Hans Journal of Data Mining 数据挖掘, 2018, 8(3), 89-95
Published Online July 2018 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2018.83010

Hans汉斯

Auxiliary Diagnosis of Breast Cancer Based on Kernel Principal Component Analysis Support Vector Machine

Keke Deng*, Wenqiang Luo, Jing Zhao, Ying Cao

School of Mathematics and Physics, China University of Geosciences, Wuhan Hubei
Email: *1427923689@qq.com, wqluo@cug.edu.cn

Received: Jun 3rd, 2018; accepted: Jun. 27th, 2018; published: Jul. 4th, 2018

Abstract

Kernel principal component analysis (KPCA) was used to extract the feature factors of breast cancer. The principal components were obtained as support vector machine (SVM) feature vector to establish support vector machine model. The model parameters were selected and optimized respectively by PSO and GA. KPCA-PSO-SVM model and KPCA-GA-SVM model were constructed to classify the breast masses as malignant. The experimental results show that the KPCA-PSO-SVM model and KPCA-GA-SVM model both improve the classification accuracy and the operation speed compared with the PSO-SVM model and GA-SVM model, which shows that the principal component analysis support vector machine can be used in the auxiliary diagnosis of breast cancer and can provide strong decision-making support for the diagnosis of breast cancer in medical institutions.

Keywords

Support Vector Machine (SVM), Kernel Principal Component Analysis (KPCA), Auxiliary Diagnosis, Classification

基于核主成分分析支持向量机的乳腺癌辅助诊断

邓珂珂*，罗文强，赵　静，曹　颖

中国地质大学，数学与物理学院，湖北 武汉
Email: *1427923689@qq.com, wqluo@cug.edu.cn

收稿日期：2018年6月3日；录用日期：2018年6月27日；发布日期：2018年7月4日

*通讯作者。

文章引用：邓珂珂，罗文强，赵静，曹颖. 基于核主成分分析支持向量机的乳腺癌辅助诊断[J]. 数据挖掘, 2018, 8(3): 89-95. DOI: 10.12677/hjdm.2018.83010

摘　要

本文利用核主成分分析法对乳腺癌的影响因子进行特征提取，以获取的主成分作为支持向量机的特征向量建立支持向量机模型，其中模型参数分别通过粒子群算法和遗传算法进行选择优化，分别构建出KPCA-PSO-SVM模型和KPCA-GA-SVM模型，对乳腺肿块是否为恶性进行二分类。实验结果显示：KPCA-PSO-SVM模型和KPCA-GA-SVM模型相比PSO-SVM模型和GA-SVM模型在分类准确率方面和运行速度方面均有所提高，表明核主成分分析支持向量机可以用于乳腺癌疾病的辅助诊断，可以为医疗机构对乳腺癌疾病的诊断提供有力的决策支持。

关键词

支持向量机，核主成分分析，辅助诊断，分类

1. 引言

乳腺癌是女性最易患的恶性肿瘤之一。在我国，其发病率以每年 3%的递增速度发展，且有日益增长之势，其死亡率增幅已达到 38.9%。同时更令人担忧的是发病年龄也呈现出年轻化的趋势，形势非常严峻[1] [2] [3]。乳腺癌能否治愈，关键在于发现的时期。早期诊断和早期治疗是有可能痊愈的。现代医学仍无法治愈晚期乳腺癌，因此应做到早期发现、早期诊断，还要防止误诊和漏诊，正确的鉴别诊断对降低死亡率是极其重要的[4]。

利用机器学习方法来诊断疾病是目前发展较快的一个应用分支[5]。SVM (Support Vector Machine，支持向量机)是一个被广泛使用的机器学习算法，它是在 1995 年由 Vapnik 等人提出的[6]。该理论是基于统计学 VC 维理论和结构风险最小原理提出的，它在解决小样本、非线性和高维模式识别方面展现出了较强的优势，现已广泛的应用于模式识别、回归估计等领域[7] [8] [9]。但 SVM 容易受输入变量过多和噪声的影响，过多的变量会使得运行的时间过长，变量之间的相关性和输入数据中存在的噪声会使得模型的稳定性和分类识别率降低。鉴于 KPCA (Kernel Principal Component Analysis，核主成分分析)不仅具有降维、除噪的优势，而且能够提取非线性的特征信息，提高数据质量[10]。因此本文把 KPCA 引入到 SVM 中，建立 KPCA-SVM 模型，实现乳腺癌的辅助诊断。

