9 种结合式机器学习算法在肿瘤早期诊断中的准确性比较研究

冯 利 岳小飞

（国家开放大学医药学院 北京 100039） （北京康复医院药剂科 北京 100144）

〔摘要〕 提出应用偏最小二乘法与线性判别分析、K－最近邻法、决策树等组成 9 种结合式分类器，分析幼年转基因肿瘤小鼠及正常对照组小鼠的血清蛋白质组数据集，阐述原理与方法，比较各结合式分类器的分类准确率。

〔关键词〕 机器学习；早期诊断；肿瘤；高维数据

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.05.005

A Comparative Study on the Accuracy of Nine Combined Machine Learning Algorithms in Early Diagnosis of Tumor *FENG Li, College of Medicine, The Open University of China, Beijing 100039, China; YUE Xiaofei, Department of Pharmacy, Beijing Rehabilitation Hospital, Beijing 100144, China*

〔**Abstract**〕 The paper proposes nine combined classifiers, including Partial Least Squares (PLS), Linear Discriminant Analysis (LDA), K－Nearest Neighbor (KNN) method, Decision Tree (DT), etc., which are used to analyze the serum proteome data sets of young transgenic tumor mice and normal control mice. It expounds the principle and method, and compares the classification accuracy of the combined classifiers.

〔**Keywords**〕 machine learning; early diagnosis; tumor; high－dimensional data

1 引言

1.1 研究背景

近年来组学技术如基因组学、蛋白质组学和代谢组学迅速发展。医学工作者可将组学、患者临床诊断及影像学等数据整合以提高疾病诊断的准确性，特别是恶性肿瘤等重大疾病[1]。虽然恶性肿瘤诊断方法发展迅速[2]，但其早期诊断仍较困难。组学可从系统、整体水平捕捉机体在疾病早期的生理、病理变化，为恶性肿瘤早期诊断提供重要参考依据[3－4]。组学数据通常变量数目多、样本量少，这给数据分析带来较大挑战。多元统计分析方法及机器学习算法因具有强大的数据分析处理能力，在支持临床决策及寻找早期诊断生物标志物方面发挥了越来越重要的作用[5－8]。

1.2 研究内容

本研究首先将原始数据集分为训练数据集（约为全部数据的 1/10）和测试数据集（约为全部数据的 9/10）。先采用训练数据集建立数据处理模型，即通过偏最小二乘法（Partial Least Squares，PLS）降维，提取不同数量主成分导入到线性判别分析

〔修回日期〕 2021－10－09

〔作者简介〕 冯利，博士，讲师，发表论文 20 余篇。

(Linear Discriminant Analysis, LDA), K - 最近邻法(K - Nearest Neighbor, KNN), 决策树(Decision Tree, DT), 支持向量机(Support Vector Machine, SVM), 人工神经网络(Artificial Neural Network, ANN), 装袋法(Bagging), 随机森林(Random Forest, RF), 二次判别分析(Quadratic Discriminant Analysis, QDA)及逻辑回归(Logistic Regression, LR) 9 种分类器中对数据进行分类, 采用 10 折交叉验证法优化各分类器参数及防止模型过度拟合, 使之达到最佳分类效果, 采用预测准确率等指标对其分类效果进行评价并将表现较好的几种分类器组成集合式分类器。此外对潜在生物标志物进行初步筛选。

