第 36 卷 第 4 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.4 Computer Engineering February 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)04—0158—03 文献标识码：A 中图分类号：TP18

基于 ACO-SVM 的质谱数据分析

张 蓉[1,2]，冯 斌[1]
(1. 江南大学信息工程学院，无锡 214122；2. 江苏信息职业技术学院计算机工程系，无锡 214101)

摘 要：生物信息学应用领域存在高维小样本和内部空间疏散的特性，因而数据分析面临着巨大的挑战。基于此，在蚁群算法的搜索过程中将特征的信噪比作为先验信息，结合支撑向量用于筛选血清蛋白相关生物标记物，实验结果表明，该方法建立的癌症诊断模型取得了较好的分类性能测试仿真结果，敏感度和特异度分别达到 94%和 92.4%。

关键词：表面增强激光解析电离飞行时间质谱；蛋白质组学；蚁群优化算法；特征选择技术；生物标记物

Analysis of Mass Spectral Data Based on ACO-SVM

ZHANG Rong[1,2], FENG Bin[1]
(1. School of Information Technology, Jiangnan University, Wuxi 214122;
2. Department of Computer Engineering, Jiangsu College of Information Technology, Wuxi 214101)

【Abstract】The high dimensional and small sample sizes natures of bioinformatics pose a great challenge for many modeling problems. A novel method is raised that combines using SNR as prior information in the Ant Colony Optimization(ACO) searching process. Combined with support vector machines, it is applied to identify relevant serum proteomic biomarkers. Experimental results show that the proposed method has strong power in distinguishing cancer patients from healthy individuals, and yields up to 94% sensitivity and 92.4% specificity.

【Key words】Surface-Enhanced Laster Desorption/Ionization Time-Of-Flight Mass Spectrometry(SELDI-TOF-MS); proteomics; Ant Colony Optimization(ACO) algorithm; feature selection technology; biomarker

1 概述

疾病蛋白质组学主要研究寻找各种疾病的特异性标志蛋白质，进而应用于临床诊断和药物开发等，因此也常称为临床蛋白质组学(clinical proteomics)。

近几年表面增强激光解析电离化时间飞行质谱(Surface-Enhanced Laster Desorption/Ionization Time-Of-Flight Mass Spectrometry, SELDI-TOF-MS)技术的起步和发展为临床蛋白质组学中肿瘤的研究注入活力，综合了芯片微阵列技术与质谱两者的优点，大大提高了对蛋白质的鉴定能力，可用于生物标记物发现、鉴定和蛋白质谱的分析。SELDI-TOF-MS 技术已经被认为是在肿瘤诊断中非常有前景的技术方法[1]。

质谱图中的谱峰对应个体蛋白质，一组图谱中包含有成千上万的特征(每个特定质/荷比对应的表达丰度)，但是只有少部分的特征对应的是蛋白峰(图谱中局部最大峰值)，因此海量数据的分析面临相当大的挑战，而质谱分析就是利用计算机算法(如分类决策树、人工神经网络、支持向量机等生物信息学分析方法)分析这些多维海量数据，分辨出蛋白峰，检测出健康体和癌症体之间表达差异的未知蛋白峰(丰度上升或者下降)，建立蛋白质指纹图谱模型以进一步用于肿瘤标志物的筛查。

本文设计了一种基于 ACO-SVM 的差异蛋白峰选择方法，构建了血清蛋白相关性生物标记物提取、选择及识别框架，并在独立的肝癌测试集数据上进行了验证实验。

2 SELDI-TOF-MS 技术

简而言之，SELDI 蛋白芯片技术根据层析技术和质谱技术发展而来，是利用经过特殊处理的固相支持物或芯片的层析表面，根据蛋白质物理、化学性质的不同，选择性地从待测生物样品中捕获特定类型的蛋白和多肽，将其结合到芯片的固相层析表面上，经过原位清洗和浓缩后，结合飞行时间质谱，对所结合的多肽或蛋白质进行质谱分析，可以检测到不同蛋白质的相对表达丰度及分子量。