2. SVM 分类器

SVM 的工作原理：以二分类为例，设样本集设为 $\{(x_i, y_i), i=1,2,\cdots,l\}$，其中 $x_i \in R^l$ 表示输入变量，$y_i = \{-1,1\}$ 表示输出标量，l 为样本集个数。通过引入非线性映射函数 $\varphi(x)$，将处于低维空间的输入变量映射到高维空间，可以构造出一个最优分类超平面：

$$f(x) = w \cdot \varphi(x) + b = 0, \tag{1}$$

其中 w 表示权重向量，b 表示最优超平面位移。

为了最小化结构风险，构造出的最优超平面需满足以下约束条件：

$$y_i\left(\left(\omega\cdot\varphi\left(x_i\right)\right)+b\right)\geq 1, i=1,2,\cdots,l. \tag{2}$$

若构造的最优分类超平面仍存在少量分类错误，那么可以引入非松弛变量 ξ_i 来量化分类器的误差，此时分类超平面最优化问题可表述为：

$$\min\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}y_i\left(\omega\cdot\varphi\left(x_i\right)+b\right)+\xi_i\geq 1,\xi_i\geq 0, i=1,2,\cdots,l. \tag{3}$$

其中 C 为惩罚参数，表示对错误分类的惩罚代价。C 越大，惩罚代价越大。$\|\omega\|^2$ 为结构风险，表示问题的复杂程度；$C\sum_{i=1}^{n}\xi_i$ 为经验风险，表示问题的误差。

对于该二次规划问题，采用拉格朗日乘子法将其转化为对偶形式：

$$\begin{aligned}&\min\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j\left(\varphi\left(x_i\right)\cdot\varphi\left(x_j\right)\right)-\sum_{j=1}^{l}\alpha_j\\&\text{s.t.}\sum_{i=1}^{l}y_i\alpha_i=0\\&\left(0\leq\alpha_i\leq C, i=1,2,\cdots,l\right)\end{aligned} \tag{4}$$

在上式中引入核函数，可转化为：

$$\begin{aligned}&\min\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_jK\left(x_i,x_j\right)-\sum_{j=1}^{l}\alpha_j\\&\text{s.t.}\sum_{i=1}^{l}y_i\alpha_i=0\\&\left(0\leq\alpha_i\leq C, i=1,2,\cdots,l\right)\end{aligned} \tag{5}$$

其中 $K\left(x_i,x_j\right)$ 称为核函数，$K\left(x_i,x_j\right)=\varphi\left(x_i\right)\cdot\varphi\left(x_j\right)$，最终构造出的最优分类超平面为：

$$f\left(x\right)=\sum_{i=1}^{n}\alpha_iy_iK\left(x_i,x\right)+b=0 \tag{6}$$

SVM 采用不同的核函数就会生成不同的 SVM 分类器，目前支持向量机常用的核函数主要有：径向基(RBF)核函数 $K\left(x_i,x\right)=\exp\left(-\|x_i-x\|^2/2\sigma^2\right)$、多项式核函数 $K\left(x_i,x\right)=\left(x_i\cdot x+1\right)^d$、线性核函数 $K\left(x_i,x\right)=x_i\cdot x$、Sigmoid 核函数 $K\left(x_i,x\right)=\tanh\left(\gamma x_i\cdot x+r\right),\gamma>0$ 等。其中径向基核函数只需要确定一个核参数，并且相关研究表明：它是一种在分类问题上表现较好的核函数[11]，因此本文采用径向基核函数作为 SVM 的核函数。SVM 常采用 GA (Genetic Algorithm，遗传算法)和 PSO (Particle Swarm Optimization，粒子群算法)对参数 (c,g) 进行迭代寻优。遗传算法是根据生物进化思想而启发得出的一种全局优化算法，它采用选择，交叉，变异操作对群体中的个体进行优胜劣汰操作，在问题空间中搜索最优解。粒子群也是一种基于群体叠代的算法，但它没有遗传算法用的交叉以及变异，而是粒子在解空间中追随最优的粒子进行搜索的一种优化算法。

