PCA-MPL-ANN 模型在卵巢肿瘤良恶性鉴别中的价值

秦明丽,王定玉,王 旗,李武志,王 琴
(四川省妇科及乳腺疾病治疗中心/西南医科大学附属医院妇科,四川 泸州 646000)

摘要:目的 构建一种集成降维技术和人工神经网络分类器的机器学习模型,探讨其对卵巢肿瘤良恶性的鉴别诊断价值。**方法** 收集 2013 年 1 月 28 日~2014 年 12 月 30 日西南医科大学附属医院诊断的卵巢癌患者(132 例)及卵巢良性肿瘤患者(211 例),通过电子病历获取人口学资料及 8 项血清肿瘤标志物检测指标。以主成分分析法(PCA)提取综合数据信息并采用多层感知器人工神经网络(MPL-ANN)模型进行诊断和预测分析。随机选取 2/3 患者为训练集建立诊断模型,1/3 为测试集进行预测,计算该模型的诊断和预测正确率及受试者工作特征曲线下面积(AUC)。**结果** PCA-MPL-ANN 模型对卵巢癌及卵巢良性肿瘤的诊断正确率分别为 66.33%和 92.00%,预测正确率分别为 67.74%及 83.61%;该模型 AUC 达到 0.838,优于 β-HCG(0.748)、CA153(0.680)及 CA125(0.613)单项指标的 AUC。**结论** 利用 PCA-MPL-ANN 整合多项血清肿瘤标志物可有效提升卵巢癌的诊断效能,为卵巢癌的智能化辅助诊断提供参考。

关键词:卵巢癌;PCA 分析;MPL-ANN 模型;肿瘤标志物
中图分类号:R737.31 **文献标识码:**A DOI:10.3969/j.issn.1006-1959.2021.07.018
文章编号:1006-1959(2021)07-0063-04

The Value of the PCA-MPL-ANN Model in the Differential Diagnosis of Benign and Malignant Ovarian Tumors

QIN Ming-li, WANG Ding-yu, WANG Qi, LI Wu-zhi, WANG Qin
(Department of Gynecology, Sichuan Provincial Gynecology and Breast Disease Treatment Center/ Affiliated Hospital of Southwest Medical University, Luzhou 646000, Sichuan, China)

Abstract: Objective To construct a machine learning model integrating dimensionality reduction technology and artificial neural network classifier, and to explore its value in the differential diagnosis of benign and malignant ovarian tumors. **Methods** Collected ovarian cancer patients (132 cases) and ovarian benign tumor patients (211 cases) diagnosed in the Affiliated Hospital of Southwest Medical University from January 28, 2013 to December 30, 2014. Obtaining demographic data and 8 serum tumor marker detection indicators through electronic medical records. Principal component analysis (PCA) is used to extract comprehensive data information and a multi-layer perceptron artificial neural network (MPL-ANN) model is used for diagnosis and predictive analysis. Randomly select 2/3 patients as the training set to establish a diagnostic model, and 1/3 as the test set for prediction, and calculate the diagnostic and prediction accuracy of the model and the area under the receiver operating characteristic curve (AUC). **Results** The correct diagnosis rate of PCA-MPL-ANN model for ovarian cancer and benign ovarian tumor were 66.33% and 92.00%, and the correct prediction rates were 67.74% and 83.61%, respectively; The AUC of this model reaches 0.838, which is better than the AUC of β-HCG (0.748), CA153 (0.680) and CA125 (0.613). **Conclusion** Using PCA-MPL-ANN to integrate multiple serum tumor markers can effectively improve the diagnostic efficiency of ovarian cancer, and provide a reference for the intelligent auxiliary diagnosis of ovarian cancer.

