基于支持向量机算法的葡萄酒质量检测模型

张一明[1],魏霖静[2]

(1. 甘肃农业大学 理学院;2. 甘肃农业大学 信息科学技术学院,甘肃 兰州 730070)

摘 要: 随着我国经济的不断发展,葡萄酒作为一种悦人悦己的生活媒介已经登上大众餐桌。然而,葡萄酒的质量检测仍以品酒师品尝为主,已不能满足规模化、智能化的食品工业发展需求。为此,基于支持向量机算法对葡萄酒理化指标进行建模,利用R语言实现Box-plot法对异常值进行处理,同时对RBF核的支持向量机参数进行优化,最终得到一个精度达到96.46%的葡萄酒质量检测模型,为葡萄酒的质量控制提供了一条行之有效的途径。

关键词: 支持向量机;R语言;Box-plot法;葡萄酒质量检测

DOI: 10. 11907/rjdk. 231829 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP181 **文献标识码:** A **文章编号:** 1672-7800(2024)009-0137-06

Wine Quality Detection Model Based on Support Vector Machine Algorithm

ZHANG Yiming[1], WEI Linjing[2]

(1. *College of Science, Gansu Agricultural University*;
2. *College of Information Science and Technology, Gansu Agricultural University, Lanzhou* 730070, *China*)

Abstract: With the continuous development of our country's economy, wine, as a delightful medium of life, has entered the public dining table. However, the quality inspection of wine is still mainly based on the tasting of wine tasters, which can no longer meet the needs of the large-scale and intelligent development of the food industry. Therefore, based on the support vector machine algorithm, the physicochemical indicators of wine were modeled, and the Box plot method was implemented using R language to handle outliers. At the same time, the support vector machine parameters of the RBF kernel were optimized, resulting in a wine quality detection model with an accuracy of 96.46%. This provides an effective approach for wine quality control.

Key Words: support vector machine; R language; Box-plot method; wine quality testing

0 引言

随着我国经济的高速发展,人民的物质生活水平逐渐提高。葡萄酒以其诱人的风味、独特的口感受到了大众的青睐。然而,葡萄酒市场在我国发展较晚,相关质量检测手段尚未成熟,葡萄酒掺假造假现象较为严重,进口葡萄酒质量参差不齐,利用假冒伪劣产品冒充高档产品的现象更是屡见不鲜[1]。以往的葡萄酒质量检测手段多为品酒师对葡萄酒的风味、口感、香气等进行评判,但不同品酒师对同一种葡萄酒的质量评价可能存在一定差异,缺乏客观性。此外,培养一名合格的品酒师并不容易,国内专业的品酒师数量更是有限,已不能满足现今规模化、智能化的葡萄酒工业发展需求。因此,亟需建立一套行之有效的检测体系,实现对葡萄酒质量的快速、批量检测。

葡萄酒中的酯、醇、硫化物、酸、糖、矿物质和酚类化合物使其拥有了丰富的口感。各种成分之间存在着复杂的关系,又与葡萄酒的风味有着密切联系。机器学习是了解这些成分复杂关系及其与葡萄酒质量相关性的一个有效方法[2]。例如,Kruzlicova等[3]采用聚类分析和线性判别分

收稿日期: 2023-07-27
扫描二维码阅读全文:
基金项目: 科技部外专项目(G2022042005L);兰州市人才创新创业项目(2021-RC-47)

作者简介: 张一明(1999-),男,CCF会员,甘肃农业大学理学院硕士研究生,研究方向为大数据分析;魏霖静(1977-),女,博士,CCF会员,甘肃农业大学信息科学技术学院教授、博士生导师,研究方向为智能计算、农业信息化、生物信息学。本文通讯作者:魏霖静。

析法对葡萄酒质量进行检测；Geana 等[4]采用主成分分析法对葡萄酒质量进行检测；Roberto 等[5]采用软独立类比模型对葡萄酒质量进行检测；Tang 等[6]采用偏最小二乘法对葡萄酒质量进行检测；夏铭泽等[7]采用支持向量机（Support Vector Machine，SVM）对葡萄酒质量进行检测。然而，以上研究建立的模型准确率均有所欠缺，且未对葡萄酒原始数据进行分析处理。通常来说，数据决定了机器学习的上限，而算法只是在逼近这个极限。因此，本文首先使用异常值检验方法对数据进行处理，然后再建立检测模型。根据前人研究成果，在葡萄酒分类中SVM相较其他方法具有显著优势[8]。因此，本文采用SVM进行建模，选择合适的核函数，然后通过优化模型参数获得一个精度较高的葡萄酒质量检测模型。

