基于 Logistic 回归、ANN、SVM 的乳腺癌复发影响因素研究①

饶飘雪, 叶 枫

(浙江工业大学 经贸学院, 杭州 310000)

摘 要: 为找出乳腺癌复发的影响因素, 并比较人工神经网络(ANN)型、支持向量机型(SVM)和 logistic 回归型在乳腺癌复发中的预测效能. 本文结合南斯拉夫卢布尔雅那大学医疗中心乳腺癌肿瘤研究所的 277 例数据, 对乳腺癌复发的影响因素进行研究. 分别采用了 logistic 回归、人工神经网络和支持向量机方法来建立乳腺癌复发的预测模型, 并对这三种分析方法进行了理论方法和预测效能的比较. 结果发现, 肿瘤大小、有无结节冒、肿瘤恶性程度(P<0.05)是乳腺癌术后复发的主要影响因素, 而在不同的预测方法中相对于 logistic 回归模型, 支持向量机和人工神经网络具有更好的预测效能, 其中支持向量机的预测效能最好.

关键词: 乳腺癌复发; 人工神经网络; logistic 回归; 支持向量机

Research on Risks Factors of Female Breast Cancer Recurrence Based on Logistic Regression, Artificial Neural Network and Support Vector Machine

RAO Piao-Xue,YE Feng

(Management Science and Engineering, Zhejiang University of Technology, Hangzhou 310000, China)

Abstract: In order to find out the influencing factors of breast cancer recurrence, this paper investigates the artificial neural network(ANN), support vector machine(SVM) and logistic regression for the prediction of breast cancer recurrence. A data set containing 277patients' records which is provided by the University of Wisconsin Hospitals, Madison from Wolberg is used to study the influencing factors of recurrence of breast cancer. By using logistic regression, artificial neural networks and support vector machine, it determines the important factors of breast cancer recurrence, and then compares these three methods. The results show that tumors size, nodules risk, the degree of malignancy(P<0.05) are the main factors of breast cancer recurrence. Compared to the logistic regression model, support vector machine and artificial neural network has better prediction performance, and support vector machine performs best.

Key words: breast cancer recurrence; artificial neural network; logistic regression; support vector machine

1 引言

乳腺癌是女性最常见的恶性肿瘤[1], 据美国癌症协会估计, 2012 年美国确诊的乳腺癌新发病例 226870 例, 死于乳腺癌的患者 39510 例[2]. 乳腺癌的复发是指病理证实的乳腺癌经过治疗(包括手术放疗化疗)后在原发灶附近或远隔器官出现病理性质完全相同肿瘤的现象[3]. 关于乳腺癌术后局部复发率文献报道不一, 一般认为在 10%~30%之间, 而胸壁复发占所有局部复发的 50%以上[4]. 一旦乳腺癌患者出现复发或癌转移, 将为临床治疗带来更大的难度. 乳腺癌局部复发后的 5 年生存率仅为 42%~49%[5], 5 年局部控制率为 27%~75%[6]. Andre F, Slimane K 等[7]人发现近几年凭借先进的诊断仪器和规范化系统治疗方案, 复发转移早诊率提高, 患者的生存时间有所延长, 乳腺癌复发死亡风险以每年 1%-2%的速度在下降. 于是找出乳腺癌术后复发的高危因素, 通过构建的预测模型对患者复发风险进行准确预测, 将给临床治疗带来极大的效能, 进而削弱乳腺癌复发的风险.