3 实验方法

3.1 质谱数据

肝细胞癌(Hepatocellular Carcinoma, HCC)是世界最常见恶性肿瘤之一，目前 HCC 发病率在我国呈上升趋势，其死亡率占恶性肿瘤死亡的前列。肝癌的早期诊断较为困难，目前临床上常用 AFP 作为肝癌的诊断指标，但 AFP 对肝癌诊断阳性率一般为 60%~70%，且在原发性肝癌诊断中也存在假阳性，因此缺乏可靠的早期诊断指标是其预后较差的主要原因。SELDI 蛋白芯片技术的问世使得寻找灵敏特异的生物学标志物以便于 HCC 的早期诊断成为可能。

本试验数据源于 HWR 实验室提供的 SELDI-QqTOF 肝癌研究质谱数据集[2]。该数据集总共包含 357 例血清样本，其中，肝癌患者样本 176 例，正常体样本 181 例，每个样本均由串联四极杆飞行时间质谱仪对结合在弱阳离子交换表面(Weak Cationic exchange, WCx2)芯片上的血清蛋白进行读取分析获得。

候选标记物选择的流程如图 1 所示。

基金项目：国家自然科学基金资助项目(60474030)
作者简介：张 蓉(1980－)，女，讲师、硕士研究生，主研方向：智能算法，模式识别；冯 斌，副教授
收稿日期：2009-09-16 **E-mail：**zoe_here@163.com

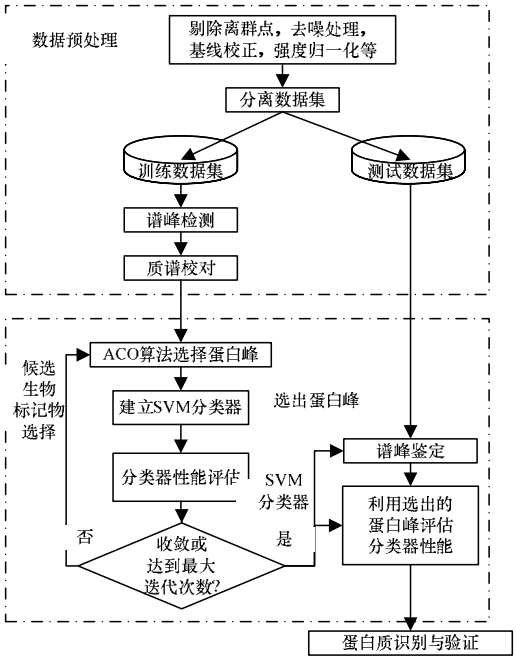


图 1　生物标记物选择流程

3.2　数据预处理

数据预处理包括剔除离群点、降维去噪处理、基线校正、强度归一化、谱峰检测、质谱校正等，缜密设计的预处理过程能够进一步增强质谱数据分析的性能，有利于生物标记的提取[3-4]。

3.2.1　降维预处理

在数据预处理中，分箱法(binning)是最简单的一种峰点检测和对齐的方法，它通过把测量的数据分组到各个箱子中达到数据降维的目的。本文对介于 1 000~11 500 的质荷比峰值，应用分箱法：设置每个箱子的宽度为 400 ppm，按每个箱子的平均强度进行滤波(smoothing)处理。

3.2.2　基线校正

在实验操作条件下，反映检测器噪音随时间变化的曲线称为基线，稳定的基线是一条直线。由于化学噪音或者离子过载，质谱数据通常会表现出一个变化的曲线，主要出现在低质荷比区域。基线校正(baseline correction)对于最小化背景噪声是非常重要的，如果没有经过充分的校正，则漂移的基线会导致电离强度严重变形。通过将 LOWESS(局部加权回归散点修匀法)滤波方法应用到质谱数据，使样条函数逼近回归适合数据，为每个测量值调整校正值，达到去除基线的目的。

3.2.3　归一化处理

归一化处理的目的是减少系统偏差，特别是大值特征对小值特征的影响。参照文献[4]介绍的方法，首先计算所有质谱数据的谱峰曲线下面积(AUC)，结果作为每个蛋白的丰度，再除以平均的 AUC 值来重新标量(即每个图谱的 AUC 值/总图谱的平均 AUC 值)。