2 原理与方法

2.1 基本原理

PCA 和 PLS 是两种常用的降维方法[9]。二者均通过对多变量数据信息调整组合提取少量综合变量来解释原数据的大部分变异, 当组间变异在总变异中不占主导地位时, PLS 分类效果往往比 PCA 更好[10]。此外 PLS 算法在处理高维、共线性、干扰强的数据时功能强大。SVM 可处理分类及回归问题, 其泛化能力优秀, 但运算量较大。RF、Bagging 和 DT 这 3 种方法较简便, 易于解释和可视化, 但有时预测准确性不高。LR 的特点是运算速度快、模型简单、易于理解, 可直接看到各个变量的权重。LDA 和 LR 相似, 二者的区别是决策边界的估计方法不同。当决策边界高度非线性时, KNN 预测结果常优于 LDA 和 LR。QDA 使用二次决策边界, 当数据集满足高斯分布假设时, 其预测结果常比 KNN 好。评价机器学习模型分类效果的常用指标有准确率、曲线下方面积(Area Under the Curve, AUC)值、召回率、精密度、*F*1 值等。其中准确率最常用, 其缺点是当两组样本数量相差太大时该指标会失真。召回率是阳性样本的检出率。精密度是阳性样本的预测准确率。AUC 值为受试者工作特征(Receiver Operating Characteristic, ROC)曲线下方面积, 在两组样本数不平衡时该指标更为客观; *F*1 值是召回率和精密度的调和平均值, 能直观评价模型对疾病患者的检出率及检测准确性。在医学研究中, 除疾病诊断外还可通过计算 PLS 模型中每个自变量的 VIP 值来筛选与样本类别密切相关的重要变量(潜在生物标志物)。一般认为, VIP 值大于 1 以及变量峰面积(峰高或表达量等)组间 *t* 检验或方差分析(Analysis of Variance, ANOVA)有显著性差异($P < 0.05$)的变量才是较为可靠的潜在生物标志物。

2.2 数据集

本研究使用美国 FDA - NCI 蛋白质组项目数据库中的蛋白质组公共数据集, 包括 SELDI - TOF - MS 高分辨质谱技术平台采集的 80 例转基因导管胰腺癌小鼠血清样本和 101 例年龄相仿的正常小鼠血清样本蛋白质组数据, 使用质荷比(扫描范围为 800 ~ 11 992.91 Da)及对应蛋白质的峰面积作为特征变量, 共 6 771 个变量[11](http://home.ccr.cancer.gov/ncifdaproteomics/ppat - terns.asp)。

2.3 数据预处理

组学数据十分复杂, 噪音信号多, 有时还有缺失值, 因此其预处理非常重要。由于该数据集已进行谱峰的质荷比(m/z)校准, 本研究首先对数据进行归一化、中心化和标度化等预处理, 调整样本间基线偏差, 消除仪器不稳定, 以及各峰间由于峰面积数值存在较大差异对分析结果的影响。在本文中数据预处理以及后续所有数据统计处理均在 R 语言(版本: 3.6.1)数据处理平台完成[12]。

2.4 分类器与降维技术相结合的分类模型

参考相关文献[10]及本研究数据初步分析结果, 选取 PLS 作为降维方法。提取 PLS 不同数量的主成分与 LDA 等 9 种分类器组成结合式分类器。在本研究中, 机器学习算法均采用 R 语言软件包完成, SVM 使用的是“e1071”软件包(版本: 1.7 - 0.1); PLS 使用的是“mixOmics”软件包(版本: 6.3.2); BAGGING 和 RF 使用的是“randomForest”软件包(版本: 4.6 - 14); ANN 使用的是“nnet”软件包(版本: 7.3 -

12)；DT 使用的是“tree”软件包（版本：1.0－39)；LDA 和 QDA 使用的是“MASS”软件包（版本：7.3－5)。LR 用 R 语言“glm”函数完成。

2.5 模型预测能力评价

对模型预测效果用准确率（Accuracy)、精密度（Precision)、召回率（Recall)、AUC、*F*1 值进行评价。

3 结果与分析

3.1 各结合式分类器的分类准确率

首先采用 PLS 和 PCA 方法选取 20 个主成分对数据集进行降维以初步观察数据，得出各主成分的累计方差贡献率，见图 1。PCA 第 1 主成分即可解释原始变量约 95% 的方差，这表明各自变量间相关性较大；PLS 第 1 主成分可解释自变量和因变量大约 50% 的方差。通过 10 折交叉验证得出，选择 25 个主成分时 PLS 的判别分析（PLS－DA）正确率为 67%，这与原始数据集的变量数目太大及与分类不相关的干扰因素较多有关。参考 PLS 对方差的解释能力，见图 1，选取 PLS 的前 5、15 及 25 个主成分构建结合式分类器，PLS－LDA、PLS－LR、PLS－QDA、PLS－ANN、PLS－SVM 的分类效果较好。使用 25 个主成分时，其分类正确率分别为 100%、100%、99%、96% 和 96%，随着主成分数目的增加其分类准确性也增加。PLS－RF、PLS－BAGGING、PLS－DT 和 PLS－KNN 的分类效果不理想，当主成分数目增大时，其分类准确率不仅没有提高，反而下降，见图 2。将 PLS－LR、PLS－LDA、PLS－ANN、PLS－SVM、PLS－QDA 几种分类器以多数投票表决法构建集合式分类器（PLS－RES)，考察其分类准确性和主成分数目的关系，PLS－RES 在使用 15 个主成分时分类准确度即可达到 100%，见图 3。