① 收稿时间:2015-10-30;收到修改稿时间:2015-11-30 [doi: 10.15888/j.cnki.csa.005181]

X 射线是乳腺癌诊断的一种传统方法, 但是这种只有 25%的 X 光检查结果是被认可的. 穿刺细胞学是另一种乳腺癌诊断方法, 这种诊断预测准确率相对较高, 但平均准确率也只有 90%[9]. 随着时代发展, 出现了一种新的诊断预测方法, 即人工智能, 目前已在乳腺癌诊断中得到了广泛应用[10]. 自 20 世纪 60 年代初, 国内外学者已经将人工智能预测模型应用到各个领域, 包括化学、金融和医学领域等[11]. 数据挖掘是近些年来广泛应用于医学领域的一种新的分析技术方法, 在疾病诊断、预后、医疗费用管理等方面表现出良好的应用价值. 由于医学数据的特殊性, 通常情况下, 临床医学数据具有复杂性、冗余重复性、多样性、时间先后性及不规范性等特点, 数据挖掘可以帮助我们从医疗复杂的数据中提取有价值的信息, 并为临床决策提供帮助.

应用在医学研究中的数据挖掘技术主要是 Logistic 回归、人工神经网络、支持向量机和自组织映射等, 关于这几种挖掘方法的比较也是呈出不穷. 岳勇, 田考聪, 汪洋等[12]调查了结核病的可能影响因素, 包括患者的一般情况、就医行为和结核病认知等维度建立了人工神经网络预测模型, 并将分析结果与 logistic 回归模型进行了比较, 结果发现 ANN 在流行病病因的探索研究中能够发挥比 logistic 回归模型更好的作用. 吴疆, 董婷[13]运用支持向量机分类算法建立了卵巢癌病变与非卵巢癌病变质谱数据的分类模型, 结果卵巢癌预测正确率达到 98%. 并且与神经网络等算法的预测结果进行了比较, 发现在癌症数据建模的应用中, 支持向量机算法具有更强的预测能力. 为充分验证这几种方法的有效性, 本文以乳腺癌复发为分析视角, 分别运用 Logistic 回归方法, 人工神经网络方法和支持向量机方法建立预测模型, 并进行结果的比较.

2 数据材料

本研究病例数据来源于南斯拉夫卢布尔雅那大学医疗中心乳腺癌肿瘤研究所, 样本中的共 286 例病人数据信息, 其中含缺失值数据的有 9 个, 为了判断的有效性, 采取了删除缺失值的方法, 最终只将无缺失值的 277 例数据纳入分析数据集. 数据集中乳腺癌复发患者有 196 个, 占总样本比例的 70.8%. 主要从患者年龄、患者绝经年龄、肿瘤大小、受侵淋巴结数、有无结节冒、肿瘤恶性程度、肿块位置、肿块所在象限和是否进行放疗这九个方面分析对乳腺癌复发的影响. 本研究将 277 份数据按 7:3 的比例随机分为训练集(190 人)与测试集(87 人)两部分. 研究中采用 SPSS20.0 建立二分类 Logistic 回归模型和人工神经网络模型, 计算各因素与乳腺癌复发联系比值比 OR 及其 95%可信区间(95%CI). 对分类变量如患者年龄、患者绝经年龄、肿瘤大小、受侵淋巴结数、有无结节冒、肿块位置、肿块所在象限和是否进行放疗先进行了自变量赋值, 运用 matlabR2012a 建立支持向量机模型, 得出训练样本和预测样本的预测准确率.

3 方法应用

3.1 Logistic 回归分析

Logistic 回归是通过假设检验来做统计推断、分析数据的一种研究方法. 它实际是一种判别分析, 主要适用于流行病学资料的危险因素分析、临床试验评价和疾病预后因素分析等方面. 通过计算分析各个变量的回归系数对因变量的影响大小来对变量进行筛选并建立回归模型. 本文中研究的是乳腺癌是否复发的问题, 是一个典型的二分类回归问题. 先对每个预测变量进行单因素方差分析, 初步找出对因变量(复发)有影响的因素, 再将上述预测结果中显著性检验分析 $P<0.05$ 的变量纳入多元模型的候选预测变量. 然后, 采用 logistic 前向逐步回归分析(LR 法)进行多因素分析, 建立 logistic 回归预测模型, 得到预测复发的准确率.