Key words: Ovarian cancer; PCA analysis; MPL-ANN model; Tumor markers

卵巢癌(ovarian cancer)是女性生殖系统三大恶性肿瘤之一,其死亡率高居妇科恶性肿瘤之首[1]。据统计,2017 年全球新发卵巢癌病例 22.4 万,其中 14.1 万患者因该病死亡[2],而我国年新发卵巢癌 5.2 万,死亡病例高达 2.3 万[3]。由于卵巢位于盆腔深处,部位隐蔽且患者临床症状不典型,故确诊时多为中晚期。研究表明,Ⅱ~Ⅳ期卵巢癌患者 5 年生存率仅为 15%~45%,而Ⅰ期患者 5 年生存率高达 90%[4],故早期、及时诊断对提高患者生存率至关重要。血清肿瘤标志物具有微创、多次采集、快速检测等诸多优点,其中癌胚抗原(CEA)、糖类抗原-125(CA125)和 β 人绒毛膜促性腺素(β-HCG)已广泛用于卵巢的鉴别诊断、疗效判断和预后评估。本研究在参考血清 CA125、CA153 和 HCG 等用于鉴别诊断卵巢癌及卵巢良性疾病患者的基础上[5-7],从卵巢肿瘤既往血清 CEA、甲胎蛋白(AFP)、CA125、CA153、CA199、CA724 及 β-HCG 等多项肿瘤标志物出发,集主成分分析(PCA)、多层感知器(MPL)及人工神经网络(ANN)等机器学习模型,试为卵巢肿瘤良恶性鉴别诊断提供便利的决策支持,现报道如下。

1 资料与方法

1.1 一般资料 回顾性分析西南医科大学附属医院 2013 年 1 月 28 日~2014 年 12 月 30 日门诊及住院的 132 例卵巢癌患者及 211 例卵巢良性肿瘤患者纸质及电子病历信息,并对患者血清 CEA、CA125、CA153、CA199、CA724 及 β-HCG 等 8 项肿瘤标志物测定结果进行分析。卵巢癌患者经病理或影像确诊,年龄 21~77 岁,其中Ⅰ~Ⅱ期 14 例,Ⅲ~Ⅳ期 60 例,未分期 58 例;浆液性卵巢癌 47 例,粘液性卵巢癌 32 例,未分型 53 例。排除其他恶性肿瘤疾病(或转移性肿瘤)、肝、肾功能显著异常、心肺功能不全、盆腔炎,高血压及糖尿病等患者。参照《体外诊断试剂临床研

基金项目:1.四川省科技厅资助项目(编号:20018/20YZTG0050);2.西南医科大学附属医院资助项目(编号:17194)
作者简介:秦明丽(1979.7-),女,四川泸州人,硕士,主治医师,主要研究领域为卵巢癌的诊断工作
通讯作者:王琴(1984.5-),女,四川仁寿人,硕士,主治医师,主要研究领域为卵巢癌的诊断与治疗工作

究指导原则》中客观上不可能获得受试者知情同意或该临床研究对受试者几乎没有风险,可以不提交伦理委员会的审核意见及受试者的知情同意书进行。

1.2 仪器与检测方法 所有患者在诊断及放化疗前空腹采血 3~5 ml 后,尽快离心分离血清检测。采用日本东曹 AIA2000 化学发光仪及配套试剂在质控在控下按标化操作规程操作。各项血清参考区间分别为 AFP:0~10.0 ng/ml,CEA:0~6.0 ng/ml,CA125:0~35.0 IU/ml,CA153:0.31~23.0 IU/ml,CA724:0.21~6.0 IU/ml,SCCA:0.011~2.5 IU/ml,CA199:0~37.0 IU/ml,β-HCG:0~3.0 mIU/ml。

1.3 PCA 模型的建立 以卵巢癌及卵巢良性肿瘤患者间有统计学差异的肿瘤标志物建立 PCA 模型,将上述血清肿瘤指标进行降维处理,利用 Z 分标准化数据后通过正交变换把相关的高维指标综合成少数几个不具相关性的新变量,提取主成分(P),在保留原来指标的大部分信息后又简化了数据结构,便于在低维度下建立疾病诊断模型。PCA 模型采用协方差矩阵进行 PCA 分析,基于特征值大小提取 PCA 并建立基于各指标的前三个 PCA(P_1~P_3)的线性方程。

1.4 PCA-MPL-ANN 模型的建立 以提取的前三个主成分(P_1~P_3)建立基于 PCA-MPL-ANN 模型,以概率 P>0.50 诊断为卵巢癌,反之为卵巢良性疾病。计算该模型的隐含层、训练时间,并随机选取 70%个体为训练集,30%个体为测试集进行预测,计算诊断及预测正确率及 PCA-MPL-ANN 模型的 ROC 曲线下面积(AUC)。

