENSEN B, YEH R, et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions [J]. BMC Bioinformatics, 2008, 9(1): 365.