在单因素方差分析结果中, 发现:肿瘤恶性程度、肿瘤大小、有无结节冒、受侵淋巴结数、是否放疗(P=0.000)对乳腺癌复发有显著影响. 以是否复发作为因变量, 以可能的影响因素为自变量进行乳腺癌复发多因素 logistic 回归分析, 采用偏最大似然估计进行逐步回归分析, 最后结果中只显示了 3 个危险因素进入 logistic 回归方程, 建立了 logistic 回归模型, 依次为:肿瘤恶性程度、有无结节冒、肿瘤大小, 而在单因素方差分析时受侵淋巴结数和是否放疗都具有统计学意义, 却未被纳入到 logistic 回归模型中, 最终进行二元 logistic 回归时得到的预测模型方程为:

Logit P=-3.961+0.157*肿瘤大小+0.944*有无结节冒+0.903*肿瘤恶性程度, 将上述公式由 Logit P 形式转

化成 logistic 形式, 得到:

$$P=\frac{1}{1+e^{-(-3.961+0.157*肿瘤大小+0.944*有无结节冒+0.903*肿瘤恶性程度)}}$$

其中, P 为 logistic 模型预测概率, e 为自然对数, 影响因素的 OR 值(95%CI)分别是 1.169(1.014~1.348)、2.570(1.337~4.939)、2.467(1.589~3.828).

表 1 logistic 回归模型样本分类表

		预测例数		准确率(%)
		无复发	复发	
观察例数				
	无复发	183	13	93.4
	复发	54	27	33.3
总体准确率				75.8

通过 logistic 回归模型样本分类表(表 1)对模型的预测效果进行评价, 得到了对乳腺癌复发的预测概率值, 结果表明样本灵敏度为 93.4%(183/196), 特异度为 33.5%(27/81), 误诊率为 32.5%(13/40), 漏诊率为 22.8%(54/237), 总体预测准确率为 75.8%.

表 3 logistic 回归模型样本分类表

3.2 BP 神经网络

BP 神经网络是一种典型的有导师学习的神经网络算法, 具有一个输入层和输出层、若干个隐含层, 是一种典型的多层前向型神经网络, 层与层之间采用全连接的方式, 隐含层中的神经元一般采用 S 型传递函数, 输出层则多采用线性传递函数.

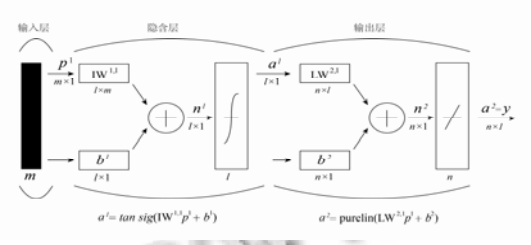

图 1 BP 神经网络结构

图 1 所示为一个典型的 BP 神经网络, 该网络具有一个隐含层, 输入层神经元数目为 m, 隐含层神经元数目为 l, 输出层数目为 n, 隐含层采用 S 型传递函数 tansig, 输出层传递函数为 purelin. BP 神经网络学习算法的指导思想是沿着负梯度方向对权值和阈值进行调整, 通过反向传播把误差分摊给各个神经元的权值和阈值, 这是 BP 神经网络的一大精髓.

本研究构造的神经网络结构为:输入层加上一个常数项单元共有 16 个神经元, 隐含层包括 6 个神经元, 输出层 2 个神经元, 对应预测变量(是否复发). 在得到的神经网络结构图中, 自变量受侵淋巴结数、肿瘤恶性程度、肿瘤大小、有无结节冒、肿块所在象限和有无放疗对模型的贡献明显较大, 且输入层有无结节冒＝1, 有无放疗＝0, 肿瘤大小, 受侵淋巴结数和肿瘤恶性程度这五个节点通过隐含层 H(1:1)节点和输出层有无复发＝1 节点有较强的连接权重, 表明, 有结节冒、无放疗、肿瘤大、受侵淋巴数多、肿瘤恶性程度高的乳腺癌患者复发的可能性更大.