3. KPCA 算法

设原始空间 R^m 中有 n 个样本 $x_1,x_2,\cdots,x_n$，由这 n 个样本构成的数据矩阵为 X，利用非线性映射函数 Φ 将原始空间 X 映射到高维特征空间 $F=\left\{h\left(x\right)\middle|x\in X\right\}$ 中，为方便讨论，设其映射空间的维数为 M (其中 M 远大于 m)。将中心化后的数据记为 $\tilde{\Phi}\left(x_i\right)$，即

$$\tilde{\Phi}(x_i)=\Phi(x_i)-\bar{\Phi} \tag{7}$$

其中 $\bar{\Phi}=\frac{1}{n}\sum_{i}^{n}\Phi(x_i)$。

中心化的数据 $\tilde{\Phi}(x_i)$ 的协方差矩阵为

$$\bar{C}=\frac{1}{n}\sum_{k=1}^{n}\tilde{\Phi}(x_k)\tilde{\Phi}(x_k)^{\mathrm{T}}=\frac{1}{n}\tilde{\Phi}(X)\tilde{\Phi}(X)^{\mathrm{T}} \tag{8}$$

将中心化后的数据对应的核矩阵记为 $\tilde{K}$，且定义 $\tilde{K}_{ij}=\left\langle\tilde{\Phi}(x_i),\tilde{\Phi}(x_j)\right\rangle=\tilde{\Phi}(x_i)^{\mathrm{T}}\tilde{\Phi}(x_j)$，且有：$\tilde{K}=K-1_nK-K1_n+1_nK1_n$，其中 K 为原始数据对应的核矩阵，1_n 为 $n\times n$ 的矩阵，其中每一个元素都是 $\frac{1}{n}$ (n 为样本数目)。对 $\bar{C}$ 进行特征向量分析，设 $\bar{C}$ 的特征值和对应的特征向量分别为 λ 和 V，设 v_k 是 V 的第 k 个特征矢量，对其进行归一化处理，即 $v_kv_k=1$，且有：

$$\bar{C}V=\lambda V \tag{9}$$

设 $\alpha_k=(\alpha_{k1},\cdots,\alpha_{kn})^{\mathrm{T}}$ 为中心化的核矩阵 $\tilde{K}$ 的第 K 个特征向量(对应的特征值为 $\tilde{\lambda}_k$)。则有 $\lambda_k=\frac{\tilde{\lambda}_k}{n}$, $v_k=\tilde{\Phi}\alpha_k$。

将 α_k 进行归一化处理，则有：

$$\|\alpha_k\|=\frac{1}{\sqrt{\tilde{\lambda}_k}} \tag{10}$$

则可得到原始空间任一样本 x 的映射数据 $\tilde{\Phi}(x)$ 在特征向量 v_k 方向上的投影为：

$$v_k^{\mathrm{T}}\tilde{\Phi}(x)=\sum_{i}^{n}\frac{\alpha_{ki}}{\sqrt{\tilde{\lambda}_k}}\left[\tilde{\Phi}(x_i)^{\mathrm{T}}\cdot\tilde{\Phi}(x)\right]=\sum_{i}^{n}\frac{\alpha_{ki}}{\sqrt{\tilde{\lambda}_k}}\tilde{K}(x_i,x) \tag{11}$$

即原始空间中任一样本数据 x 的第 K 维的非线性主成分为：

$$t_k=v_k^{\mathrm{T}}\tilde{\Phi}(x)=\sum_{i}^{n}\frac{\alpha_{ki}}{\sqrt{\tilde{\lambda}_k}}\tilde{K}(x_i,x) \tag{12}$$

4. KPCA-SVM 模型

4.1. 详细流程如下

步骤 1：数据归一化。为了消除量纲对结果造成影响，对初始输入变量数据通过 $y=\frac{x-x_{\min}}{x_{\max}-x_{\min}}$ 的方法进行处理，将初始输入数据归一化到[0,1]之间。其中 x 是某一特征的原始数值，$x_{\max}$ 是对应特征的最大值，$x_{\min}$ 是对应特征的最小值。y 为归一化后的数值。同时将乳腺癌数据中标签 M (恶性)和 B (良性)分别标记为−1 和 1。

步骤 2：随机抽样。利用随机抽样的方法随机选取 285 组数据作为训练数据，其余 284 组数据作为测试数据。

步骤 3：用 KPCA 方法降维。对 Breast Cancer Wisconsin (Diagnostic)的全部数据的 30 个初始影响因子应用 KPCA 算法降维，将降维后的数据的前 285 组作为模型训练的新样本，其余 284 组作为模型的检验数据。

步骤 4：参数寻优。分别采用 PSO 算法和 GA 算法对参数(c,g)进行迭代寻优，建立乳腺癌诊断结果与影响因素之间的非线性模型。

步骤 5：利用所建立的模型对检验样本进行预测，检验模型的预测能力，分别得到 KPCA-PSO-SVM (核主成分分析和粒子群优化支持向量机)模型和 KPCA-GA-SVM (核主成分分析和遗传算法优化支持向量机)模型的分类准确率。

