基于互信息量的生物信息数据特征标注方法

何红洲[1,2]，周明天[1]

(1. 电子科技大学计算机科学与工程学院 成都 611731; 2. 绵阳师范学院数学与计算机科学学院 四川 绵阳 621000)

【摘要】提出了一种用于排位特征变量的基于特征矩阵信息增益的无监督特征标注准则(IGC)及直接选择法(DS)、累积最大熵法(CEM)和最大信息增益法(IGM)3种新的特征过滤方法来降低聚类的复杂度。使用经典的QC或*K*-means聚类算法，在杆状病毒数据集(RSV)、混合血统白血病数据集(MLL)和急性白血病患者数据集(ALP)等3种不同的生物信息数据集上测试并对比了这些特征过滤方法和目前的偏差选择(VS)和基因修剃(GS)过滤方法对聚类结果的影响。试验结果表明，3种特征过滤方法在加速聚类过程及保持初始数据的聚类结构上都具有明显的优势。

关 键 词 特征标注; 特征过滤; 信息增益; Jaccard群落系数; 奇异值分解熵

中图分类号 TP391 文献标志码 A doi:10.3969/j.issn.1001-0548.2013.06.020

Feature Annotation Method of Biological Information Data Based on Mutual Information

HE Hong-zhou[1,2] and ZHOU Ming-tian[1]

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731;

2. College of Mathematics & Computer Science, Mianyang Normal University Mianyang Sichuan 621000)

Abstract A unsupervised feature annotation criterion-information gain criterion (IGC)-based on feature matrix information gain is proposed to rank the feature variable. According to this rank, three new feature filtering methods：direct selection (DS), cumulate maximum entropy (CEM), and information gain maximum (IGM) are given to reduce clustering complexity. The clustering results of these three filtering methods with two existing variance selection (VS) and gene shaving (GS) methods were tested and compared by using classic QC or *K*-means algorithm and three biological datasets: rod-shaped viruses (RSV), mixed-lineage leukemia (MLL), and acute leukemia patients (ALP). The experiment results show our feature filtering method has obvious superiority in accelerating the clustering procedure and preserving the clustering structure of initial data.

Key words feature annotation; feature filtering; information gain; Jaccard score; SVD-entropy

在临床诊断决策数据分析、序列分析、微阵列数据分析[1-2]及基因表达序列数据分析[3]等生命科学领域的多类别应用领域的分类和聚类分析中，会遇到抽样数据多达成百上千的包括噪音在内的特征维数的问题，将这些数据直接进行分析，一方面会将问题变得非常复杂(如计算的复杂性)；另一方面分析结果的准确性也会受到噪音的干扰。从而在分析前进行特征选择，即从大维数初始特征集合中选择少量而“有用”的特征子集就成为一个研究课题。这里的“有用”表示选择的特征(集)不仅能够保持初始数据的结构特性，而且仅使用选出的特征就能够容易将数据分成有意义的类，即来自同一类别的数据更“相似”，而来自不同类别的数据更“相异”。文献[4]回顾了生物信息的特征选择技术，其中最重要的两种技术是有监督的特征封装(feature wrapper)和无监督的特征过滤(feature filtering)技术。前者需要精心定义一个目标函数，并通过选择进行动态优化，使目标函数满足一定的条件或收敛；后者不需要参照已有的经验分类或定义任何目标函数，因而可能更难于实现，但它们具有几个重要的理论优势：一是当没有先验知识供参考时，不需要专家或数据分析师就可以预先形成好的结果；二是它降低了数据过度拟合的危险，这正是有监督学习的缺陷。

动植物的遗传基因、病毒基因和生物医学数据是生物信息数据库中的一个庞大的类别，这些数据具有维度高(特征数量大)，噪音种类庞杂而分散，以

收稿日期: 2012 – 06 – 26；修回日期：2013 – 05 – 16
基金项目: 四川省教育厅自然科学研究项目(12ZB070)
作者简介: 何红洲(1967 –), 男, 博士生, 副教授, 主要从事生物信息数据挖掘及复杂网络方面的研究.

