

网格搜索的支持向量机方法在乳腺癌诊断中的应用

曹君杰, 冯爱芬*, 常芳欣, 杨双杨, 蒋智涵, 王世杰

河南科技大学数学与统计学院, 河南 洛阳

收稿日期: 2025年4月21日; 录用日期: 2025年5月13日; 发布日期: 2025年5月23日

摘要

支持向量机(SVM)作为一种高效的分类模型, 其性能在很大程度上取决于超参数的选择。本文利用支持向量机分类原理对乳腺癌诊断问题进行建模分析, 采用网格搜索方法和五折交叉验证的方法对SVM的超参数进行优化选择, 评估优化后模型的性能。实验结果表明, 得到模型的准确率为98.24%、召回率为96.83%、F1分数为0.976、ROC AUC为0.9965。表明采用最优超参数组合的SVM模型在乳腺癌数据集上取得了很好的效果。

关键词

乳腺癌诊断, 支持向量机, 网格搜索, 超参数选择, 交叉验证

Application of Grid Search-Optimized Support Vector Machine in Breast Cancer Diagnosis

Junjie Cao, Aifen Feng*, Fangxin Chang, Shuangyang Yang, Zhihan Jiang, Shijie Wang

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang Henan

Received: Apr. 21st, 2025; accepted: May 13th, 2025; published: May 23rd, 2025

Abstract

As an efficient classification model, the performance of Support Vector Machine (SVM) largely depends on the selection of hyperparameters. This study employs SVM classification principles for modeling

*通讯作者。

文章引用: 曹君杰, 冯爱芬, 常芳欣, 杨双杨, 蒋智涵, 王世杰. 网格搜索的支持向量机方法在乳腺癌诊断中的应用[J]. 应用数学进展, 2025, 14(5): 238-243. DOI: 10.12677/aam.2025.145252

and analyzing breast cancer diagnosis problems, utilizing grid search method and five-fold cross-validation to optimize the selection of SVM hyperparameters and evaluate the performance of the optimized model. Experimental results demonstrate that the model achieved an accuracy of 98.24%, recall of 96.83%, F1-score of 0.976, and ROC AUC of 0.9965. The findings indicate that the SVM model with optimal hyperparameter combinations achieved excellent performance on the breast cancer dataset.

Keywords

Breast Cancer Diagnosis, Support Vector Machine, Grid Search, Hyperparameter Selection, Cross-Validation

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着人工智能的发展,全球信息化进一步推进,数据挖掘和机器学习技术开始逐渐登上历史的舞台,从理论研究走向生活的方方面面,其中智慧医疗得到了长足发展。由于医疗行业的特殊性,时时刻刻都在产生海量的医疗数据,数据挖掘和机器学习技术为这些海量医疗数据的分析和应用提供了新的思路 and 手段[1] [2]。在人工智能化、信息全球化的今天,利用人工智能和数据挖掘中相对较为成熟的算法与医疗数据的特点进行结合来处理医疗数据,更好地满足人民的健康医疗需求。目前,乳腺癌主要通过临床触诊、X 射线、超声波、CT、核磁共振、细针穿刺细胞病理学检查等方法进行诊断。但由于医学图像质量低、患者个体差异和医护人员疏忽等,上述的传统诊断方法存在漏诊、误诊的情况。随着计算机技术和医疗数据智能化技术的发展,计算机辅助诊断技术应运而生,并得到了广泛研究,目前已取得了较好的诊断效果。采用神经网络、SVM、K 近邻算法(K-Nearest Neighbor algorithm, KNN)等数据挖掘技术,对乳腺癌各项生理指标进行数学建模,从而区分出乳腺癌的良性与恶性[3] [4]。数据挖掘与医疗诊断的结合,一定程度上将诊断过程从复杂的病理分析中解救出来,可以有效避免医护人员主观因素的影响,从而减少乳腺癌的错诊和漏诊。

