

基于主成分分析和支持向量机的卵巢癌预测

刘永超^{1,2}, 王卫兵¹, 徐倩³, 郭艳宏², 吴超²

(1. 哈尔滨理工大学 计算机科学与技术学院, 哈尔滨 150080;
2. 齐齐哈尔医学院附属第二医院, 黑龙江 齐齐哈尔 161000;
3. 黑龙江省供电公司哈尔滨分公司配电运检室, 哈尔滨 150001)

摘要: 卵巢癌是我国女性发病率较高的癌症之一, 并且有逐年增加的趋势。基因芯片被广泛用于癌症早期检测, 癌症的早发现可以确保生存率增加超过 97%。本文针对基因芯片的质谱数据, 提出了一种基于主成分分析(PCA)和支持向量机(SVM)的卵巢癌预测方法。重点进行了模型设计和仿真对比实验。实验结果表明: 本文方法在预测精度上达到了 89.1%, 而只消耗 4.791 s 的 CPU 时间, 是一个较好的平衡点。

关键词: 主成分分析; 支持向量机; 卵巢癌预测; 基因芯片

DOI: 10.15938/j.jhust.2021.03.014

中图分类号: TP391.4 文献标志码: A 文章编号: 1007-2683(2021)03-0099-04

A Cancer Prediction Method Based on Principal Component Analysis and Support Vector Machine

LIU Yong-chao^{1,2}, WANG Wei-bing¹, XU Qian³, GUO Yan-hong², WU Chao²

(1. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China;

2. The Second Affiliated Hospital of Qiqihar Medical University, Qiqihar 161000, China;

3. Distribution and Operation Inspection Room of Harbin Branch of Heilongjiang Power Supply Company, Harbin 150001, China)

Abstract: Ovarian cancer is one of the most common cancers contracted by women in China, and it has a tendency to increase year by year. The gene chip is widely applied to the early detection of cancer, which ensures an increase in survival rate over 97%. In this paper, by virtue of a mass spectrometry data of gene chip, an ovarian cancer prediction method based on principal component analysis (PCA) and support vector machine (SVM) is proposed. The model is designed and the simulation comparison experiments are carried out, which are at the core of the research. The experimental results verify that the proposed method has presented the superior performance with prediction accuracy of 89.1% and CPU time of 4.791 s.

Keywords: principal component analysis; support vector machine; ovarian cancer prediction; gene chip

收稿日期: 2019-07-24

基金项目: 国家自然科学基金(61305001)。

作者简介: 刘永超(1986—), 男, 硕士研究生;

徐倩(1969—), 女, 高级工程师。

通信作者: 王卫兵(1964—), 男, 教授, 硕士研究生导师, E-mail: wangweibing163@163.com.

0 引言

癌症是人类生命和健康的主要威胁之一,是世界上最致命的疾病^[1]。在我国,癌症患者逐年增加。在ACS的《2018年全球癌症统计数据》报告中指出:癌症发病率和死亡率我国均居全球第一!2018年在全球新增癌症患者1 810万人,我国占21%;全球死于癌症的人数是960万人,我国占23.9%。卵巢癌是我国女性发病率较高的癌症之一,并且有逐年增加的趋势,虽然可以通过降低风险因素来预防,但仍高发。癌症的诊断是一项极其艰巨而复杂的任务,但早期发现癌症可以确保生存率增加超过97%^[2]。可见,卵巢癌早期发现对我国女性健康尤为重要。

基因芯片是利用核酸互补杂交原理按二维结构将大量探针分子固定于支持物上,与标记的样品分子进行杂交反应,通过对杂交信号的监测分析获取样品分子的数量和序列信息^[3]。目前,广泛应用基因芯片检测各种癌症,但其产生的微阵列数据具有:样本小、数据量大、维度高等特点,很容易导致预测癌症存在与否的分类失败。所以,在预测前必须对数据进行特征提取,这是提高分类质量和降低分类算法的计算复杂性的重要步骤。