3.2.4　谱峰检测

峰点检测即检测出质谱中局部最大值。在基线的讯号水平上，预设一个“ 阈值 ”，超过该值时，判别为峰可开始检测。一般采用下面 2 种方式判别峰讯号的变化：依照信号斜率的变化检测信号或是依照积分面积检测峰信号。考虑到质谱中噪音随质荷比线性上升，本文采用线性上升的阈值去噪，通过计算质谱强度斜率的符号函数，当值由正数变化为负数时视为峰点，同时去掉质荷比小于 1 000 的峰点。

3.2.5　质谱校正

质谱校准即在每个质谱图中和整个训练集之间，通过合并相邻峰来生成 *m/z* 窗口的方法来校正所有的谱峰。步骤如下：(1)定义 2 条阈值线，第 1 条由 2.5 线性下降至 1，第 2 条由 1.5 线性下降至 0.1；(2)筛选出位于第 1 条阈值线上的谱峰，将差异低于 2 个箱子或质量相差最多为 0.000 8 的相邻峰合并成一个 *m/z* 窗口；(3)筛选出介于 2 条阈值线之间的谱峰，把符合步骤(2)判断条件的谱峰加入相邻 *m/z* 窗口中；(4)剔除那些低于 5 个质谱图共有的峰点。

3.3　生物标记物选择

3.3.1　回归支撑向量机

回归支撑向量机(SVM)最吸引人的地方是采用了核化技术和结构风险最小化思想。SVM 学习得到的模型不但经验风险很小，而且泛化误差也很小，即结构风险最小，这也是 SVM 对于经典的神经网络的最大优势，对于小样本数据集建模问题，SVM 应该是最值得优先考虑的。

3.3.2　蚁群算法

蚁群算法(ant colony algorithm)是由意大利学者 Dorigo 于 20 世纪 90 年代初提出的，它是根据蚂蚁觅食原理而设计的一种群体智能算法[5]。该算法的基本思想如下：(1)一群蚂蚁随机从出发点出发，遇到食物，衔住食物，沿原路返回。(2)蚂蚁在往返途中，在路上留下外激素标志。(3)外激素将随时间逐渐蒸发(一般可以用负指数函数来描述)。(4)由蚁穴出发的蚂蚁，其选择路径的概率与各路径上的外激素浓度成正比。

这样，每只蚂蚁经过 *n* 次迁移后就得到一条一定长度的回路，再根据相关公式重新计算各条路径的外激素浓度，可进行下一步搜索。利用同样原理可以描述蚁群进行多食物源的觅食情况。

3.3.3　ACO-SVM 特征选择方法

ACO 算法用于获取最优谱峰集合，每只蚂蚁根据下列给出的转移概率函数从 *L* 个候选谱峰中选择出 *n* 个不同的特征组成特征向量：

$$\boldsymbol{p}_i(t)=\frac{(\tau_i(t))^{\alpha}\eta_i^{\beta}}{\sum_i(\tau_i(t))^{\alpha}\eta_i^{\beta}} \tag{1}$$

其中，$\tau_i(t)$ 代表第 *i* 个特征在 *t* 时刻的外激素量；η_i 代表先验信息(例如，基于单变量 *t*-统计的先验信息)；α和β则分别代表外激素和先验信息的权重因子。*t*=0 时刻，所有特征的外激素量的 $\tau_i(t)$ 设为一个常量。因此在第 1 次迭代中，每只蚂蚁只依据与先验信息对应成比例的概率从 *L* 维特征中选择 *n* 维不同的特征。S_j 代表第 *j* 只由 *n* 个特征构成的蚂蚁。使用 S_j 中的特征来构造 SVM 分类器并通过 10-交叉验证方法来计算分类准确率。再按照式(2)，依据 S_j 分类的性能更新 S_j 中每个特征的外激素量：