1 SVM

SVM 由 Vapnik 等[9]于 1995 年提出，其建立在结构风险最小理论的基础上，因此能很好地扩大最优分类面与训练样本之间的距离，从而减少分类误差的上界。其核心思想为通过非线性映射将低维度难以分类的问题映射到高纬度空间中。如此以来，原本寻找最优分类直线的问题转化为在高维空间寻找最优分类面的问题。SVM是一种可用于回归与分类问题的有监督机器学习模型，可以输出较为准确的预测结果。其基本步骤为：

输入：训练数据集 $T=\{(X_1,Y_1),(X_1,Y_1),\cdots,(X_1,Y_1)\}$，其中 $x_i\in\mathcal{X}=R^n, y_i\in\mathcal{Y}=\{-1,+1\}, i=1,2,3,\cdots,N$；

输出：分类决策函数

选取合适的核函数 $K(x,z)$ 和参数，构造并求解最优化问题。用公式表示为：

$$\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i \tag{1}$$

$$s.t.\ \sum_{i=1}^{N}\alpha_i y_i=0 \tag{2}$$

$$0\leqslant\alpha\leqslant C,\ i=1,2,\cdots,N$$

由式(1)、式(2)求得最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^T$。

选择 α^* 的一个正分量 $0\leqslant\alpha_j^*\leqslant C$，计算：

$$b^*=y_i-\sum_{n=1}^{\infty}\alpha_i^* y_i K(x_i,x_j) \tag{3}$$

构造决策函数：

$$f(x)=sign(\sum_{n=1}^{\infty}\alpha_i^* y_i K(x,x_i)+b^*) \tag{4}$$

当 $K(x,z)$ 为正定核函数时，以上问题会转化成凸二次规划问题，解是存在的[10]。

SVM采用核函数的方法，数据由低维向高维空间映射时没有增加计算的复杂性，有效克服了维数灾难的问题。其还基于结构风险最小化原理有效解决了传统方法过拟合和陷入局部最小的问题[11]。然而，SVM的性能依赖于核函数的选择、数据质量和特征提取质量，其在应对大规模数据集时存在存储容量不足和计算速度不佳等问题，因此训练效率和泛化性能有待进一步提升[12]。

2 检测模型建立

2.1 数据预处理

选择 UCI 数据集中的 Wine Quality Data Set，共包含 4 898 个白葡萄酒样本。对数据进行汇总处理，展示其分布特征，具体如表 1 所示。可以看出，除了葡萄酒密度稳定在 1 g/mL 左右，其他指标均有较大极差，由此可见不同质量的葡萄酒各理化指标有所差异。

Table 1 Summary of wine data

表 1 葡萄酒数据汇总

指标	Min	1st Qu	Median	Mean	3rd Qu	Max
fixed.acidity	3.8	6.3	6.8	6.9	7.3	14.2
volatile.acidity	0.08	0.21	0.26	0.28	0.32	1.10
citric.acid	0	0.27	0.32	0.33	0.39	1.66
residual.sugar	1	2	5	6	10	66
chlorides	0.01	0.04	0.04	0.05	0.05	0.35
free.sulfur.dioxide	2.00	23.00	34.00	35.31	46.00	289.00
total.sulfur.dioxide	9	108	134	138	167	440
density	0.99	0.99	0.99	0.99	1.00	1.04
pH	2.7	3.1	3.2	3.2	3.3	3.8
sulphates	0.22	0.41	0.47	0.49	0.55	1.08
alcohol	8.0	9.5	10.4	10.5	11.4	14.2
quality	3.0	5.0	6.0	5.9	6.0	9.0

2.2 数据可视化

对数据中各个变量的相关系数进行研究。结果见图 1。可以看出，葡萄酒的密度、酒精浓度、残糖等变量与质量之间相关性较强，其余变量有一定的相关性，但并不强，还需要进一步分析[13]。其中，与葡萄酒质量关系最密切的

