·方法研究·

基因表达数据中加权 SAM 法的基因选择和分类预测研究

任雨冬,陆震,李婧惟,刘艳
哈尔滨医科大学卫生统计学教研室,黑龙江 哈尔滨 150081

摘要: **目的** 使用高斯核函数和欧式距离函数改进微阵列显著分析法(significance analysis of microarray,SAM)得到 MSAM1 法(modified significance analysis of microarray-1,MSAM1)和 MSAM2 法(modified significance analysis of microarray-2,MSAM2),与 SAM 法、Relief 法、支持向量机递归特征消除法(support vector machine recursive feature elimination,SVM-RFE)进行对比,评价在基因表达数据中 MSAM1 法、MSAM2 法的基因选择和分类预测能力。 **方法** 从 Bioconductor 中的 golubEsets 包获得 leukemia 数据集(Golub 等人给出了该数据集所包含的 50 个差异基因),运用 R 软件实现 5 种算法,分别用正确率和 ROC 曲线下面积即 *AUC* 值评价基因选择能力和分类预测能力,用 Kruskal-Wallis *H* 检验比较 5 种方法的正确率和 *AUC* 值的组间差异,进一步的两两比较采用 SNK-*q* 检验。 **结果** 正确率和 *AUC* 值均表现为 MSAM1 和 MSAM2 最优,SAM 和 SVM-RFE 法次之,Relief 法排在最后;5 种方法的组间差异有统计学意义($H=150.333$,$P<0.0001$ 和 $H=293.2579$,$P<0.0001$),两两比较结果显示虽然 MSAM1 和 MSAM2 之间差异无统计学意义($P>0.05$),但两种方法与其他 3 种方法之间差异均有统计学意义($P<0.05$)。 **结论** 用高斯核函数和欧式距离函数改进的加权 SAM 法提高了 SAM 法的基因选择和分类预测能力,在实际基因表达数据的应用中可以得到更为稳定的分析结果。

关键词: SAM;基因表达数据;基因选择;分类预测
中图分类号:R181.2 **文献标识码:**A **文章编号:**1006-3110(2020)12-1537-04 DOI:10.3969/j.issn.1006-3110.2020.12.036

Gene selection and classification prediction of weighted SAM method in gene expression data

REN Yu-dong, LU Zhen, LI Jing-wei, LIU Yan
Department of Health Statistics, Harbin Medical University, Harbin, Heilongjiang 150081, *China*
Corresponding author: LIU Yan, E-mail: liuyan@ ems.hrbmu.edu.cn

Abstract: Objective The modified significance analysis of microarray-1 (MSAM1) method and the modified significance analysis of microarray-2 (MSAM2) method are obtained by using the Gaussian kernel function and the Euclidean distance function to improve the significance analysis of microarray (SAM) method, respectively. The original SAM method, the support vector machine recursive feature elimination (SVM-RFE) method, and the Relief method were compared to evaluate the gene selection and classification prediction ability of the MSAM1 method and the MSAM2 method in gene expression data. **Methods** The leukemia data set was obtained from the golubEsets package in Bioconductor (Golub, et al. gave 50 differential genes contained in the data set). Five kinds of gene selection methods were implemented using R software. The gene selection ability and classification prediction capability were evaluated by the correct rate and the area under the ROC curve, namely, the AUC value. Kruskal-Wallis H test was used to compare the between-group differences in the correct rate and AUC value among the five methods, and SNK-q test was used for further pairwise comparison. **Results** Both the correct rate and the AUC value were optimal for MSAM1 and MSAM2, followed by the SAM and SVM-RFE methods, and the Relief method was ranked last. The between-group differences among the five methods were statistically significant ($H=150.333$, $P<0.0001$; $H=293.2579$, $P<0.0001$). The results of the pairwise comparison showed that there was no statistically significant difference between MSAM1 and MSAM2 ($P>0.05$), but the differences between the above-mentioned two methods and the other three methods were statistically significant ($P<0.05$). **Conclusions** The weighted SAM method modified by Gaussian kernel function and Euclidean distance function improves the gene selection and classification prediction ability of SAM method, and can obtain more stable analysis results in the application of actual gene expression data.