$$\tau_i(t+1)=\rho.\tau_i+\triangle\tau_i(t) \tag{2}$$

其中，ρ 是介于 0~1 之间的常数，为外激素蒸发系数；$\triangle\tau_i$ 是与 S_j 的分类精度成正比的一个值，如果 $f_i\notin S_j$，则 $\triangle\tau_i$ 值为 0。此信息更新方程作用于所有蚂蚁(S_1,S_2,…,S_N)，*t*=0 时刻，$\triangle\tau_i$ 值为 0，因此概率函数只考虑权重因子。在随后的迭代过程中，外激素量和权重因子共同影响转移概率函数。重复这样的进化迭代，那些具有强外激素和强先验信息的特征就会具有较高的转移概率函数从而吸引蚂蚁向其转移，实现目标函数的最优化。

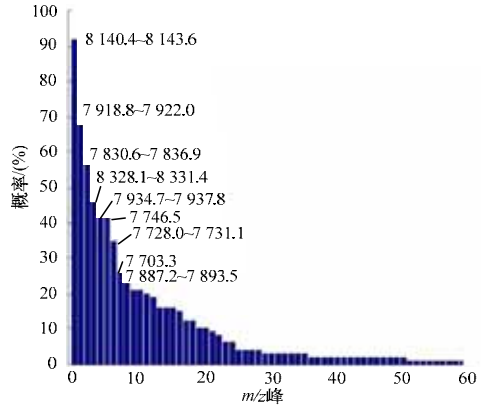
4 实验结果及分析

4.1 数据预处理

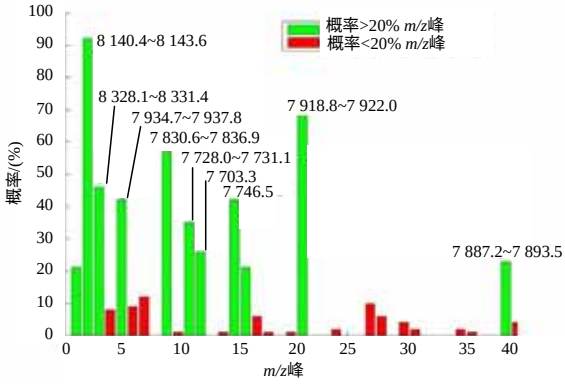
176 个肝癌患者样本和 181 个正常体样本被分成训练集与测试集，其中训练集有 200 个样本(100 个患者样本和 100 个正常体样本)，测试集包含 157 个样本数据(76 个患者样本和 81 个正常体样本)。分箱法处理原始样本数据使得质荷比维度从 340 000 降至 6 107。随后的谱峰检测和校正从 6 107 个箱子中生成了 368 个质荷比窗口。依次对 357 个质谱图谱计算找出每个窗口的最大电离强度，生成 368×357 数据矩阵用于进一步的生物标记物的选择。

4.2 生物标记物选择

谱峰检测的目的是选择出具有合理电离强度和信噪比的谱峰，而生物标记物选择所要解决的问题是识别出能够用于分类癌症患者和正常体的谱峰。本文中 ACO-SVM 算法的参数设定如下：N =50, L =368, n =5, α =1, β =1, ρ =0.1。采用 Golub[6] 提出的信噪比排序作为先验信息，即 $\eta_i = |u_{1i} - u_{2i}| / (\sigma_{1i} + \sigma_{2i})$，其中，$u_{1i}$ 和 u_{2i} 分别代表第 1 组和第 2 组中谱峰 i 的平均强度；σ_{1i} 和 σ_{2i} 分别为其对应的样本标准差。整个算法运行 350 次，最大迭代次数设置为 500，算法终止条件为 SVM 分类器收敛或已达最大迭代次数。图 2 分别按频率和信噪比排序描绘了质荷比峰被选中的概率。选用前 9 个质荷比峰建立了肝癌蛋白指纹图诊断模型，其敏感度为 94%，特异度为 92.4%。



(a)按频率排序



(b)按信噪比排序

图 2 质荷比峰被选中的概率

表 1 列出了分别使用 SNR 排序、ACO-SVM 和 PSO-SVM 方法从 368 个候选质荷比峰中选择出的前 9 个质荷比峰，PSO-SVM 一栏中的数据来自文献[2]，ACO-SVM 实验结果中有 7 个质荷比峰与 PSO-SVM 实验结果一致，其中有 4 个与 SNR 排序实验结果一致。通过比较，可以看出 ACO-SVM 算法的性能要好于 SNR 排序和 PSO-SVM 算法。