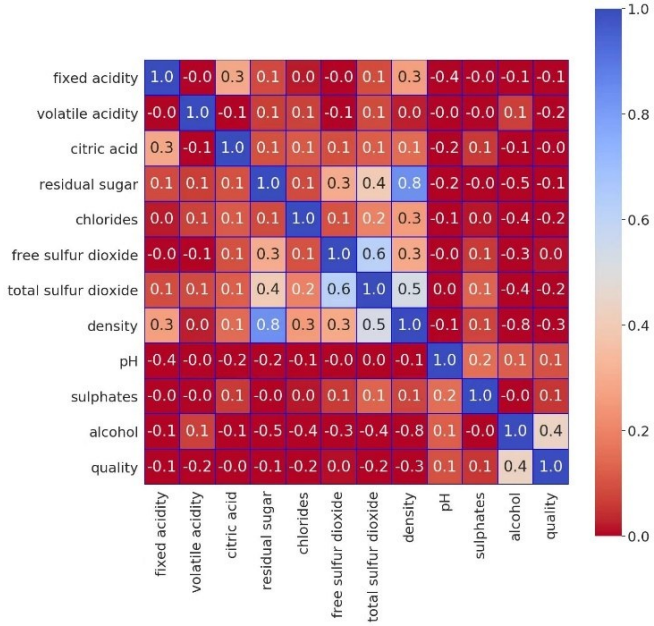

Fig. 1 Correlation coefficients of various variables in wine data

图 1 葡萄酒数据各变量相关系数

变量为酒精，通常酒精浓度越高，酒体越饱满，在舌尖上的回味也越清爽。

核密度图是一种观察连续型变量分布的有效方法。图 2 为酒精浓度与葡萄酒质量的核密度图。可以看出，不同质量葡萄酒酒精浓度分布有所不同，普通葡萄酒酒精浓度呈正偏态分布，优质葡萄酒酒精浓度呈负偏态分布，劣质葡萄酒酒精浓度近似地呈对称分布。

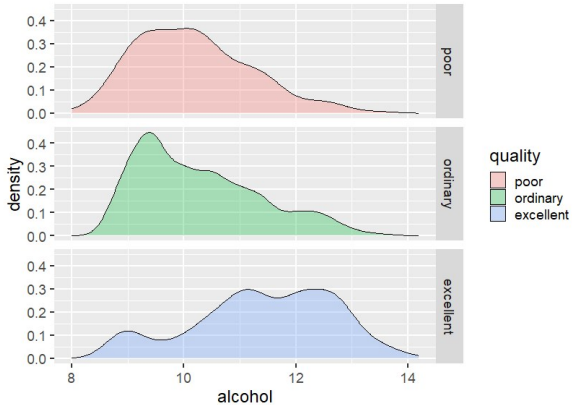

Fig. 2　Nuclear density diagram of alcohol concentration and wine quality

图 2　酒精浓度与葡萄酒质量的核密度图

图 3 为不同质量葡萄酒的酒精和逸出酸度浓度散点图。可以看出，大部分葡萄酒逸出酸度水平较低，因为葡萄酒中酒精所具有的甘甜味会平衡其中的酸味。了解理化指标对葡萄酒风味的影响有助于对相关数据进行预处理[14]。

2.3　异常值处理

葡萄酒数据中部分变量极差较大，如二氧化硫浓度范围为[9,440]。由于 SVM( ) 函数在生成模型时会默认对每个变量进行标准化处理，需检查数据是否存在异常值。异常值是与其他观察结果显著不同的点，可能是由于传感器故障、人工录入错误等导致。异常值会导致数据分布偏斜，拉高或拉低数据整体情况并严重影响数据集的均值和标准差。如果忽略这些异常值，在机器学习建模时可能会导致错误结论，因此需要识别异常值并对其进行预处理。

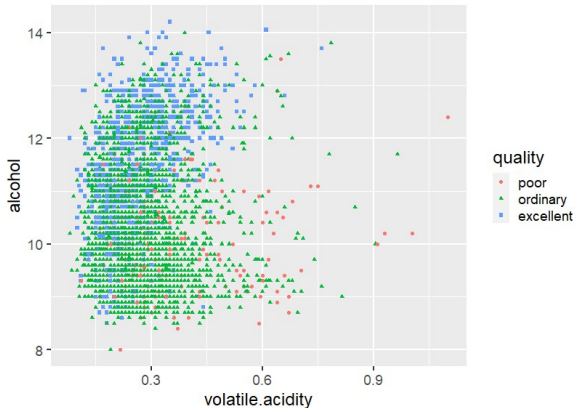

Fig. 3　concentration scatter diagram of alcohol and escaping acidity of different quality wines