1.5 统计学方法 采用 SPSS 17.0 软件进行统计学分析,血清 8 项肿瘤标志物水平均呈偏态分布,以中位数和四分位数间距表示[$M(P_{25},P_{75})$],两组比较采用独立样本的秩和检验,诊断价值采用受试者工作特征(ROC)曲线分析,P<0.05 表示差异有统计学意义。

2 结果

2.1 卵巢癌与卵巢良性肿瘤患者血清 8 项指标比较 卵巢癌患者血清 AFP、CEA、CA125、CA153、CA724 及 β-HCG 水平均高于卵巢良性肿瘤患者,差异有统计学意义(P<0.05),见表 1。

表 1 卵巢癌与卵巢良性肿瘤患者血清 8 项指标比较[$M(P_{25},P_{75})$]

项目	卵巢癌(n=132)	卵巢良性肿瘤(n=211)	Z	P
AFP	2.70(2.12,3.62)	2.26(1.68,3.69)	2.341	0.019
CEA	2.27(1.53,3.01)	1.77(1.24,2.56)	2.812	0.005
CA125	29.70(11.40,159.20)	16.70(9.25,40.90)	3.517	0.000
CA153	16.90(11.00,28.60)	11.80(8.99,18.10)	5.615	0.000
CA199	14.40(10.00,24.10)	16.10(10.30,27.60)	1.196	0.232
CA724	4.22(2.39,9.72)	3.49(1.95,5.21)	2.706	0.007
β-HCG	1.00(0.54,2.15)	0.21(0.10,1.00)	7.753	0.000
SCCA	0.36(0.22,0.66)	0.35(0.18,0.67)	0.833	0.404

2.2 卵巢癌与卵巢良性肿瘤患者血清 8 项指标 ROC 曲线分析 两类患者的 AFP、CEA、CA125、CA153、CA724 及 β-HCG 的 AUC 比较,差异有统计学意义(P<0.05),其中 β-HCG 最高(AUC=0.748),其次为 CA153(AUC=0.680)及 CA125(AUC=0.613),见表 2。

表 2 血清肿瘤标志物诊断卵巢癌与卵巢良性肿瘤的效能

项目	AUC	95%置信区间	P
AFP	0.575	0.515~0.636	0.019
CEA	0.590	0.528~0.652	0.005
CA125	0.613	0.547~0.678	0.000
CA153	0.680	0.622~0.739	0.000
CA199	0.462	0.523~0.650	0.232
CA724	0.587	0.523~0.650	0.007
SCCA	0.473	0.409~0.537	0.405
β-HCG	0.748	0.695~0.800	0.000

2.3 PCA 分析 共提取 3 个主成分(P_1~P_3),表达式如下:P_1=-0.103AFP-0.199CEA+0.723CA125+0.40CA153 +0.423CA724 +0.659β -HCG;P_2 = 0.501AFP -0.092CEA -0.341CA125 -0.580CA153 + 0.584CA724+0.402β-HCG;P_3=0.452AFP+0.808CEA+0.007CA125+0.271CA153-0.15CA724+0.238β-HCG。其中 P_1 主要反映 CA125、CA153 和 β-HCG 特征,可归纳为妇科肿瘤标志物及激素水平,P_2 主要反映 CA724、AFP 特征,归纳为肿瘤的胃及肝脏转移,P_3 主要反映 CEA 特征,即存在恶性肿瘤,见表 3。

表 3 前三个主成分对应的特征向量

项目	主成分 1(P_1)	主成分 2(P_2)	主成分 3(P_3)
AFP	-0.103	0.501	0.452
CEA	-0.199	-0.092	0.808
CA125	0.723	-0.341	0.007
CA153	0.40	-0.58	0.271
CA724	0.423	0.584	-0.15
β-HCG	0.659	0.402	0.238