目前有很多机器学习的方法用于各种癌症分类。文[4]采用一个无监督的基因过滤算法以降低用于分型计算的数据噪声,首先提出了一个概率模型对样本中的分类结构进行建模,然后基于聚类的结果采用相对熵的方法获得对分类贡献大的基因作为特征基因,最后根据选出的特征基因做聚类来进行癌症分类。文[5]提出了一项关于前列腺癌疾病的调查研究,并使用神经模糊分类系统进行模式识别,以便尽早进行治疗计划。文[6]通过基于内核的学习和特征选择提出了完整的癌症诊断过程。使用SVM-RFE用于预过滤基因。文[7]提出了对数回归的方法来预测和分类癌症。文[8]提出了使用贝叶斯网络方法来预测和分类癌症。文[9]首先使用基因的分类信息指数来选取分类特征基因,去除无关基因进行数据的预处理,然后用支持向量机来进行癌症的分类预测。文[10]提出了一种基于SVM模型诊断卵巢癌,使用模糊粗糙集理论进行特征选择,使用SVM来处理复杂真实的数据,进行快速学习,此方法具有良好的分类性能。本文提出一种基于PCA和SVM的卵巢癌预测方法,使用PCA进行卵巢癌微阵列数据样本的特征提取,使用适合小样本分类的SVM作为区分卵巢癌患者和非卵巢癌患者的分类器。

1 预测模型

假设总共有 n 个卵巢癌患者的微阵列数据记录,每个记录包含 p 个特征属性。根据每个患者的特征数据,通过分类实现预测患者是病态还是正常。本文提出的方法如图1所示,其中,PCA可以在保证一定的信息利用率的基础上提取简化特征属性的维数;SVM用于分类及预测。对卵巢癌患者的微阵列数据而言,一个数据包含上万个或几万个特征属性,可见使用PCA对这样的数据进行降维处理是有必要的。从模型的数据流上说,PCA输出的数据作为SVM的输入数据,看似串行,实这则并行。对单个数据而言,PCA和SVM的工作是串行级的;对整体数据而言,PCA和SVM的工作是同步的,并行级的,SVM处理上一个数据的同时PCA可以降低下一个数据,这样可以提高模型的预测效率。

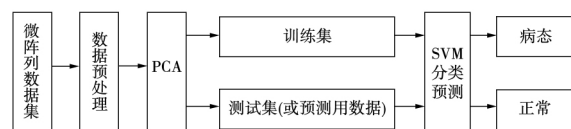


图1 预测模型框图

Fig. 1 Block diagram of prediction model

1.1 数据预处理

首先,处理缺失值和属性的冗余值。历史数据可能存在一些遗漏或冗余的属性值,这可能会影响分类的准确性和计算速度。

其次,数据标准化,消除其尺寸差异的影响。根据原始样本中存在 n 个具有 p 个属性的数据,构造 $n \times p$ 的变量矩阵,每个列在矩阵中标准化,获得归一化值 z_{ij} ,其计算方法为

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (i=1,2,\dots,n \quad j=1,2,\dots,p) \quad (1)$$

$$\bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n} \quad (i=1,2,\dots,n \quad j=1,2,\dots,p) \quad (2)$$

$$s_j = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}} \quad (i=1,2,\dots,n \quad j=1,2,\dots,p) \quad (3)$$

1.2 主成分分析

属性中的太多变量将增加大量的计算负载。PCA可以删除相关变量,并创建一些彼此无关的新变量。更重要的是,原始信息能最大限度地被这些新变量保留。PCA的实现过程描述如下:

计算标准化变量矩阵 Z 的协方差矩阵 C ,即

$$C = \frac{ZZ^T}{n-1} \quad (4)$$

计算协方差矩阵 C 的特征方程 $|C - \lambda I_p| = 0$, 得到特征值 λ_j 和相应的特征向量 α_j , 其中 $j = 1, 2, \dots, p$ 。

根据式(5) 确定主成分 k 的数值, 确保信息利用率高于 95%。

$$\frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^n \lambda_j} \geq 95\% \quad (5)$$

特征值按从大到小排列, 取前 k 个最大特征值对应的特征向量, 得到一个 k 行 p 列的特征向量矩阵 V 。

根据式(6) 计算得到 n 行 k 列的矩阵 X , 这个矩阵 X 就是降维后的变量矩阵, 包含 k 个主要成分的数据。

$$X = (VZ^T)^T = ZV^T \quad (6)$$

1.3 支持向量机

SVM 是一种判别分类器, 由分类超平面所定义。也就是说, 使用标记的训练样本训练模型, 然后通过输出最佳超平面来实现测试样本分类。SVM 描述如下:

卵巢癌预测问题为非线性问题, 所以 SVM 的目标函数 $f(x)$ 表示为

$$f(x) = w^T \cdot \varphi(x) + b \quad (7)$$

其中: x 为输入; w 为权向量; b 为偏置向量。把输入数据映射到高维数据空间。

利用极大化间隔的求解思想, 最终得到分类决策函数为

$$f(x) = \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \quad (8)$$

其中: $K(x_i, x)$ 为核函数, 表示映射到的高维特征空间的两个点的内积 $\langle \varphi(x_i) \cdot \varphi(x) \rangle$ 。这里选择学习能力强且误差小的 Gauss 径向基(RBF) 核函数为

$$K(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right) \quad (9)$$

其中, σ 为核函数系数, 具有高的灵活性。

2 仿真实验与分析

2.1 数据集

本文所有仿真实验使用同一个数据集, 是来自于 AICR 的卵巢癌微阵列质谱数据集, 样本共有 253 个, 包括卵巢癌患者阳性(病态) 和阴性(健康), 且每个样本的属性有 15 000 之多。在这 253 个卵巢癌数据样本中随机选其中的 170 个为训练集, 剩余的 83 个为测试集, 卵巢癌微阵列质谱数据如表 1 所示。

表 1 卵巢癌微阵列质谱数据

Tab. 1 Mass spectrometry data of ovarian cancer microarray

样本	数据集	训练集	测试集
样本总数	253	170	83
阳性样本数	91	61	30
阴性样本数	162	109	53

2.2 仿真实验设计

设计 3 个仿真实验来证明本文提出的基于 PCA 和 SVM 的癌症预测方法的有效性, 即主成分 k 值确定、预测精度和算法效率。

仿真实验中的计算使用安装了 Libsvm-3.21 工具包的 MATLAB R2010b 软件完成。模型中的 SVM 参数随机选择, 要求 $C \in [0, 100]$ 、 $\sigma \in [0.1, 10]$, 用 MATLAB 仿真的时候用经验值或试凑法调节。

仿真实验都是在实验室的 PC 机上完成。PC 机的 OS 为 Win7 Sp1, 硬件运行环境为: 处理器(CPU): Intel(R) Core(TM) i3-3110 CPU @ 2.4 GHz, 物理四核; 内存(RAM): 8.00 GB; 磁盘驱动器: ATA WDC WD5000LPVX-0; 显示适配器: NVIDIA GeForce GT 635M。

2.3 结果与分析

原始数据集总共有 15 000 个属性。通过本文 1.2 节计算协方差矩阵特征方程的特征值和特征向量的方法进行了主成分分析, 经过分析, 数据的特征值大大减少, 290 个属性可以保证 100% 的信息利用率(通过计算贡献率可以近似获得)。在这个实验中, 我们相信 95% 的信息利用率可以确保结果的正确性, 即最少由 98 个数据属性表示。然后, 根据式(5) 计算主成分 k 值。信息利用率选择如图 2 所示。在图 2 中, 属性个数被表示为横轴, 属性个数对应的信息利用率被表示为纵轴。

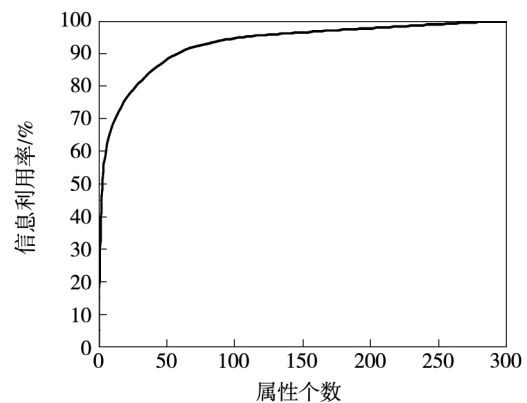


图 2 信息利用率选择

Fig. 2 Information utilization selection

在数据集上,应用所提出的方法进行预测实验,并与数据集上的 KNN 和 ESOM 等典型分类方法在预测精度上进行了比较,如图 3 所示。可以看出,预测精度上最高是本文所提出的方法,预测精度为 89.1%,高于 KNN 的 71.7% 和 ESOM 的 62.5%。

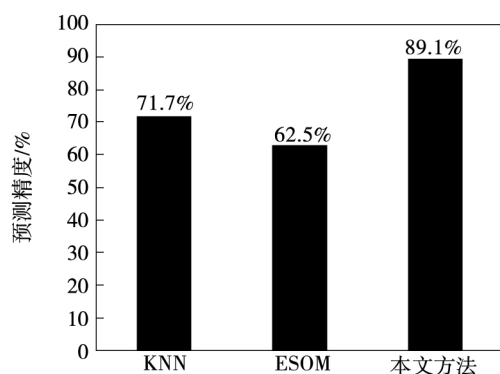


图3 预测精度对比

Fig. 3 Comparison of prediction accuracy

通过消耗的 CPU 时间来比较本文方法、KNN 和 ESOM 的效率,CPU 时间对比如图 4 所示。由于本文中使用的数据集拥有大量的属性值,因此需要花费很长时间来训练和预测分类结果。在这 3 种方法中,KNN 是最耗时的,所需的 CPU 时间高达 1 225.433 s。由于本文方法和 ESOM 都对数据集进行了降维处理,所以消耗的 CPU 时间较少。本文方法应用了主成分分析,效率得到了极大的提高,只消耗 4.791 s 的 CPU 时间。