4.2. KPCA-SVM 模型的介绍

KPCA-PSO-SVM 模型是先用 KPCA 算法对乳腺癌数据进行特征提取，提取原始数据中的非线性的特征信息，这些特征向量去除了原始数据中的噪声，数据质量有所提高。将 KPCA 算法提取出来的特征信息作为支持向量机的特征向量，SVM 基于这些特征向量进行建模，在建立模型的过程中采用 PSO 算法对 SVM 的参数进行迭代寻优。KPCA-GA-SVM 模型与 KPCA-PSO-SVM 模型类似，只是 KPCA-GA-SVM 模型是采用 GA 算法对 SVM 的参数进行迭代寻优。

5. 仿真实验

5.1. 数据源

数据集采用 UCI 网站(http://archive.ics.uci.edu/ml/datasets.html)中的 Breast Cancer Wisconsin (Diagnostic)数据。该数据集是美国威斯康辛大学麦迪逊分校采用针吸细胞学方法所得到的乳腺肿块的样本。该样本具有 10 个属性，分别为半径、结构、周长、面积、圆球度、体密度、轮廓的凹陷度、轮廓的凹陷点数量、对称性、碎形维度。这 10 个属性又分别从均值、标准差、最大值三方面来描述，所以总共有 30 个特征。该数据集不包含缺失值，共有 569 组样本，构成一个 569×30 的样本空间。同时对应的得到了 569×1 的标签向量矩阵，分别用 B 表示良性(benign)，M 表示恶性(malignant)。在数据预处理时，使用 1 表示良性样本标签，使用−1 表示恶性样本标签。并随机选择 285 组样本作为训练集，其余 284 组样本作为测试集。

5.2. 实验设置

本文中参数寻优的方法分别采用粒子群算法、遗传算法，惩罚参数 c 和核参数 g 的参数范围均设置为默认。除此之外，设定 KPCA 的核函数为高斯核函数，参数 sigma 设定为 7.5。本文采用了台湾林智仁教授开发的 Libsvm 3.22 工具箱来进行实验。实验平台为 Intel 奔腾双核处理器(1.8 GHz)，4 GB 内存，Windows 7 操作系统，所用软件为 Matlab 2014a。由于模型计算结果具有一定的波动性，为了公平性和准确性，我们随机取十次样本，对每次取得的样本分别建立 SVM 模型、KPCA-PSO-SVM 模型和 KPCA-GA-SVM 模型，并分别计算 SVM、KPCA-PSO-SVM 模型和 KPCA-GA-SVM 模型的分类准确率和运行时间，取十次随机抽样模型下分类准确率的平均值作为最终的平均分类准确率，取十次运行时间的平均值作为最终的平均运行时间。

5.3. 实验结果分析

为了使分类的结果得到更好的评估，引入灵敏度(Sen)、特异度(Spe)、F 分数和平均分类准确率来评价 SVM、KPCA-PSO-SVM 和 KPCA-GA-SVM 三个分类器的性能[12]。灵敏度是指将实际有病的人正确地判定为有病的比例，灵敏度越高，表示该模型越好。特异度是指将实际无病的人正确地判定为无病的比例，特异度越高，表示该模型越好。F 分数是判定该模型优良性的指标，F 值越大，表示该模型越好。

下面给出灵敏度(Sen)、特异度(Spe)、F 分数的计算公式。

Table 1. Comparison of experimental results
表 1. 实验结果对比

模型	参数(c,g)	准确率/%	平均运行时间/s	灵敏度/%	特异度/%	F 分数/%
PSO-SVM	(17.23,0.01)	97.89	173.36	100	94.03	87.37
KPCA-PSO-SVM	(0.48,68.92)	98.60	101.04	100	96.43	92.63

Table 2. Comparison of experimental results
表 2. 实验结果对比

模型	参数(c,g)	准确率/%	平均运行时间/s	灵敏度/%	特异度/%	F 分数/%
GA-SVM	(25.36,1.38)	96.13	72.22	96.59	96.87	95.98
KPCA-GA-SVM	(16.85,9.63)	97.54	24.27	98.86	98.03	97.12

$$Sen = \frac{TP}{TP + FN} \tag{13}$$

$$Spe = \frac{TN}{TN + FP} \tag{14}$$

$$F\text{-score} = 2 \times \frac{TP}{TP + FP} \times \frac{TP}{TP + FN} \div \left(\frac{TP}{TP + FP} + \frac{TP}{TP + FN} \right) \tag{15}$$