Keywords: significance analysis of microarray; gene expression data; gene selection; classification prediction

基金项目:黑龙江省自然科学基金(LH2019H005)
作者简介:任雨冬(1994-),男,黑龙江海伦市人,硕士在读,主要从事医学领域统计学方法的应用与研究工作。
通信作者:刘艳,E-mail:liuyan@ ems. hrbmu. edu. cn。

基因表达数据广泛地应用于生物医学领域癌症细胞分子水平上的分类和预测研究[1]。由于高维度和小样本量这两个问题,促进了基因表达数据分析中基因选择方法的发展。基因选择方法大致分四类:过滤法、封装法、嵌入法和混合法。过滤法因其结果直观且容易理解、有很好的泛化能力以及与分类器相互独立等优点一直被广泛使用[2],其中微阵列数据显著分析法(significance analysis of microarray, SAM)是基于过滤法应用最广泛的基因选择方法之一。为了降低离群值(高通量测序实验中因生物学或技术错误会使数据中存在大量离群值)对差异基因识别的影响,本研究将两种加权函数改进的 SAM 法即 MSAM1 和 MSAM2[3]应用到实际的基因表达数据即 leukemia 数据集[4]中,从基因选择和分类预测两个方面评价改进 SAM 法的能力。

1 原理与方法

1.1 SAM SAM 法由 Tusher 于 2001 年开发[5],是根据基因表达的变化相对于重复测量的标准差为每个基因打分,通过使用基因特异性校正 t 检验来识别基因表达数据中的差异基因[3],即在传统 t 检验公式的基础上加上一个较小的正数 S_0(取值通过样本数据计算),从而避免将表达水平和变异程度较低的无生物学意义的基因识别为差异表达基因[6-9]。SAM 法计算公式为 $d_i=\frac{\bar{x}_i-\bar{y}_i}{s_i+s_0}$,其中 $\bar{x}_i$ 和 $\bar{y}_i$ 分别是两组中基因 i 的表达水平,$\bar{x}_i=\sum_{j=1}^{n_1}x_{ij}/n_1$,$\bar{y}_i=\sum_{j=1}^{n_2}y_{ij}/n_2$;$n_1$ 和 n_2 分别是两组样本量;s_0 是模糊因子,通过调节使 d_i 的变化系数最小;s_i 是基因特异性分散度,$s_i=\sqrt{a\{\sum_{j=1}^{n_1}(x_{ij}-\bar{x}_i)^2+\sum_{j=1}^{n_2}(y_{ij}-\bar{y}_i)^2\}}$ $a=(1/n_1+1/n_2)/(n_1+n_2-2)$。

1.2 加权 SAM 法 为降低离群值对差异基因筛选的影响,使用中位数代替均值并使用权重函数 w 来计算统计量,此时 $\tilde{s}_i$(改进后的 s_i)的计算公式为:$\tilde{s}_i=\sqrt{\sum_{j=1}^{n_1}w(x_{ij})(x_{ij}-median_j(x_{ij}))^2+\sum_{j=1}^{n_2}w(y_{ij})(y_{ij}-median_j(y_{ij}))^2}$ 相应的统计量 $\tilde{d}_i$(改进后的 d_i)的计算公式为 $\tilde{d}_o=\frac{\bar{x}_i-\bar{y}_i}{\tilde{s}_i+s_0}$。

1.2.1 高斯核函数(Gaussian kernel)加权 SAM 法(MSAM1) 高斯核函数(Gaussian kernel)是一种广泛应用的加权函数,权重是随着距离中心的距离增加而平稳减少到 0。权重函数为 $W(x_{ij};\mu_i,\sigma)=\frac{1}{\sigma}\phi(\frac{x_{ij}-\mu_i}{\sigma})$,其中 ϕ 表示的是标准正态分布的概率密度函数,$\phi(x)=e^{-x^2/2}/\sqrt{2\pi}$;$\mu_i$ 表示的是基因特异性参数,例如 $\mu_i=median_j(x_{ij})$;σ 表示标准差,依赖于基因表达数据的恒定值。