及结构复杂的特性，要对这些数据进行有效分类，其任务首先是在高效地缩减数据维度的过程中有效地排除噪音的干扰，从而根据数据本身的特征信息对其进行分类。目前应用于上述生物信息数据领域的属于无监督特征过滤范畴的方法主要见于文献[5]中的是VS(variance selection)方法和文献[6]中的GS(gene shaving)方法。前者按偏差对特征进行排位，而后者则按第一主分量的最高排序选择特征，其实它们都是半监督的特征过滤方法，即在第二阶段仍然需要使用有监督的特征选择，而且其计算复杂度的降低会以破坏聚类结果的质量为代价，因而它们在排除噪音数据的干扰上存在着极大的局限性。

本文提出了一种新型的基于奇异值分解熵的信息增益的特征排位及相应的完全无监督特征过滤方法有效地克服了上述缺陷，从而能够在无监督的情况下对给出的生物信息数据进行更加有效地分类。

1　问题的一般化描述

设$\boldsymbol{Z}_{\text{ini}}=\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n\}$是具有$n$个实例或观测值的初始数据集，其中$\boldsymbol{Z}_j=(z_{1j}, z_{2j},\cdots, z_{dj})^{\mathrm{T}}(1\leqslant j\leqslant n)$是具有$d$个特征属性值的列向量，称$\boldsymbol{Z}=(z_{ij})_{d\times n}$为初始数据矩阵，其第$i$行表示$n$个数据在第$i$个特征$f_i$上的取值。特征选择的目标是：定义$\boldsymbol{D}=\{f_1, f_2,\cdots, f_d\}$的一个真子集$\boldsymbol{D}_{\text{s}}$，使得仅保留这$t_{\text{s}}=|\boldsymbol{D}_{\text{s}}|$个特征属性值的数据集$\boldsymbol{Z}^*=\{\boldsymbol{Z}_1^*, \boldsymbol{Z}_2^*,\cdots, \boldsymbol{Z}_n^*\}$能够被更简单而有效地分类。

2　信息增益准则(IGC)特征排位法

2.1　特征标注

特征标注即针对某种特定的应用(如对初始数据集进行分类)，使用某种方法将特征集$\boldsymbol{D}$中所有特征按其重要性程度标识出来，以确定在该应用中起作用的特征排序的过程。

设s_j表示初始数据集$\boldsymbol{Z}_{\text{ini}}$对应数据矩阵$\boldsymbol{Z}$的奇异值，从而$\lambda_j=s_j^2$为$d\times d$矩阵$\boldsymbol{P}=\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$的特征值[7]，定义表征数据集信息量的奇异值分解熵如下[8]：

$$E(\boldsymbol{Z})=-\sum_{j=1}^{d} w_j \log_d(w_j)=-\frac{1}{\log(d)}\sum_{j=1}^{d} w_j \log(w_j) \quad (1)$$

式中，$w_j=\lambda_j/\sum_{k=1}^{d}\lambda_k$ 表示λ_j的相对权重。熵值的取值区间为[0,1]，其中0表示数据集$\boldsymbol{Z}_{\text{ini}}$的意义可以由$\boldsymbol{Z}$的某一个特征列向量来唯一解释，1表示在解释数据集$\boldsymbol{Z}$的意义时，每一个特征向量所提供的信息量是均恒的。如图1所示给出了高熵分布(熵为0.868 7的灰色条)和低熵分布(熵为0.101 4的黑色条)的数据矩阵特征值分布比较。

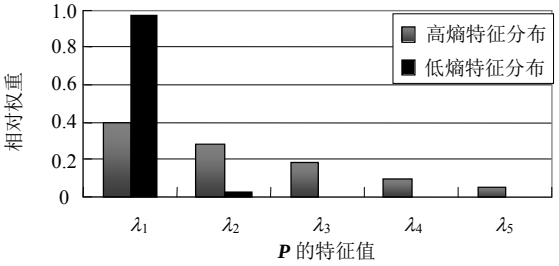

图1　不同熵值的两种特征值分布比较

为了描述特征集$\boldsymbol{D}$中的每个特征对初始数据集分类的影响和贡献程度，从而将每个特征在所有特征中的影响因子标注出来，受信息论中互信息描述两个随机变量相互影响各自随机程度概念的启发，针对$d\times n$列的初始数据矩阵$\boldsymbol{Z}$，定义$\boldsymbol{Z}^{(i)}$为$\boldsymbol{Z}$去掉第i个特征行后所得到的$(d-1)\times n$矩阵，而标注值G_i表示第i个特征行与初始数据矩阵$\boldsymbol{Z}$的互信息量，即去掉第i个特征行代替$\boldsymbol{Z}$后按式(1)计算的熵的减少量(注意这时d比原来少1)，这种减少量表示第i个特征属性对数据集的信息增益或信息损耗，即：