表 2 ANN 训练样本分类表

		预测例数		准确率(%)
		无复发	复发	
观察例数				
	无复发	118	14	89.4
	复发	29	29	50.0
总体准确率				77.4

表 3 ANN 测试样本分类表

		预测例数		准确率(%)
		无复发	复发	
观察例数				
	无复发	58	6	90.6
	复发	17	6	26.1
总体准确率				73.6

应用人工神经网络对全部样本进行预测, 以 0.5 作为预测拟概率分界值, 表 2, 表 3 显示训练样本的灵敏度为 89.4%(118/132), 特异度为 50.0%(29/58), 误诊率为 32.6%(14/43), 漏诊率为 19.7%(29/147), 总体准确率为 77.4%. 测试样本的灵敏度为 90.6%(58/64), 特异度为 26.1%(6/23), 误诊率为 50.0%(6/6), 漏诊率为 29.3%(17/75), 总体准确率为 73.6%,

3.3 支持向量机

支持向量机(SVM)算法是 Vapnik 等人基于基础研究出来的统计学方法, 它采用了结构风险最小化准则, 在最小化样本点误差的同时, 保证结构风险的最小化, 具有最佳的分类效果和泛化效果. 支持向量机每个中间节点对应一个支持向量, 输出是中间节点的线性组合, 这一点在结构上与神经网络类似.

支持向量机的基本思想是希望能够找到一个最优平面, 这个最优平面能够使所有训练样本离该最优分类面的误差最小, 设线性可分的训练样本集为 $\{(x_i,y_i),i=1,2,\cdots l\}$, 其中 x_i 是第 i 个训练样本的输入列向量, y_i 为对应输出值. 分类超平面方程为

$wx+b=0$. 进行归一化后, 使所有样本满足 $|w\chi_i+b|\geq 1$, 也就是说, 离分类平面最近的样本应该满足 $|w\chi_i+b|=1$, 使得分类间隔为 $\frac{2}{\|w\|}$. 能够满足上式的分类面就是最优分类面. 而满足此条件的训练样本就是支持向量. 而要使分类间隔最大, 就要保证 $\|w\|$ 最小, 且 $y_i(w\chi_i+b)\geq 1$, $i=1,2,\mathrm{L}\ l$.

本研究利用 matlabR2012a 创建支持向量机模型, 采用默认的 RBF 核函数, 对乳腺癌复发进行预测, 得出乳腺癌对训练样本和测试样本的预测准确性. 但由于训练集和测试集是随机产生的, 所以程序每次运行的结果都会不同, 某次运行的预测结果如图 2, 图 3 所示, 训练集的准确率为 79.47%, 测试集的准确率为 78.16%.

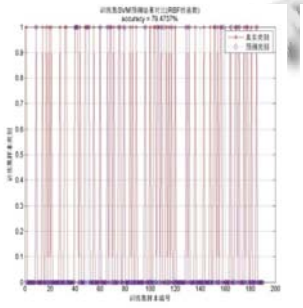

图 2　测试集 SVM 预测

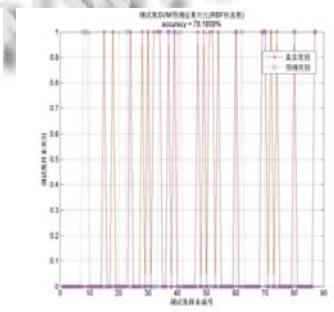

图 3　训练集 SVM 预测

4　预测模型比较

4.1 Logistic 回归与人工神经网络的比较

本研究中, 乳腺癌复发的样本量达到 196 例, 研究影响因素有 9 个, 纳入模型的却只有肿瘤恶性程度、有无结节冒、肿瘤大小这 3 个自变量, 因此 Logistic 回归假设检验及建模过程非常清楚, 而人工神经网络可解释性却相对模糊, 人工神经网络通过隐含层的数学模型来构建输入层与输出层之间的关系. 而与传统的统计方法的自变量相比, 人工神经网络的网络参数要多很多, 各个参数通过各种各样的组合方式影响输出的结果. 通过查看 ROC 曲线中曲线下面积 AUC 的值可以判定 Logistic 回归和 ANN 预测模型结果的比较. 曲线下的面积越大, 诊断的准确性越高.