支持向量机(Support Vector Machines, SVM)的初始模型源于 1992 年由 Boser 等人提出的最优边界分类训练算法;1996 年, Vapnik 等正式提出基于统计学习理论的支持向量机算法[5] [6]。支持向量机是基于统计学习理论提出的着重解决小样本、非线性和高维度的新型机器学习方法,是一种有效克服“维数灾难”、“过学习”、“局部极值”等难题的机器学习方法。因其坚实的理论基础、模型的稀疏性以及极值解必是全局最优解等优点被广泛应用于模式识别和回归预测两大领域中,并有效推广于预测和综合评价等领域和学科,同时有效解决了神经网络局部极值问题。如果将 SVM 数据挖掘方法与各个研究领域的专业知识相结合,可将 SVM 扩展到不同的领域,但大量的实践证明, SVM 惩罚系数、核函数的选取以及参数的优化对分类性能有极大的影响,因此如何利用科学有效的方法实现对 SVM 参数的优化是个很重要的问题[7]-[9]。

SVM 其性能高度依赖于超参数的选择。超参数选择方法有随机搜索(Random Search)、贝叶斯优化(Bayesian Optimization)、经验法则(Rule of Thumb/Empirical Rule)、网格搜索(Grid Search)。其中网格搜索具有很强的实用性,其通过穷举搜索超参数空间中的不同组合,以找到最优的超参数设置。本文利用支持向量机分类原理对乳腺癌诊断问题进行建模分析,采用网格搜索方法对 SVM 的超参数进行优化选择,

并评估优化后模型的性能。

2. 支持向量机的基本原理

支持向量机(Support Vector Machine, SVM)是由 Vapnik 等人基于统计学习理论中的结构风险最小化和 VC 维理论提出的一种机器学习方法。近年来, SVM 已经被广泛应用于多个领域, 如故障检测、医疗诊断和图像分类等。设数据集为 (x_i, y_i) , 其中 $i=1, 2, \dots, N$, $x_i \in R^n$, $y_i \in (1, -1)$, y_i 为标签值, N 为样本量。SVM 的核心目标是找到一个能够最大化分类间隔的决策边界, 即超平面 $w^T x + b$ 。当数据是线性可分时, SVM 可以找到决策边界。然而, 在现实世界中, 数据往往不是完全线性可分的。为了解决这一问题, SVM 引入了软间隔的概念, 即允许一些数据点违反间隔规则, 以此让模型拥有更好的泛化能力。SVM 通过决策函数 $f(x) = \text{sgn}(w^T x + b)$ 进行分类, 其中 w 和 b 分别是超平面的系数和偏置项。根据 SVM 的目标可以得到如下含有不等式约束的凸二次规划问题。

$$\begin{aligned} \min \quad & \frac{1}{2} \|W\|^2 + \left(C \sum_{i=1}^N \xi_i \right) \\ \text{s.t.} \quad & y_i [W^T x_i + b] \geq 1 - \xi_i, i = 1, \dots, n \\ & \xi_i \geq 0, i = 1, \dots, n \end{aligned}$$

其中, C 叫作正则化参数, 控制对超出边界样本的惩罚程度。

在选择了适合的核函数 $k(x_i, x_j)$ 、适合的精度以及正则化参数后, 构造求解在约束条件下的凸二次规划优化问题。

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^n \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0, i = 1, \dots, n \\ & 0 \leq \alpha_i \leq C_i, i = 1, \dots, n \end{aligned}$$

当数据为非线性不可分时, 通过核函数实现从低维空间转换到高维空间实现线性可分, 而核函数的选取需满足 Mercer 定理: 任何半正定函数都可以作为核函数。

核函数决定了分类器模型性能的优劣。常见核函数如下:

1) 线性核函数: $K(x_i, x_j) = x_i \cdot x_j$ 。

2) 多项式(Poly)核函数: $K(x_i, x_j) = (\gamma \cdot x_i \cdot x_j + r)^d$, $\gamma > 0$, $d = 1, 2, \dots, n$ 。其中, d 为多项式阶数, r 为常数。 d 越大核矩阵元素值越大, 核函数的维度越大, 计算越复杂, 相对而言对数据的分类能力更好, 但是泛化性能较低, 容易过拟合。