$$G_i=I(i;\boldsymbol{Z})=E(\boldsymbol{Z})-E(\boldsymbol{Z}^{(i)}) \quad (2)$$

设μ和σ分别为所有G_i的均值和标准差，对所有G_i将d个特征属性相对于区间$[\mu-\sigma, \mu+\sigma]$分为3组：增益组($G_i>\mu+\sigma$)、损耗组($G_i<\mu-\sigma$)和无偏组($G_i>\mu-\sigma$且$G_i<\mu+\sigma$)。增益组中的特征属性对揭示初始数据的信息结构是决定性的，因为去掉它们中的任意一个特征属性都会损失信息量，因此认为它们属于优选对象；损耗组中的特征属性对提示初始数据的信息结构产生了副作用，其影响可能会一致地分布到所有数据点，因此认为它们属于优舍(弃)对象；无偏组中的特征属性的标注值在一个正常的估值区间，因而从统计意义上讲，它们属于冗余特征，因此认为它们属于次舍(弃)对象。

2.2　特征过滤

按照上述的分析方法，给出下面3种特征过滤方法，其中第一个方法就是在增益组中按照G_i的排位(降排)选出前面的部分特征，这时初始数据矩阵$\boldsymbol{Z}$将由新的秩为r_{s}的(所选特征属性数)矩阵$\boldsymbol{Z}_{\text{s}}$所代替。

1) 直接选择法(DS)：按照G_i值的降排位选出前r_{s}个特征属性.

2) 累积最大熵法(CEM)：选出G_i值最大的第一个特征，在剩余的特征中选出第二个特征，使其与第一个特征所形成的2特征n数据点的矩阵的奇异值分解熵(按式(1)计算，这时d=2)最大，继续$r_{\text{s}}-2$轮这样的循环(注意d相继变为3到r_{s})，直到所有的r_{s}个特

征选择完毕。算法如下(Φ表示空集)：

① $\boldsymbol{D}_S=\Phi$;

②按式(1)和式(2)计算$\boldsymbol{D}$中各f_i 的增益G_i，并令$G_k=\max\limits_{1\leqslant i\leqslant d}\{G_i\}$;

③ $\boldsymbol{D}_S\leftarrow\boldsymbol{D}_S\cup\{f_k\}$;

④若$|\boldsymbol{D}_S|<r_s$, $f_k=\arg(\max\limits_{f\in\boldsymbol{D}-\boldsymbol{D}_S}\{E(\boldsymbol{D}_S\cup\{f\})\})$，转步骤③；(这里$E(\boldsymbol{D}_S\cup\{f\})$表示从初始数据矩阵$\boldsymbol{Z}$中仅取$\boldsymbol{D}_S\cup\{f\}$集中的特征行后构成的新矩阵的奇异值分解熵)

⑤输出$\boldsymbol{D}_S$。

3) 最大信息增益法(IGM)：第一个特征的选择如CEM，去掉选出的特征后用式(1)和式(2)重新计算新的G_i值，再从中选出新的G_i值最大的特征作为第二个选择的特征，继续下去直到所有r_s个特征选择完毕。算法如下：

① $\boldsymbol{D}_S=\Phi$;

②按式(1)和式(2)计算$\boldsymbol{D}$中各f_i 的增益G_i, 并令$G_k=\max\limits_{1\leqslant i\leqslant d}\{G_i\}$;

③ $\boldsymbol{D}_S\leftarrow\boldsymbol{D}_S\cup\{f_k\}$, $\boldsymbol{D}\leftarrow\boldsymbol{D}-\{f_k\}$;

④若 $|\boldsymbol{D}_S|<r_s$, 转步骤②;

⑤输出$\boldsymbol{D}_S$。

以上3种方法同时也反映了特征属性的3种不同的排位原则。

3 试 验

3.1 Jaccard群落系数

为了表示无监督的特征过滤方法对聚类结果的影响，首先给出表征物种分类有效性的Jaccard群落系数(Jaccard分数)。设F表示来源于数据集世系结构中的家族分类，C表示来源于用某种特征过滤算法选出特征后对数据集所作的分类，分类使用QC[9]或K-means[10]聚类算法，n_{11}、n_{10}和n_{01}分别表示两种分类共有的种类数，仅在F中的种类数和仅在C中的种类数，则Jaccard分数为：

$$J=\frac{n_{11}}{n_{11}+n_{01}+n_{10}} \tag{3}$$

Jaccard群落系数反映了算法分类和家族分类的交并比，其值介于0(不匹配)到1(完美匹配)之间。

3.2 数据集及结果分析

为了验证本文方法对生物信息数据的分类效果，在3组数据集上对本文方法与两种经典的特征过滤方法VS及GS的性能进行了比较。在对第一组数据集的分类中，本文方法还进一步区分了构成杆状病毒氨基酸的物化特性；第二组和第三组数据集是典型的高维度(成千上万的基因探针)生物医学信息数据，对它们进行特征过滤进一步体现本文方法对提高分类效率的显著性。