在(13) (14) (15)式中 TP 是指在工作集中实际是良性样本，预测结果为良性样本的数目。FN 是指在工作集中实际是良性样本(即标签为 1 的样本)，预测结果是恶性样本(即标签为−1 的样本)的数目；TN 是指在工作集中实际是恶性样本，预测结果为恶性样本的数目；FP 是在工作集中实际是恶性样本，预测结果为良性样本的数目[12]。

SVM、KPCA-PSO-SVM 和 KPCA-GA-SVM 分类测试结果如表所示，其中准确率指测试数据的平均分类准确率。

由表 1 可知，KPCA-PSO-SVM 分类器比 PSO-SVM 分类器在平均分类准确率上提高了 0.71%，KPCA-PSO-SVM 模型在运行时间方面相比 PSO-SVM 模型缩短了 72.32 秒。由表 2 可知，KPCA-GA-SVM 分类器比 GA-SVM 分类器在平均分类准确率上提高了 1.41%，KPCA-GA-SVM 模型在运行时间方面相比 GA-SVM 模型缩短了 47.95 秒。因此在乳腺癌的辅助诊断中 KPCA-SVM 方法比 SVM 方法更为有效。

6. 结语

本文使用了 SVM、KPCA-PSO-SVM 和 KPCA-GA-SVM 三种分类方法，并对其结果进行对比。从实验结果来看，不论是用粒子群优化算法还是遗传算法，KPCA-SVM 在乳腺癌辅助诊断方面的效果均要优于 SVM。KPCA 算法不仅提高了 SVM 的建模效率，并且提高了模型预测的准确率、灵敏度和特异度，获得了较好的识别结果。以上结果表明 KPCA-SVM 算法可以用于乳腺癌疾病的辅助诊断，为乳腺癌的诊断提供了一种新途径。

参考文献

[1] 彭建兵, 焦莉. 基于极小 T-不变量增加的 Petri 网可达性分析[J]. 计算机应用研究, 2010, 27(10): 3798-3802.

[2] 江志斌. Petri 网及其在制造系统建模与控制中的应用[M]. 北京: 机械工业出版社, 2004.

[3] 白杨, 朱金福. 基于随机 Petri 网的航空货运出港系统分析[J]. 数理统计与管理, 2012, 31(2): 199-206.

[4] 刘兴华, 蔡从中, 袁前飞, 肖汉光, 孔春阳. 基于支持向量机的乳腺癌辅助诊断[J]. 重庆大学学报(自然科学版): 2007, 30(6): 140-144.

[5] 章永来, 史海波, 尚文利, 周晓锋, 纪晓楠. 面向乳腺癌辅助诊断的改进支持向量机方法[J]. 计算机应用研究, 2013, 30(8): 2373-2376.

[6] Vapnik, V.N. (1997) The Nature of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, **8**, 1564. https://doi.org/10.1109/TNN.1997.641482

[7] 刘洛霞. 基于 SVM 的多变量函数回归分析研究[J]. 电光与控制, 2013, 20(6): 50-57.

[8] Huanrui, H. (2016) New Mixed Kernel Functions of SVM Used in Pattern Recognition. *Cybernetics & Information Technologies*, **16**, 5-14. https://doi.org/10.1515/cait-2016-0047

[9] Korpela, J., Miyaji, R., Maekawa, T., Nozaki, K. and Tamagawa, H. (2016) Toothbrushing Performance Evaluation Using Smartphone Audio Based on Hybrid HMM-Recognition/SVM-Regression Model. *Journal of Information Processing*, **24**, 302-313. https://doi.org/10.2197/ipsjjip.24.302

[10] 彭令, 牛瑞卿, 赵艳南, 邓清禄. 基于核主成分分析和粒子群优化支持向量机的滑坡位移预测[J]. 武汉大学学报(信息科学版), 2013, 38(2): 148-152.

[11] 宋晖, 薛云, 张良均. 基于 SVM 分类问题的核函数选择仿真研究[J]. 计算机与现代化, 2011, 2011(8): 133-136.

[12] 袁前飞, 蔡从中, 肖汉光, 刘兴华, 孔春阳. 基于支持向量机的乳腺癌预后状态预测和疗效评估[J]. 北京生物医学工程, 2007, 26(4): 372-376.