3) 高斯径向基(RBF)核函数: $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, $\gamma > 0$ 。其中, γ 为高斯核函数的核宽度, 控制作用范围, 高斯核函数具有较好的局部性, 应用范围广, 对大样本和小样本都有较好的分类效果, 参数少。

4) Sigmoid 核函数: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$, $\gamma > 0$, $r < 0$ 。其中, γ 是尺度, r 是衰减参数。相当于两层感知机函数。

面对不同的研究对象以及核函数, 其构造的 SVM 模型的性能各不相同。

3. 模型的建立

基于网格搜索法 Grid Search CV, 建立 SVM 的乳腺癌诊断模型主要分为两个过程, 第一个过程运用

数据集训练得到最优参数对，第二个过程运用测试集检验模型的分类性能，实现乳腺癌良性恶性的诊断。构建参数优化的支持向量机诊断模型具体实现步骤如下：

数据预处理：输入原始数据，对数据进行预处理，标准化数据；将一部分数据作为训练集，另一部分作为测试集。

利用网格法寻找网格最优参数：确定参数 C 和核函数的范围，根据范围可以构建网格搜索平面，采用 K 折交叉验证寻找网格面的最优参数对，找到分类精度最高的参数对为最优参数对。如果有多组同样精度的参数对，取 C 值最小的那一组。利用训练集找到最优参数对后，构建 SVM 诊断模型，采用交叉验证法测试模型的性能。

支持向量机的不同模型参数分类性能可以通过 K 折交叉验证方法来检验。具体步骤：先将实验训练的数据集成 K 等分， K 为分组数，每次取其中的 $K - 1$ 份数据作为训练数据，剩下的一份数据作为测试数据进行验证。重复 k 次实验，直到每组都进行训练完成，根据 K 次计算后得到的平均值作为分类器性能的评价指标，最后选择其中分类效果最好的一组参数作为最优参数对。 K 折交叉验证可以有效地避免过拟合(over-fitting)以及欠拟合(under-fitting)情况的发生，因此被称为是对泛化误差的无偏估计，这种方法选择的参数能够使分类器具有良好的训练精度和泛化性能。交叉验证既可以用来验证支持向量机模型的性能，又可以通过多次验证寻找最好的支持向量机参数对，使其对乳腺癌数据的分类性能最好，有效防止过学习现象。

4. 网格搜索求解超参数

网格搜索是一种通过穷举给定的参数组合来寻找最优超参数的方法。在 SVM 超参数优化中，首先确定需要调优的参数范围。例如，选择不同的核函数类型(线性、多项式、高斯等)，以及正则化参数 C 和高斯核函数宽度参数 σ 的取值范围。然后对这些参数值进行排列组合，生成一个参数网格。对于参数网格中的每一组参数，网格搜索都会训练一个 SVM 模型，然后通过评估性能来选取最佳参数组合。我们选取交叉验证来评估该模型的性能。交叉验证[6]是一种将数据集划分为多个子集，然后轮流用其中一个子集作为测试集，其余子集作为训练集进行训练和测试的方法，它可以帮助我们更准确地评估模型的性能。

网格搜索会选择在交叉验证中表现最好的参数组合作为最优超参数。这种方法虽然计算量大，但能够系统地搜索参数空间，找到全局最优解或近似最优解。

5. 实证分析

5.1. 数据集及预处理

本文采用公开的“UCI Breast Cancer Dataset”[10]数据集进行实验。该数据集包含 569 个实例，涉及 30 个特征指标，包括细胞核半径、纹理、周长等统计量(均值、标准差、最大值)。分为良性和恶性肿瘤两类：良性(B/Malignant) 357 例，恶性(M/Benign) 212 例。