1) 杆状病毒数据集(RSV)。

RSV(rod-shaped viruses)[11]是由61个杆状病毒所构成的数据集，它们影响多种农作物，如烟叶、西红柿、黄瓜等的生长和发育。用这些病毒的表皮蛋白中的18种氨基酸成分作为18个特征测度，根据家族分类结构，这些病毒分为4类：游牧病毒组(hordeviruses)3个、烟草脆裂病毒组(tobraviruses)6个、烟草花叶病毒组(tobamoviruses)39个和真菌传杆状病毒组(furoviruses)13个。取r_s的初始值为3，并对所有特征测试了各种过滤方法的性能。如图2所示给出了5种特征过滤方法的特征排位(横坐标括号里的值是所有18种氨基酸成分按式(1)和式(2)计算的G_i值，G_i的均值μ= –0.001 49及标准差σ=0.007 42)。显然，DS方法的特征排位与G_i的大小一致(折线是严格单调增的函数曲线)，DS、CEM及IGM方法在前3个特征的选择上是一致的(参看图2的前3列)。图2也画出了VS方法及GS方法的特征排位曲线。特别注意，尽管VS选出的第一个特征属性与本文方法一致，但选出的第二个和第三个特征ASX(代表天冬酰胺ASN及天冬氨酸ASP)和GLX(代表谷氨酰胺GLN及谷氨酸GLU)在本文的方法中排位靠后；而GS方法选出的第一个特征与本文方法刚好相反。

有趣的是：位于排位前三的甘氨酸GLY、苏氨酸THR和赖氨酸LYS并非表皮蛋白的主要成分，它们共占表皮蛋白的18%左右。进一步的研究表明无论是从大小，还是极性亦或是电荷特性都不足以把这3种氨基酸从其余的氨基酸中区分开。然而，由于GLY、THR、LYS和甲硫氨酸MET(CEM方法排位第四种氨基酸)分别代表了不同的功能分组，因此可以得出结论：CEM过滤方法与携带不同物理化学特性的氨基酸的选择是一致的，而VS和GS不能反映这样的物化特性。

用Jaccard群落系数对氨基酸的选择作评价。使用QC聚类算法对RSV数据集的61个病毒进行聚类，如图3所示的DS、CEM及VS方法在取r_s =3～18的性能。由该图可以看出，当对具有18个特征的61个病毒进行分类时，其分类的Jaccard群落系数仅为0.6(All features曲线)；而使用DS、CEM后其分类的精度远大于0.6，特别是当r_s=3时，J>0.9，而VS方法的Jaccard系数为0.38。如果选择3个以上的特征，CEM方法是唯一性能最好的方法。使用K-means算