1.2.2 欧式距离(Euclidean distance)加权 SAM 法(MSAM2) 欧式距离的权重函数为 $w(x_{ij})=\frac{1}{\sum_k d_E(x_{ij},x_{iK})}$,其中 $d_E(x_{ij},x_{iK})$ 表示的是基因 i 的第 j 个和第 k 个样本之间的欧式距离。其原理是某基因的所有样本中,如果一个观测样本距离其他样本越远,这个样本被赋予的权重越小,使得异常值得到一个较小的权重,从而避免对差异基因的影响。

1.3 支持向量机(support vector machine, SVM) SVM 法由 Vapnik 于 1995 年提出[10],是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的有监督学习算法。SVM 在进行基因选择时通常需要结合其他算法来达到目的,目前使用最多的方法就是支持向量机递归特征消除法(SVM-RFE)[11-13]。SVM-RFE是基于支持向量机最大间隔原理的序列后退选择算法。它通过模型训练样本,然后对每个基因的得分进行排序,去掉得分最小的基因,然后将余下的基因再次进行模型训练,进行下一次迭代,最后选出需要的基因数。

1.4 Relief 法 Relief 法是由 Kira 和 Rendell 在 1992 年提出的[14],是一种基于过滤法的基因选择算法。Relief 系列算法根据基因对近距离样本的区分能力来评估重要基因,其基本思想为:重要基因应该使同类样本接近,使不同类样本远离。具体表现为基因的权重越大,该基因的分类能力越强;反之,该基因分类能力越弱。其中权重为负表示不相关。Relief 算法优点在于比较简单,但是运行效率高,对数据类型没有限制,并且结果也比较好,因此得到广泛应用。

1.5 统计学方法 运用 R 3.5.3 统计软件中的“samr”、“CORElearn”、“sigFeature”等软件包实现上述 5 种基因选择方法,分别用正确率和 ROC 曲线下面积即 *AUC* 值评价基因选择能力和分类预测能力,用 Kruskal-Wallis H 检验比较 5 种方法的正确率和 *AUC* 值的组间差异,进一步的两两比较采用 SNK-q 检验。其中,正确率=筛选出真实的差异基因的个数/全部差异基因个数,正确率越高则方法的基因筛选能力越好;同理,*AUC* 值越大说明方法的分类预测能力越好。

2 结 果

本研究从 Bioconductor 中的 golubEsets 包获得 leukemia数据集,数据集中包含 27 例急性淋巴细胞白血病(ALL)患者和 11 例急性髓细胞白血病(AML)患者的 7 129 个基因;依据 Golub 等[4]给出的 leukemia 数据集所包含的 50 个差异基因列表用来对 5 种方法的基因选择和分类预测能力进行评价。

2.1　*数据的离群值情况*　计算每个基因的 38 个数值的四分位数间距,基因表达量介于 1.5~3 倍四分位数间距的数值即为离群值,依据含离群值的比例定义含离群值基因,用以分析数据的离群值情况。分析结果显示(见图 1):左侧第一个直条表示,若将含离群值的比例超过 0.01 即 1%的基因当作含离群值基因,则 7 129个基因中 66%的基因即 4 705 个基因属于含离群值基因;右侧第二个直条表示,若将含离群值的比例超过 0.1 即 10%的基因当作含离群值基因,则 9.58%的基因即 623 个基因属于含离群值基因。