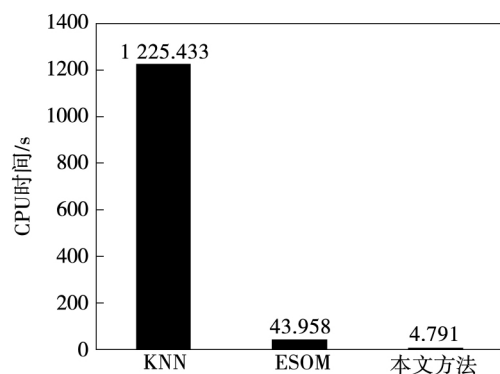


图4 CPU 时间对比

Fig. 4 CPU time comparison

3 结 论

根据微阵列质谱数据的特点,本文提出一种基于主成分分析(PCA)和支持向量机(SVM)的卵巢

癌预测方法。使用 PCA 在预测前对数据进行特征提取,来提高分类质量和降低分类算法的计算复杂性。使用 SVM 作为区分卵巢癌阳性和阴性的分类器,其更适合小样本分类。仿真实验结果表明:本文方法在预测精度上达到了 89.1%,而只消耗 4.791 s 的 CPU 时间,是一个较好的平衡点,但仍需要在其他卵巢癌数据集上进行验证。同时,本文方法也可以应用在其他高维数据分类的领域。

参 考 文 献:

- [1] SIEGEL R L, MILLER K D, FEDEWA S A, et al. Colorectal Cancer Statistics, 2017. [J]. CA: A Cancer Journal for Clinicians, 2017, 67(3): 177.
- [2] MILLER K D, SIEGEL R L, LIN C C, et al. Cancer Treatment and Survivorship Statistics, 2016 [J]. CA: A Cancer Journal for Clinicians, 2016, 66(4): 271.
- [3] 沈锐, 谢杨, 李磊, 等. 基因芯片技术在于癌症诊断研究发展 [J]. 世界最新医学信息文摘, 2017, 17(33): 124.
- [4] SHEN Rui, XIE Yang, LI Lei, et al. Gene Chip Technology Lies in the Development of Cancer Diagnosis Research [J]. World Latest Medicine Information, 2017, 17(33): 124.
- [5] 李泽, 包雷, 黄英武, 等. 基于基因表达谱的肿瘤分型和特征基因选取 [J]. 生物物理学报, 2002, 33(4): 413.
- [6] LI Ze, BAO Lei, HUANG Yingwu, et al. Cancer Subtype Discovery and Informative Gene Identification with Gene Expression Profiles [J]. Acta Biophysica Sinica, 2002, 33(4): 413.
- [7] SUBRATA Kar, MAJUMDER D. Dutta. An Investigative Study on Early Diagnosis of Prostate Cancer Using Neuro-fuzzy Classification System for Pattern Recognition [J]. International Journal of Fuzzy Systems, 2016, 19(2): 1.
- [8] MEDJAHED, SEYYID Ahmed. Kernel-based Learning and Feature selection Analysis for Cancer Diagnosis [J]. Applied Soft Computing, 2017, 51: 39.
- [9] ZHOU X B, LIU K Y, WONG S T C. Cancer Classification and Prediction Using Logistic Regression with Bayesian Gene Selection [J]. Journal of Biomedical Informatics, 2004, 37: 249.
- [10] HELMAN P, VEROFF R, ATLAS S R, et al. A Bayesian Network Classification Methodology for Gene Expression Data [J]. J. Comput. Biol., 2004, 11: 581.
- [11] 李颖新, 阮晓钢. 基于基因表达谱的肿瘤亚型识别与分类特征基因选取研究 [J]. 电子学报, 2005, 33(4): 651.
- [12] LI Yingxin, RUAN Xiaogang. Cancer Subtype Identification and Feature Gene Selection with Gene Expression Profiles [J]. Acta Electronica Sinica, 2005, 33(4): 651.
- [13] BADRIA, F A. A Framework for Ovarian Cancer Diagnosis Based on Amino Acids Using Fuzzy-rough Sets with SVM [J]. International Conference on Advanced Machine Learning Technologies and Applications, 2014.

(编辑: 王 萍)