图 3　不同质量葡萄酒酒精和逸出酸度浓度散点图

常用异常值检测方法包括 3sigma 准则法、Z-score 法、箱线图法（Box-plot）、K-最近邻分类算法等。本文通过绘制各变量的箱线图进行数据异常值处理。箱线图通过绘制连续型变量的最大值、上四分位数、中位数、下四分位数和最小值来描述其分布，能够显示出可能的离群点（即大于上四分位数 1.5 倍四分位差或小于下四分位数 1.5 倍四分位差的值）[15]。

图 4 为葡萄酒各指标箱线图。图中黑色的点即为离群值点。可以看出，固定酸度、逸出酸度、柠檬酸、氯化物、二氧化硫、硫酸盐等指标异常值较多，而残糖、密度、酒精浓度等指标分布较为均匀。

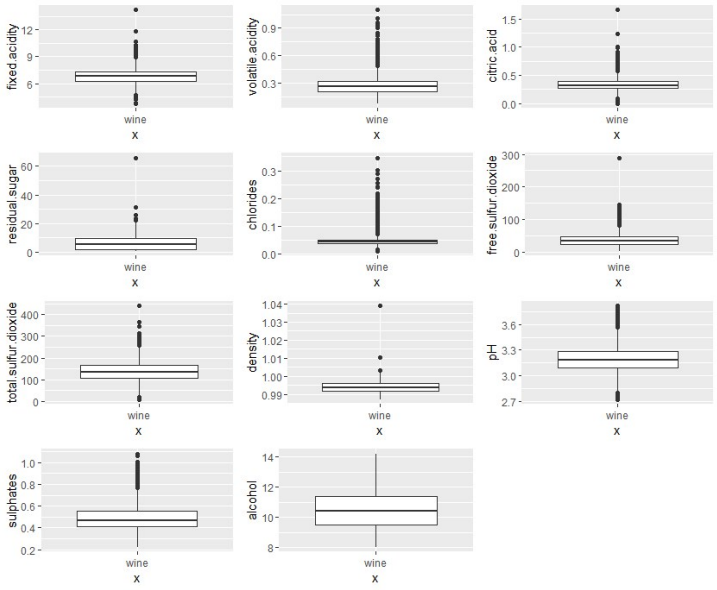

Fig. 4　Box chart of various indicators of wine

图 4　葡萄酒各指标箱线图

数据集中的葡萄酒分为劣质、普通、优质3个等级，数据中的极端值很可能是造成葡萄酒品质优劣的关键因素。直接对总体数据进行异常值检验并删除可能会造成分析结果不显著，甚至将预测结果引入歧途[16]。因此，本文首先将数据划分为3类，定位到异常值所在位置，采用四分位数进行替换，分别进行异常值处理，然后进行合并分析，如此以来可最大程度地减少对原始数据的影响。图5为葡萄酒质量分布直方图。由于模型需要，将原本的数值型变量转化为多变量因子，感官评分小于5的葡萄酒被标记为质量较低，共有183个数据；评分为5和6的葡萄酒被标记为质量中等，共有3 655个数据；评分大于6的葡萄酒被标记为质量较高，共有1 060个数据。

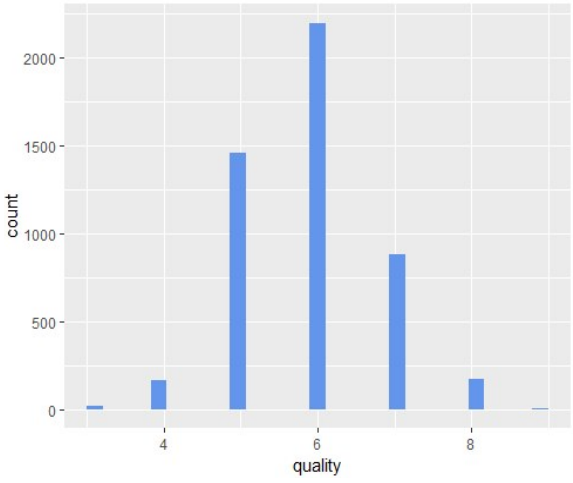

Fig. 5 Histogram of wine quality distribution

图5 葡萄酒质量分布直方图

根据分类准则将数据分为3类后分别绘制不同指标的箱线图。结果如图6所示。根据图6的可视化结果，首先找出所有离群点并结合葡萄酒的理化性质对数据集进行异常值处理，将其分别用上四分位数和下四分位数代替。处理完后的数据汇总如表2所示。处理后的数据对原始数据的分布状况改变幅度很小，采用该数据建立机器学习模型。