首先对数据预处理，方法如下。

1) 空缺值处理：对于缺失较多的数据集，常用的方法是利用均值处理、回归等方式对空缺位置进行缺值填充。这种方法的基本原理是结合上下文对缺失值进行填充，使填充值最可能逼近原值；此外，针对缺少较少的，一种较为简便的方式是直接删除缺失值。本研究中，乳腺癌数据样本比较全面，缺失值很少，所以直接删除了缺失值的整条记录。

2) 异常值处理：数据中的噪声是异常值常见的形式。这些噪声通常来源于数据采集、记录过程中随机产生的误差和偏差。本文使用命令删除极端异常值和明显非正常医疗数据值的孤立点。数据集成与变

换：如果数据来源于多个集合，则需要按照特定的方式将数据集集成合并，本文的数据已经集成，无需再做特别处理。数据变换是指将形态各异的原始数据变换成一致的格式，使其满足数据挖掘的要求。对于乳腺癌数据来说，每一项特征的检查数据有一定的冗余性或者规模较大，会影响分类器训练的能力，所以预测模型的输入指标需做相应的变换。这样大的属性值可能会使训练模型的输入维度变得非常高，导致计算量变大，训练所花时间变长，影响模型的分类准确性。为避免以上情况的发生，本文对原始数据进行了归一化处理。并且将指标是否是患者用 1 和 0 替换。

5.2. 超参数优化及交叉验证

本文使用网格搜索方法对 SVM 的超参数进行优化。网格搜索通过定义超参数的取值范围，遍历所有可能的超参数组合，并通过交叉验证评估每种组合的性能，最终选择性能最优的组合。

交叉验证方法具体实现：先将实验训练的数据集成 5 等分为 5 组，每次取其中的 4 份数据作为训练数据，剩下的一份数据作为测试数据进行验证。重复 5 次实验，直到每组都进行训练完成，根据 5 次计算后得到的平均值作为分类器性能的评价指标，最后选择其中分类效果最好的一组参数作为最优参数对。五折交叉验证可以有效地避免过拟合(over-fitting)以及欠拟合(under-fitting)情况的发生，选择的参数能够使分类器具有良好的训练精度和泛化性能，使其对乳腺癌数据的分类性能最好，有效防止过学习现象。

经过网格搜索，我们找到了最优的超参数组合：{'C': 2, 'kernel': 'RBF'}。

即 C 参数我们选择的是 2，核函数我们选择的是高斯径向基(RBF)核函数。

5.3. 模型评估

我们使用准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 分数(F1 Score)以及 ROC AUC 作为模型性能的评估指标[9]。最优模型性能指标见表 1。

ROC 曲线展示了模型在不同阈值下的真正例率(TPR)和假正例率(FPR)，如图 1 曲线下的面积(AUC)为 0.9965，表明模型具有出色的分类性能。

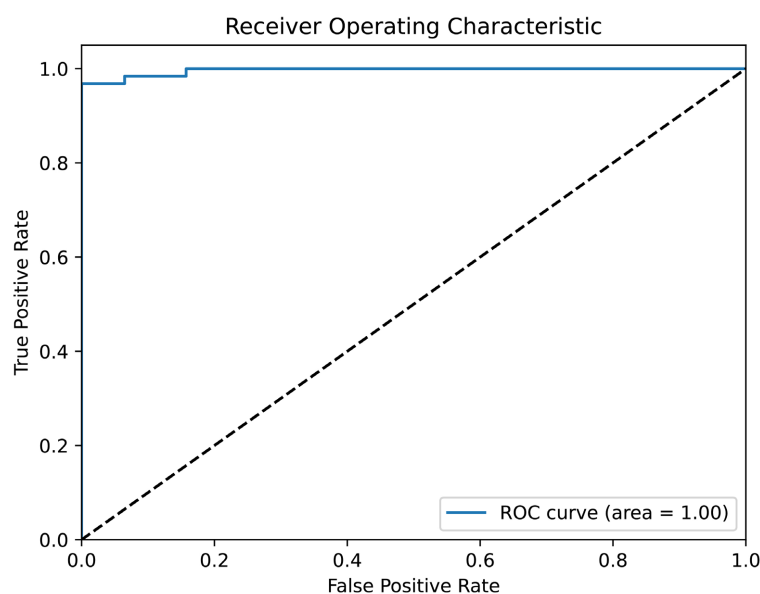


Figure 1. ROC curve
图 1. ROC 曲线

Table 1. Performance indicators of support vector machine model for breast cancer diagnosis**表 1.** 乳腺癌诊断支持向量机模型性能指标