法也能得到相似的结果。

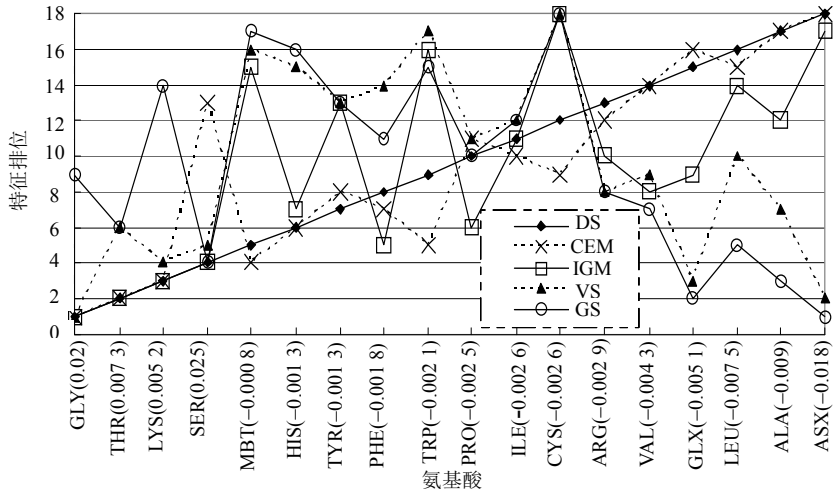

图2　RSV病毒表皮蛋白18种氨基酸成分使用5种特征过滤方法的特征排位曲线

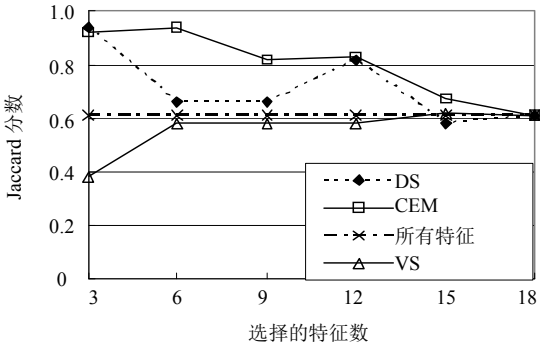

图3　3种特征过滤方法对RSV病毒数据集聚类质量的影响(使用参数σ=0.5的QC聚类算法)

2) 混合血统白血病数据集MLL。

MLL数据集[12]涉及3类白血病(leukemia)：带染色迁移的混合血统白血病(MLL)、传统的急性淋巴细胞白血病(ALL)及急性骨髓性白血病(AML)。试验记录了72个病人的Affymetrix U95A基因切片，共包括125 82个人体基因探针，其中20人被诊断为MLL，24人被诊断为ALL，28人被诊断为AML。文献[12]表明这3类白血病可以按照某种基因表达结构来划分。然而当以一种无监督的方式选择8 700个基因时，其聚类效果远远次于有监督的方式下选择500组基因的效果，而这500组基因很容易地就把癌症病人区分开。如图4所示对比了DS方法与VS方法对聚类结果的影响，聚类使用参数K=3的K-Means算法，并取100轮的均值，从图4的ALL features曲线可以看出，K-means算法的渐近Jaccard系数为J=0.426，VS方法破坏了聚类质量而DS方法将使用排位靠前的250～500(按照增益组得到的r_s的最大值为254)个特征就将Jaccard系数提高到0.7～0.8之间，使聚类质量得到了明显的改善。

3) 急性白血病患者数据集ALP。

ALP数据集[13]来源于具有两种类型白血病ALL和AML的38个病人(ALL有27人，小孩；AML有11人，成人)的骨髓吸出物，其中ALL分为T细胞白血病和B细胞白血病两组，而AML又分为已接受治疗和没有接受治疗两组，针对每一位病人，Affymetrix基因切片记录了681 7个人体基因探针，本文的任务就是将用681 7×38阶基因表达矩阵表示的38位病人分为4个正确的组别。

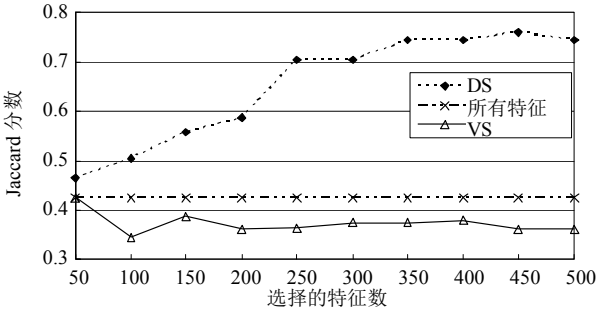

图4　两种特征过滤方法对MLL数据集聚类质量的影响(使用参数K=3的K-means聚类算法)

仍然使用前面提到的QC算法，如图5所示表示一旦所选择的基因超过100，DS和CEM方法就能取得一个好的结果。如用CEM和DS方法选择120～200个基因，则聚类结果的Jaccard系数就大于0.707，该值为所有特征的渐近Jaccard分数线。

综上，可以得出以下的结论：

1) DS和CEM方法一方面大大降低了聚类的复杂度(选出的特征仅为初始特征集中的一个很小部分)，另一方面也提高了聚类质量(即Jaccard分数提高

了)，而VS方法在降低聚类复杂度的同时也破坏了聚类的质量(即Jaccard分数降低了)。

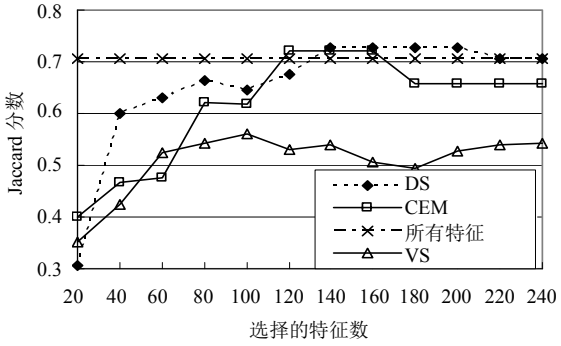

图5 3种特征过滤方法对ALP数据集聚类质量的影响(使用参数σ=0.5的QC聚类算法)