2.4 PCA-MPL-ANN 模型分析 该模型的隐含层数为 1,训练时间为 0:00:00:120,训练集为 251 人,测试集为 92 人。该模型对卵巢癌及卵巢良性疾病的诊

断正确率分别为 66.33%(67/101) 和 92.00%(138/150),预测正确率分别为 67.74%(21/31)及 83.61%(51/61),见图 1。以建立的 PCA-MPL-ANN 模型绘制 ROC 曲线,该模型的诊断效能较高(0.838),AUC 优于 β-HCG、CA153 及 CA125 (0.748 >0.680 > 0.613),敏感性和特异性分别为 72.60%和 88.90%,见图 2。

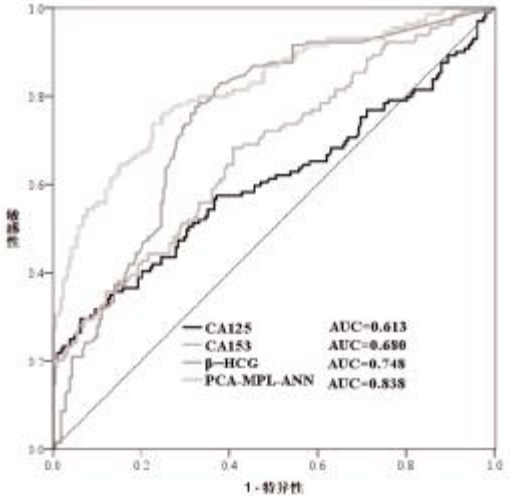

图 1 PCA-MPL-ANN 模型及 CA125、CA153 及 β-HCG 的 ROC 曲线

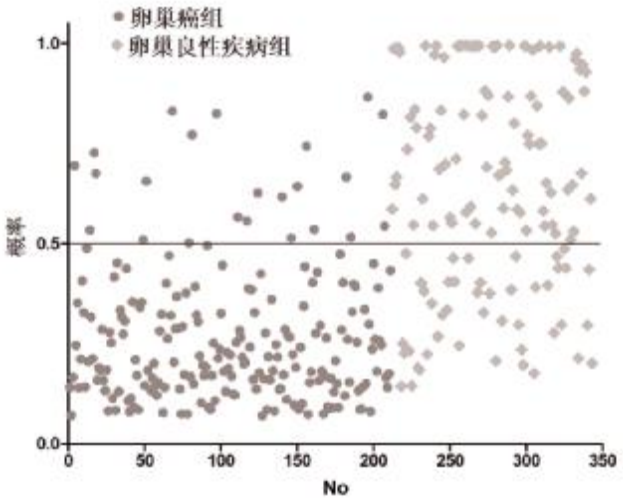

图 2 基于 PCA-MPL-ANN 模型的卵巢癌及卵巢良性肿瘤的概率分布

3 讨论

卵巢癌是最致命的妇科肿瘤,腹痛、腹部增大、腹胀或恶心等症状通常是非特异性的,直到疾病发展晚期才出现,导致诊断延迟[8]。阴道超声和血清 CA125 筛查是常用的卵巢癌筛查技术,但敏感性和特异性均不太理想[9]。临床上,联合多指标进行分析是卵巢癌常用的辅助手段,但传统的并联试验在提高诊断敏感性的同时,降低了特异性,而串联实验则在提升特异性的同时降低了敏感性,两者均不能较好地对未知患者进行预测分析。因此,寻找更多的肿瘤标志物联合检测新模式,成为提升诊断效能和进一步诊断和预测分析的突破口[10]。

近年来,人工智能和机器学习算法的蓬勃发展为分析复杂的生物数据集提供新的方法[11]。人工神经网络(ANN)作为机器学习领域最经典及最活跃的方法,通过模仿人脑神经元的拓扑结构建立的计算机学习网络系统,可以解决复杂的非线性映射问题而广受关注[12]。张桐硕等[9]研究发现,采用误差负反馈(BP)-ANN 模型综合肿瘤标志物、血细胞分析、性激素等 6 类共计 28 项实验室检测指标能很好地鉴别诊断卵巢癌、其他恶性妇科肿瘤、卵巢良性疾病及正常对照人群, 其 AUC、敏感性和特异性分别为 0.948,91.9%和 86.9%。本研究采用 PCA 提取 6 项卵巢癌及卵巢良性肿瘤患者间存在差异的肿瘤标志物,建立 PCA 及 PCA-MPL-ANN 模型,通过 PCA 降维处理,将 6 维空间的数据形象、直观地展现在三维空间,从妇科肿瘤标志物、激素水平、胃肠及肝脏转移等多方面揭示了数据规律。借助 PCA-MPL-ANN 模型能较好地鉴别诊断卵巢癌及卵巢良性疾病,该模型诊断的 AUC、敏感性和特异性分别为 0.838,72.60%和 88.90%。由于本研究仅纳入 6 项有统计学差异的肿瘤标志物,故诊断效能较张桐硕等[9]相关报道低。因此,本研究也提示仅采用现有肿瘤标志物建立机器学习模型诊断效能有限, 需要探索和挖掘更多有价值的标志性建立诊断和预测模型,提高诊断效能。