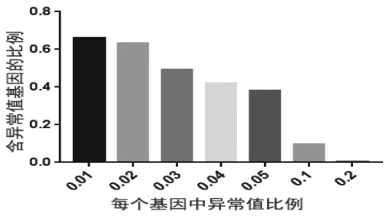


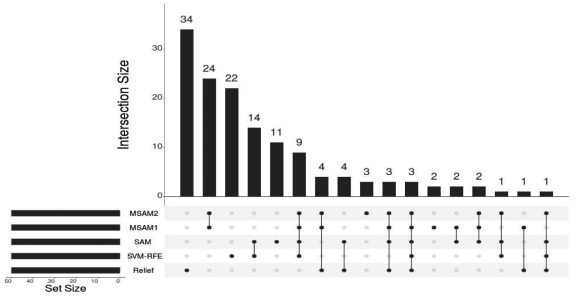
图 1　Leukemia 数据集中离群值比例

2.2　*5 种方法正确率和 AUC 值的比较*　5 种方法通过 5 折交叉验证计算迭代 100 次的正确率和 *AUC* 值(见表 1),正确率和 *AUC* 值均表现为 MSAM1 和 MSAM2 最优,SAM 和 SVM-RFE 法次之,Relief 法排在最后;5 种方法的组间差异有统计学意义($H=150.333$,$P<0.0001$ 和 $H=293.2579$,$P<0.0001$),两两比较结果显示虽然 MSAM1 和 MSAM2 之间差异无统计学意义($P>0.05$),但两种方法与其他 3 种方法之间差异都有统计学意义($P<0.05$)。

表 1　5 种方法的分析结果

方法	正确率($n=100$)		*AUC* 值($n=100$)	
	$\bar{x}\pm s$	*M* (*IQR*)	$\bar{x}\pm s$	*M* (*IQR*)
MSAM2	0.5964±0.1112	0.5851(0.5168~0.6809)	0.9339±0.0370	0.9331(0.9068~0.9638)
MSAM1	0.5679±0.1151	0.5542(0.4720~0.6456)	0.9304±0.0394	0.9349(0.9054~0.9606)
SAM	0.5090±0.2119[ab]	0.5064(0.3507~0.6823)	0.8673±0.0499[ab]	0.8629(0.8358~0.9038)
SVM-RFE	0.4820±0.2004[ab]	0.4698(0.3410~0.6262)	0.8601±0.0498[ab]	0.8556(0.8224~0.8957)
Relief	0.3241±0.1040[ab]	0.3241(0.2576~0.3904)	0.7546±0.0758[ab]	0.7587(0.6976~0.8203)

注:5 种方法的正确率和 *AUC* 值数据的总体均呈正态分布,但总体方差不齐;a 表示与 MSAM1 法相比差异有统计学意义($P<0.05$);b 表示与 MSAM2 法相比差异有统计学意义($P<0.05$)。



注:图中每个直条表示筛选出的差异基因个数,直条下面的●表示能筛选出该直条的差异基因的方法,每个方法标识●的直条的差异基因个数的总和都为 50。

图 2　5 种方法筛选差异基因异同情况的 upset 图

2.3　*5 种方法筛选出的差异基因的比较*　选取 5 种方法各自筛选出的排名在前 50 位的差异基因,与 Golub 等[4]给出的 leukemia 数据集所包含的 50 个差异基因列表进行对比,MSAM2、MSAM1、SAM、SVM-RFE 和 Relief 的正确率分别为 0.68(34/50)、0.64(32/50)、0.40(20/50)、0.28(14/50)和 0.22(11/50);*AUC* 值分别为 0.9741、0.9582、0.8754、0.8346 和 0.7891。将 5 种方法各自筛选出的排名在前 50 位的差异基因绘制成 upset 图(见图 2),5 种方法筛选出的相同差异基因有 3 个,每种方法筛选出不同于其他 4 种方法的差异基因个数表现为 MSAM2 有 3 个、MSAM1 有 2 个、SAM 有 11 个、SVM-RFE 有 22 个、Relief 有 34 个,综合说来 MSAM1 和 MSAM2 两种方法的能力更好。

3　讨　论

随着基因组测序水平技术的发展,高维度小样本量的数据促进了基因选择技术的发展。前文提到基因选择包含四类方法,在基因组数据分析中,大多数研究致力于开发基于过滤法的方法,因为该方法最简单,最快,计算效率最高。且过滤法有很大的灵活性,不仅可以与任何学习算法相结合,还可以与其他基因选择方法相结合,比如封装法,进而产生混合方法。这种混合方法可通过过滤法选择相关基因,剔除冗余基因,然后通过封装法验证这些基因,并确定了具有较高分类精度的最终特征集,从而提高了数据分析的效率。

SAM 法是基于过滤法应用最广泛的基因选择方法之一,但当基因中包含大量离群值时,通过 SAM 法进行基因选择时,差异基因的排名比预期的要低很多,甚至可能无法识别出差异基因,导致 SAM 法筛选差异基