2) 本文的特征过滤方法能够有效地加速聚类过程。一方面，本文方法能够快速选择出对分类有效的特征(因为这些选择出的特征排位靠前)；另一方面本文方法能够在选出的特征占原有特征比例很小(RSV：16.67%；MLL：1.99%～3.97%；ALP：1.76%～2.93%)的情况下取得较好的分类效果(RSV：Jaccard分数>0.9；MLL：Jaccard分数>0.7；ALP：Jaccard分数>0.707)，这在很大程度上降低了分类计算的复杂度。

3) 结合信息增益的分组及图3～图5的曲线的趋势，很容易确定出在DS和CEM方法中所选特征数r_s的最优值，这在VS及GS方法中根本不可能实现。

4 结 束 语

使用生物数据的分类方法来研究生命信息的医学特征是生物信息学研究的一个重要课题，而特征选择又是分类前对生物数据进行预处理的一个重要手段，本文针对目前的VS和GS特征过滤方法的局限，提出了一种基于互信息量的无监督特征过滤方法对杆状病毒数据集、混合血统白血病数据集和急性白血病患者数据集进行分类，试验表明该方法在3种不同的生物信息数据集上都取得了较好的分类效果。

参 考 文 献

[1] KERR G, RUSKIN H J, CRANE M, et al. Techniques for clustering gene expression data[J]. Computers in Biology and Medicine, 2008, 38(3): 283-293.

[2] WANG Yan-fei, YU Zu-guo, ANH V. Fuzzy C-means method with empirical mode decomposition for clustering microarray data[C]//IEEE International Conference on BIBM. Hong Kong, China: IEEE, 2010: 192-197.

[3] WANG Hai-ying, ZHENG Hui-ru, AZUAJE F. Clustering-based approaches to SAGE data mining[J]. BioData Min, 2008(1): 5-16.

[4] SAEYS Y, INZA I, LARRANAGA P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23(19): 2507-2517.

[5] ZHU Dong-xiao. Semi-supervised gene shaving method for predicting low variation biological pathways from genome -wide data[J]. BMC Bioinformatics, 2009, 10(Suppl): 54 -65.

[6] SHENG Jin-hua, DENG Hong-wen, CALHOUN V, et al. Integrated analysis of gene expression and copy number data on gene shaving using independent component analysis[J]. IEEE/ACM Trans Comput Biol Bioinform, 2011, 8(6): 1568-1579.

[7] SANTOS A R, SANTOS M A, BAUMBACH J, et al. A singular value decomposition approach for improved taxonomic classification of biological sequences[J]. BMC Genomics, 2011, 12 (Suppl 4): 11-25.

[8] PONNAPALLI S P, MICHAEL A S, CHARLES F V, et al. A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms[J]. PLoS One, 2011, 6(12): e28072.

[9] HORN D, GOTTLIEB A. Algorithm for data clustering in pattern recognition problems based on quantum mechanics [J]. Physical Review Letters, 2002, 88(1): 1-4.

[10] VARSHAVSKY R, HORN D, LINIAL M. Clustering algorithms optimizer: a framework for large datasets[C]// Bioinformatics Research and Applications, Lecture Notes in Computer Science. Berlin Heidelberg: Springer-Verlag, 2007(4463): 85-96.

[11] LLOYD S P. Least squares quantization in PCM[J]. IEEE Transaction on Information Theory, 1982, 28(2): 129-137.

[12] TSENG G C. Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data[J]. Bioinformatics, 2007, 23 (17): 2247-2255.

[13] FAUQUET C, DESBOIS D, FARGETTE D, et al. Classification of furoviruses based on the amino acid composition of their coat proteins[C]//Viruses with Fungal Vectors. Edinburgh: Association of Applied Biologists, 1988, 19-38.

[14] ARMSTRONG S A, STAUNTON J E, SILVERMAN L B, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia[J]. Nature genetics, 2002(30): 41-47.

[15] GOLUB T R, SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring[J]. Science, 1999(286): 531-537.

编 辑 张 俊