表 1 质荷比峰排序表

SNR 排序	ACO~SVM	PSO~SVM
敏感度：86%	敏感度：94%	敏感度：92%
特异度：80%	特异度：92.4%	特异度：91%
8 163.2~8 166.4	8 140.4~8 143.6	7 918.8~7 922.0
8 140.4~8 143.6	7 918.8~7 922.0	7 746.5
8 328.1~8 331.4	7 830.6~7 836.9	8 140.4~8 143.6
7 669.5	8 328.1~8 331.4	7 934.7~7 937.8
7 934.7~7 937.8	7 934.7~7 937.8	7 887.2~7 893.5
7 953.7	7 746.5	4 473.4~4 475.2
7 818.1	7 728.0~7 731.1	7 830.6~7 836.9
7 684.8~7 687.9	7 703.3	7 728.0~7 731.1
7 830.6~7 836.9	7 887.2~7 893.5	7 718.7~7 721.8

在 ACO-SVM 实验中值得注意的是，当用前 3 个质荷比峰(8 140.4~8 143.6; 7 928.8~7 922; 7 830.6~7 836.9 *m/z*)建立诊断模型时，其敏感度为 91%，特异度为 88%。这与 PSO-SVM 算法选择出的前 7 个质荷比峰(其中也包括这 3 个质荷比峰)在同等条件下建立的诊断模型性能是完全一致的，提示存在不同质荷比峰之间有相互削弱的可能，要进一步寻求其对应的生物病理学来解释。

5 结束语

本文设计的 ACO-SVM 特征选择方法具有一定的优越性，利用其筛选出的差异蛋白峰建立的肝癌诊断模型获得了较好的敏感度和特异度。但是，SELDI-TOF 蛋白质组学数据分析有许多方面值得探讨[3,7-8]，比如，质谱数据本身的高维、小样本和样本不平衡等问题，以及蛋白质的序列、构象、提纯、特性无法给出描述，还需要和其他技术，如二维电泳、酵母双杂交、生物信息学等结合使用，才能给出蛋白质的生化特征。下一步的工作是扩大样本例数、进行独立的实验方法验证和对找到的差异蛋白做进一步的分析和鉴定。

参考文献

[1] 陈主初. 疾病蛋白质组学[M]. 北京: 化学工业出版社, 2006.

[2] Ressom H. Analysis of Mass Spectral Serum Profiles for Biomarker Selection[J]. Bioinformatics, 2005, 21(6): 4039-4045.

[3] Diamandis E P. Mass Spectrometry as a Diagnostic and a Cancer Biomarker Discovery Tool: Opportunities and Potential Limitations[J]. Mol. Cell Proteomics, 2004, 6(3): 367-378.

[4] Sauve A C, Speed T P. Normalization, Baseline Correction and Alignment of High-throughput Mass Spectrometry Data[C]//Proc. of the Genomic Signal Processing and Statistics Workshop. Baltimore, USA: [s. n.], 2004.

[5] Dorigo M, Maniezzo V, Colomi A. Ant System：Optimization by a Colony of Cooperating Agent[J]. IEEE Transactions on Systems, Man and Cybernetics, 1996, 26(1): 29-41.

[6] Golub T R. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring[J]. Science, 1999, 286(8): 531-537.

[7] Meuleman W. Comparison of Normalisation Methods for Surface-enhanced Laser Desorption and Ionisation(SELDI) Time-Of-Flight(TOF) Mass Spectrometry Data[EB/OL]. (2008-02-07). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2258289/.

[8] Meng Haohua, Li Guozheng, Wang Ruisheng, et al. The Imbalanced Problem in Mass-spectrometry Data Analysis[C]//Proc. of the 2nd International Symposium on Optimization and System Biology. Lijiang, China: [s. n.], 2008: 128-135.

编辑 任吉慧