大数据时代,由于海量的数据信息不断产生,基因组学、蛋白组学及代谢组学等大样本数据信息为机器学习模型精准鉴别诊断卵巢癌提供了极大的机遇, 但需要昂贵的仪器设备及有经验技术人员且尚未形成像常规检验项目的流水线检测。因此,借助简便、易得的血清肿瘤标志物、血常规及生化指标建立机器学习诊断和预测模型值得探索研究。由于本研究中 CA125 在卵巢癌患者中表达不明显,单项指标 β-HCG、CA153 及 CA125 的诊断效能在 0.613~0.748, 故该模型也较好地弥补了 CA125 对卵巢癌早期诊断能力的不足。尽管如此,本研究纳入人群及血清肿瘤标志物项目类型有限, 尚需要大样本验证分析,使结论更严谨、可靠。

综上,PCA-MPL-ANN 模型可有效提升卵巢癌的诊断效能,取得了较好的效果,为卵巢癌的智能化辅助诊断提供了新思路。

参考文献:

[1]Gao L,Chen M,Ouyang Y,et al.Icaritin induces ovarian cancer cell apoptosis through activation of p53 and inhibition of Akt/mTOR pathway[J].Life Sciences,2018(202):188-194.

[2]Siegel RL,Miller KD,Jemal A.Cancer Statistics,2017 [J].CA,2015,60(1):277-300.

[3]Chen W,Zheng R,Baade PD,et al.Cancer statistics in China,2015[J].CA,2016,66(2):115-132.
[4]Meng W,Ying W,Qichao Z.Clinical value of combining transvaginal contrast-enhanced ultrasonography with serum human epididymisprotein-4 and the resistance index for early-stage epithelial ovarian cancer [J].Saudi Medical Journal,2017,38(6):592.
[5]Sal V,Kahramanoglu I,Bese T,et al.Is serum level of nestin useful in detecting epithelial ovarian cancer[J].J Obstet Gynaecol Res,2017,43(2):371-377.
[6]Kucera C,Cox-Bauer C,Miller C.Apparent ectopic pregnancy with unexpected finding of a germ cell tumor:A case report[J].Gynecologic Oncology Reports,2017,21(C):31-33.
[7]Lin ZY,Fang YZ,Jin H W,et al.Performance evaluation of a chemiluminescence microparticle immunoassay for CK-MB[J].Journal of Clinical Laboratory Analysis,2018,32(6):e22426.
[8]Shintani D,Yoshida H,Imai Y,et al.Acute pancreatitis induced by paclitaxel and carboplatin therapy in an ovarian cancer patient [J].European Journal of Gynaecological Oncology,2016,37(2):286.
[9]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.
[10]Kawakami E,Tabata J,Yanaihara N,et al.Application of Artificial Intelligence for Preoperative Diagnostic and Prognostic Prediction in Epithelial Ovarian Cancer Based on Blood Biomarkers[J].Clinical Cancer Research,2019,25(10):3006-3015.
[11]Bakkar N,Kovalik T,Lorenzini I,et al.Artificial intelligence in neurodegenerative disease research: use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis[J].Acta Neuropathologica,2018,135(2):227-247.
[12]Alexakis DD,Mexis FK,Vozinaki AK,et al.Soil Moisture Content Estimation Based on Sentinel-1 and Auxiliary Earth Observation Products.A Hydrological Approach [J].Sensors,2017,17(6):1455.

收稿日期:2020-12-10;修回日期:2020-12-22
编辑/成森