如表 4 所示, Logistic 回归模型的曲线下面积 AUC 的值为 0.748, 95%可信区间为 0.683~0.814, 人工神经网络模型 AUC 的值为 0.756, 95%可信区间为 0.691~0.821, 以 0.05 进行显著性检验标准, 前者 *P* 值 <0.001, 后者 *P* 值<0.001. 可看出, Logistic 回归模型和 ANN 预测模型曲线下面积 AUC 的值都要大于 0.7, 说明这两个模型的预测都具有一定的准确性. 而且 ANN 预测模型的 AUC 要大于 Logistic 回归模型中 AUC 的值, 这就说明人工神经网络预测模型在对乳腺癌复发的研究中的预测准确率要比 Logistic 回归模型高.

表 4　ANN 模型测试样本分类表

	AUC	*P*	*OR(95%CI)*
logistic回归	0.748	0.000	0.683~0.814
ANN	0.756	0.000	0.691~0.821

4.2 BP 神经网络与支持向量机的比较

支持向量机算法在 SRM 准则下, 避免了传统学习算法的过度学习现象, 能够得到有限样本信息下的最优解, 保证模型具有最佳的泛化能力, 同时, 通过对原问题的对偶化, 能够得到一个全局最优解, 避免了人工神经网络算法中陷入局部极小值的问题, SVM 还将输入空间通过非线性变换转换到高维的特征空间, 使得算法的复杂度也降低了.

在对 Logistic 回归和人工神经网络预测准确率比较后, 进一步运用均方误差 MSE 和决定系数 R^2 来验证两个模型的预测效能. 在 matlabR2012a 下对 BP 神经网络和支持向量机算法进行了仿真, 得到了两个模型在研究乳腺癌复发影响因素应用中的 MSE 和 R^2 的值.

表 5　支持向量机和 BP 神经网络预测结果比较

	训练样本		测试样本	
	MSE	R^2	MSE	R^2
BP 网络	0.16254	0.83795	0.27611	0.72354
SVM	0.02500	0.88298	0.21169	0.79868

支持向量机对乳腺癌复发训练样本的预测的均方误差为 0.025, 决定系数为 0.88298, 对乳腺癌复发测试样本的预测的均方误差为 0.21169, 决定系数为 0.79868. BP 神经网络对乳腺癌复发训练样本的预测的均方误差为 0.16254, 决定系数为 0.83795, 对乳腺癌复发测试样本的预测的均方误差为 0.27611, 决定系数为 0.72354. 可以看出, 支持向量机对训练样本和测试样本进行预测的均方误差 MSE 都比 BP 神经网络的均方误差要小, 决定系数 R^2 的值比 BP 神经网络的都要大, 说明相比之下 SVM 的误差更小, 拟合效度更好. 也正说明了支持向量机比 BP 神经网络具有更好的预

测效能.

5 讨论

Logistic 回归分析属于非线性概率模型中的一种, 它的理论基础完善, 模型所用的假设简单, 不要求自变量完全符合正态分布, 且可解释性强. 但对研究影响因素众多的疾病会存在很大的局限性, 例如建模过程的繁琐, 以及数据结构中的空单元和多重共线性等问题. 与 Logistic 回归不同的是, ANN 模型避开了传统线性处理模式中复杂的参数估计过程, 能够解决一系列不能用函数表达的分类回归问题, 具有自学习和联想存储功能, 能够为每位研究对象给出一个特定的预测结果. 但是, 神经网络算法容易陷于局部最优而不能找到全局最优参数, 从而也会导致模型预测效果不佳. 相比之下, SVM 对样本量没有要求, 可以在有限样本的情况下获得最优解,也不需要像神经网络那样反复的确定网络结构, 具有更好的泛化能力.