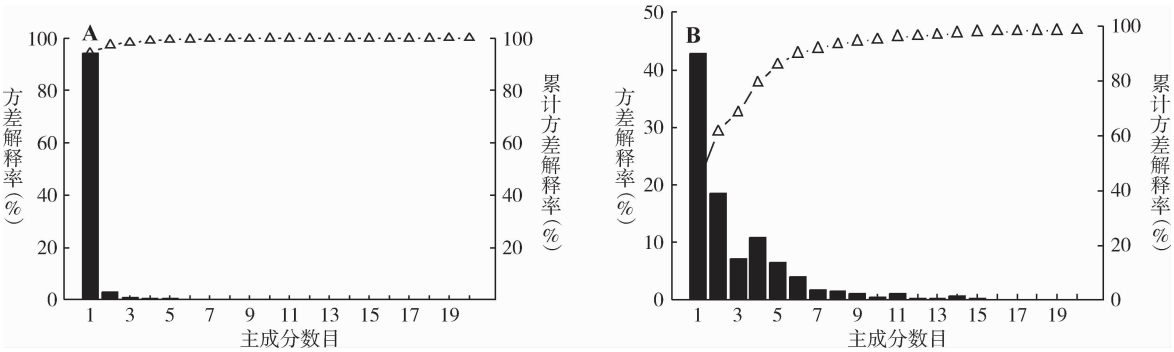

图 1 PCA（A）及 PLS（B）提取的前 20 个主成分的方差贡献率和累积方差贡献率

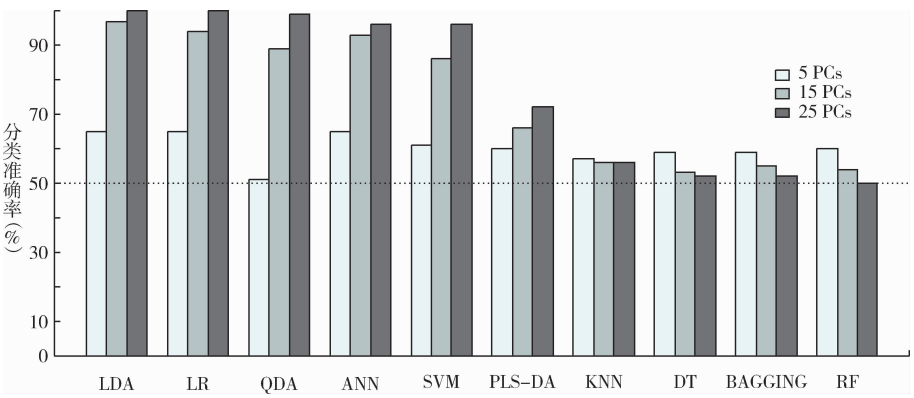

图 2 10 种分类器在主成分为 5（10PCs)、15（15PCs）和 25（20PCs）时的分类准确率

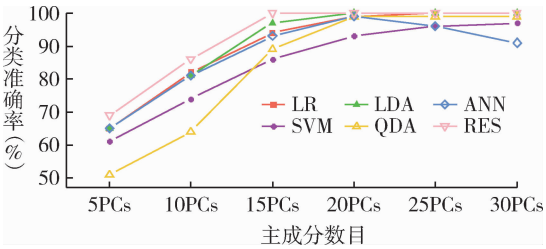

图 3 采用不同主成分数时 5 种结合模型的分类正确率

3.2 模型预测指标的评价

当主成分数为 5、10、20 时 PLS－ANN 等 5 种分类器的 5 种评价指标预测值，见表 1。当主成分数目为 2 和 10 时 5 种分类器各评价指标预测值，见图 4。当预测正确率接近 100% 时，5 种评价指标的差别不大。当预测正确率逐渐降低时 *F*1 值和召回率显著下降。选择 5 个主成分时 QDA 的预测正确率为 51%，其 *F*1 值和召回率分别仅为 19% 和 12%。