评估指标	数值	说明
准确率(Accuracy)	0.9825 (98.25%)	模型整体预测正确率非常高, 表明在大多数情况下能够正确分类良性和恶性肿瘤。
精确率(Precision)	0.9839 (98.39%)	模型预测为恶性肿瘤的样本中, 98.39%确实是恶性肿瘤, 误判率很低。
召回率(Recall)	0.9683 (96.83%)	模型能够识别出 96.83%的真实恶性肿瘤, 漏诊率较低, 适合对恶性肿瘤的高敏感性要求。
F1 分数(F1 Score)	0.976 (97.6%)	精确率和召回率的调和平均值, 表明模型在精确性和全面性之间取得了很好的平衡。
ROC AUC	0.9965 (99.65%)	模型区分良性和恶性肿瘤的能力非常强, 接近完美分类器 (AUC = 1)。

6. 结论

支持向量机是建立在结构风险最小化原则基础上的一种新的机器学习方法, 主要借助于最优化方法解决数据挖掘中的分类问题, 其模型和算法的改进是本文研究的主要内容, 本文通过网络搜索方法成功找到了 SVM 模型的最优超参数组合, 并在乳腺癌数据集上取得了优异的性能。实验结果表明, 网络搜索是一种有效的超参数优化方法, 能够显著提升 SVM 模型的分类性能。

基金项目

河南省“专创融合”特色示范课程项目(2023), 河南科技大学教育教学改革研究与实践项目(2024BK089), 河南科技大学大学生创新创业计划项目(2024238)。

参考文献

- [1] 尚丹. 基于参数优化的 SVM 分类器在肺癌早期诊断中的应用[D]: [硕士学位论文]. 郑州: 郑州大学, 2014.
- [2] 常小改, 冯爱芬, 范静雅. 基于机器学习的抗乳腺癌候选药物的分类预测[J]. 四川文理学院学报, 2023, 33(2): 54-60.
- [3] 何梦. “丁香医生”微信公众号乳腺癌信息传播的框架研究[D]: [硕士学位论文]. 北京: 北京外国语大学, 2022.
- [4] 李华, 张伟, 王丽. 乳腺癌早期诊断中多模态影像特征与机器学习算法的结合应用[J]. 中国医学影像技术, 2023, 39(6): 854-859.
- [5] Vapnik, V., Golowich, S.E. and Smola, A. (1996) Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. *Proceedings of the 10th International Conference on Neural Information Processing Systems*, Denver, 3-5 December 1996, 281-287.
- [6] Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A. and Vapnik, V. (1996) Support Vector Regression Machines. *Proceedings of the 10th International Conference on Neural Information Processing Systems*, Denver, 3-5 December 1996, 779-784.
- [7] 霍罕妮. 支持向量机中参数选取的一个问题[D]: [硕士学位论文]. 大连: 大连理工大学, 2007.
- [8] 丁勇, 秦晓明, 何寒晖. 支持向量机的参数优化及其文本分类中的应用[J]. 计算机仿真, 2010, 27(11): 187-190.
- [9] 陈旭东, 许忠平, 童凯, 等. 基于网络搜索优化支持向量机多分类参数识别不同工艺酱酒的应用研究[J]. 中国酿造, 2024, 43(6): 213-217.
- [10] 阿里云. 威斯康星乳腺癌数据分析及自动诊断[EB/OL]. <https://tianchi.aliyun.com/dataset/106831>, 2021-07-21.