研究结果显示:肿瘤大小、有无结节冒、肿瘤恶性程度是乳腺癌术后复发的主要影响因素. 其中 Logistic 回归预测模型总准确率为 75.8%; ANN 预测模型训练样本的总准确率为 77.4%, 测试样本的总准确率为 73.6%; 支持向量机模型中训练样本的总准确率为 79.5%, 测试样本的总准确率为 78.2%. 这表明 SVM 预测模型、ANN 预测模型可获得比 Logistic 回归分析更好的预测效果, 并且 SVM 模型预测效能更高, 能够较准确地根据乳腺癌患者的肿瘤情况判定是否复发, 为个体预测提供了一种新方法. 但是, 为充分验证这三种模型的预测结果有效性, 本研究还存在一些不足. 首先, 本研究建立的预测模型主要从数据上反映乳腺癌复发的发展变化趋势, 若相关参数发生变化或无法获得相应参数, 都无法做出有效预测. 而且模型中的训练样本和测试样本是随机的, 每次的预测结果都不一样, 也会对三者间的效能评价造成一定影响. 另外, 研究构建的预测模型中只纳入了九个输入变量, 在输入的变量选择方面仍可进行深入研究探索. 而且, 由于医疗数据分析的特殊性, 输入变量并不仅仅限于数据变量, 还包括很多医学图像、文字描述等信息, 引入其他医学数据对于提高乳腺癌复发的预测准确率, 降低复发率将产生深远意义.

参考文献

1 Khanfir A, Frikcha M, Kallel F, Meziou M, Trabelsi K, Boudawara T'Mnif J, Daoud J. Breast cancer in young women in the south of Tunisia. Cancer Radiother, 2006, 10(8): 565−571.

2 Siegel R, Naishadham D, Jemal A. Cancer statistics. CA Cancer J Clin, 2012, 62(1): 10−29.

3 东星,于波,周卫东等. maspin 和 bax 联合检测对乳腺癌复发的研究.中国现代普通外科进展,2012,15(1):20-25.

4 Komoik Y, Akiyama F, Iino Y. Analysis of ipsilateral breast tumore currencies after breast-conserving treatment based on the classification of true recurrences and new primary tumors. Breast Cancer, 2005, 12(2): 104−111.

5 Willner J, Kilikuta IC. Locoregional recurrence of breast cancer following mastectomy: Always a fatal event:Results of univariate and multivariate analysis. Int J Radiat Oncol Biol Phys, 1997, 37(4): 853−860.

6 Aberizk WJ, Silver B, Hederson IC. The use of radiotherapy for treatment of isolated locoregional recurrence of breast carcinoma after mastectomy. Cancer, 1986, 58: 1214−1218.

7 Andre F, Slimane K, Bachelot T. Breast cancer with synchronous metastases:trends in survival during a 14-year period. J Clin Oncol, 2004, 22(16): 3302−3308.

8 Elmore JG. Variability in radiologists' interpretations of mammograms. N Engl J Med, 1994, 331(22): 1493−1499.

9 Fentiman IS. Detection and treatment of breast cancer. Martin Duntiz, 1998, 8: 255−262.

10 Kovalerchuk B. Fuzzy logic in computer-aided breast cancer diagnosis: Analysis of lobulation. Artif Intell Med, 1997, 11 (1): 75−85.

11 Laurikkala J, Juhola M. A genetic-based machine learning system to discover the diagnostic rules for female urinary in continence. Comput Methods Programs Biomed, 1998, 55(3): 217−228.

12 岳勇,田考聪,汪洋.BP 神经网络在结核疑似病患者就诊延迟影响因素分析中的应用探讨.中国卫生统计,2007,24(6): 590−592.

13 吴疆,董婷.支持向量机算法用于癌症数据建模.科技技术与工程,2007,7(20):5363−5365.