表 1 主成分数为 5、10 和 20 时各分类器 5 种评价指标的预测值（%）

主成分数	评价指标	LR	SVM	LDA	QDA	ANN
5	准确率	67	61	65	51	65
	AUC	67	61	65	50	62
	精准度	77	55	61	12	77
	召回率	62	57	60	47	57
	*F*1 值	67	54	60	19	66
10	准确率	83	74	81	64	81
	AUC	83	73	86	63	94
	精准度	88	71	80	37	84
	召回率	88	70	82	75	77
	*F*1 值	87	69	79	50	80
20	准确率	99	93	100	99	99
	AUC	99	97	100	98	100
	精准度	99	91	100	98	96
	召回率	99	94	100	99	100
	*F*1 值	99	92	100	98	98

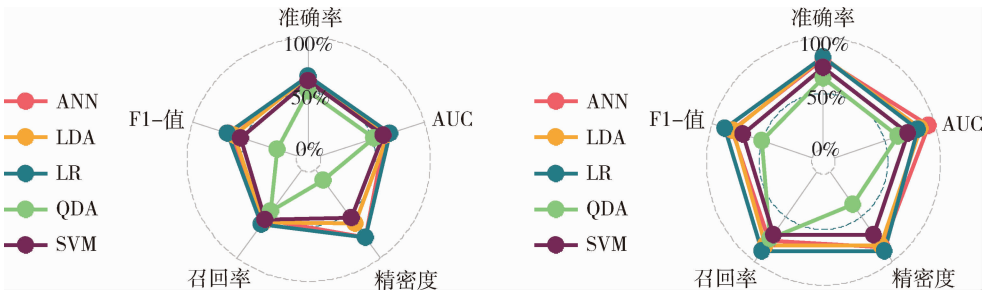

图 4 主成分数目为 2（A）时和 10（B）时 5 种结合分类器各评价指标的预测值

3.3 潜在生物标志物筛选

当主成分数为 20 时采用 PLS－DA 结合模型，筛选得到前 20 个主成分的 VIP 均值 > 1 且 *t* 检验 *P* < 0.05 的变量（潜在生物标志物）105 个，见表 2。本研究主要目的是构建一种处理多维数据的结合式算法以对不同生理功能的生物样本进行分类，因此筛选出的潜在生物标志物为何种蛋白质及其具有何种生物学功能需要进一步鉴定和分析。

表 2 通过 PLS－DA 结合模型筛选出的潜在生物标志物信息（部分）

序号	m/z 值	VIP 值	*P* 值	序号	m/z 值	VIP 值	*P* 值
1	11 755.479 5	4.59	0.02	11	11 699.199 0	3.42	0
2	11 741.385 0	4.39	0	12	11 694.521 1	3.38	0
3	11 736.690 2	4.00	0	13	11 689.846 2	3.37	0
4	11 731.998 2	3.75	0	14	11 685.172	3.28	0
5	11 727.307 0	3.68	0	15	11 680.501 0	3.28	0
6	11 722.617 1	3.63	0	16	11 675.830 3	3.22	0
7	11 717.930 3	3.61	0	17	11 671.161 1	3.22	0
8	11 713.244 1	3.60	0	18	11 666.494 2	3.12	0
9	11 708.561 2	3.46	0	19	11 661.829 0	3.11	0
10	11 703.879 0	3.45	0	20	11 657.166 1	3.09	0

4 结语

研究[11]发现，$KRAS^{G12D}$基因表达与成年（9 周龄）小鼠侵入性胰腺导管癌密切相关，携带该致癌基因的小鼠成年后全部患癌。本研究中的数据集为携带 $KRAS^{G12D}$基因的幼年（5 周龄）转基因小鼠及年龄相仿的正常对照组小鼠血清蛋白质组学数据。采用本研究建立的结合式分类器在癌症未发病时即可将携癌基因幼年小鼠与正常对照组加以区分，表明本研究具有较大潜在应用价值。在对本研究中数据集进行 PCA 分析时发现各变量之间具有较高相关性，当变量之间高度相关时 PLS 的分类准确性明显优于 PCA[9]。此外有研究[13]发现，当变量之间相关性较高时，基于特征提取的 SVM 比单独使用 SVM 的分类效果好，这与本研究结果一致。本研究建立的方法也可用于基于光谱[9]、色谱、基因组、代谢组、影像等高维数据及包括少数几种临床诊断指标的低维数据的肿瘤辅助诊断。低维数据可不降维直接进行分类。此外本研究提出的潜在生物标志物的筛选方法有助于通过测定少数指标即可对肿瘤进行早期诊断。

参考文献

1 张桐硕，任鹤菲，曹瑾，等. 基于集成机器学习的卵巢癌多检验指标联合诊断模型［J］. 临床检验杂志，2018，36（12）：908－913.

2 刘洪璐，王熙才. 外周血 miRNA 应用于肿瘤早期诊断的研究进展［J］. 中国肿瘤生物治疗杂志，2018（2）：109－117.

3 孙权，高福，蔡建明. 组学技术在肿瘤诊断中的应用［J］. 中国肿瘤临床，2010，37（17）：1012－1015.

4 韩企夏. 早期发现是治愈乳腺癌的关键［J］. 抗癌，2001（3）：29.

5 张婷婷，渠宁，郑璞，等. 机器学习在甲状腺肿瘤诊疗中的应用［J］. 中国癌症杂志，2017，27（12）：992－995.

6 李喆，吕卫，闵行，等. 机器学习在乳腺肿瘤分类检测中的应用研究［J］. 计算机工程与科学，2016，38（11）：2303－2309.

7 李建更，李萍，李君，等. PCA 和 PLS 应用于胃癌亚型分类研究［J］. 生物物理学报，2009，25（2）：141－147.

8 孙小宇，姚晨，康晓平. 支持向量机在建立冠心病早期诊断模型中的应用［J］. 中国卫生统计，2011，28（2）：122－125.

9 刘平年. PLS 法和 PCA 法在近红外光谱定量分析中的应用研究［J］. 广州食品工业科技，2004，20（4）：106－107，34.

10 刘文慧. PCA 与 PLS 用于高维数据分类的比较性研究［C］. 西安：2011 年中国卫生统计学年会会议，2011.

11 Hingorani S R，Petricoin E F，Anirban M，et al. Preinvasive and Invasive Ductal Pancreatic Cancer and Its Early Detection in the Mouse［J］. Cancer Cell，2003，4（6）：437－450.

12 R Core Team（2018）. R：A Language and Environment for Statistical Computing. R Foundation for Statistical Computing，Vienna，Austria［EB/OL］.［2018－12－20］. https：//www. R－project. org/.

13 于春梅，杨胜波，陈馨，等. SVM 和基于 PCA、PLS 的 SVM 在非线性辨识中的比较研究［J］. 计算机应用研究，2007，24（6）：85－86.

（上接第 24 页）

2 朱秘平，邓朝华. 互联网医院在患者就医中的优势：系统综述［J］. 中国卫生事业管理，2020，40（11）：90－92.

3 廖生武，薛允莲，谭碧慧，等. “互联网＋”人工智能时代医院智慧诊疗管理策略［J］. 中国医院管理，2019，39（10）：5－8.

4 王雨，李巍，冯磊. 供需匹配框架下的互联网医院生成逻辑、运行困境及路径优化［J］. 中国卫生经济，2019，38（8）：24－26.

5 国家统计局. 中国统计年鉴－2018［M］. 北京：中国统计出版社，2019.

6 单莹，孔凡磊，时涛，等. 我国公共卫生财政投入现状的时空分析［J］. 中国卫生经济，2020，39（9）：41－44.

7 束雅春，宁丽琴，陈列红，等. 公立中医院建设互联网医院实践与思考［J］. 中国医院，2021，25（4）：28－30.

8 健康界智库. 2019 互联网医院发展研究报告［EB/OL］.［2019－03－02］. https：//www. doc88. com/p－1408459166340. html？r＝1.

9 王丽敏，陈志华，张梅，等. 中国老年人群慢性病患病状况和疾病负担研究［J］. 中华流行病学杂志，2019（3）：277－283.

10 伍曦，杨凤，李佳芮，等. 基于 4 C 理论模型的互联网医院发展策略探讨［J］. 医学信息学杂志，2021，42（4）：13－16.

11 杜学鹏，吴晓丹，贾宏明. 互联网医院发展的问题识别与对策［J］. 卫生经济研究，2021，38（1）：22－25.