2.4 模型预测结果

基于Window.10系统平台，利用R语言中e1071包中的svm( )函数对模型进行拟合，使用高斯核为模型的核函数进行建模。表3为模型预测结果。

3 模型优化

3.1 核函数优化

核函数与参数选择直接决定SVM模型的运行效率[17]。在建立模型前并不知道什么样的核函数适合需要分析的数据，而核函数也只是隐式地定义了特征空间。

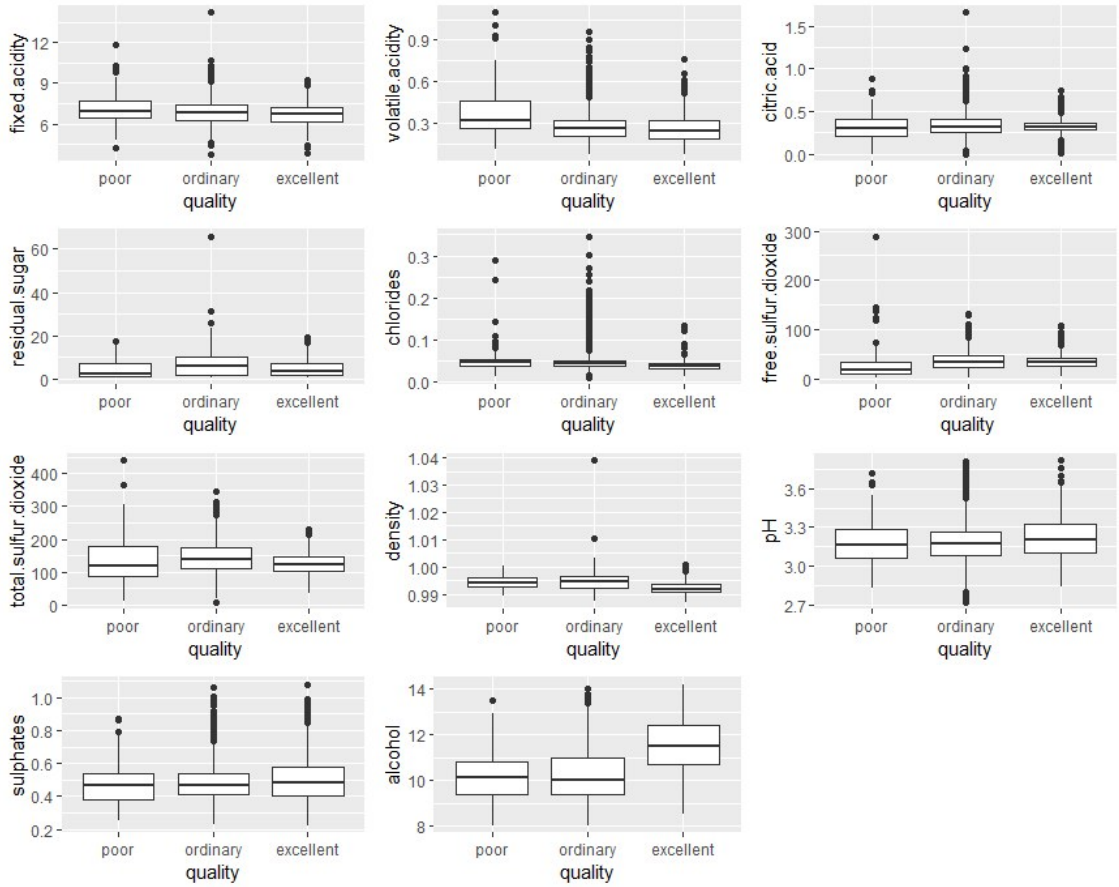

Fig. 6 Box chart of various indicators of classified data

图6 分类数据各指标箱线图

若核函数选择得不合适,样本将会被映射到一个不合适的特征空间,导致模型预测精度不高[18]。常用核函数见表4。

Table 2　Data summary after abnormal value processing

表 2　异常值处理后数据汇总

指标	Min	1st Qu	Median	Mean	3rd Qu	Max
fixed.acidity	3.800	6.300	6.800	6.851	7.300	14.200
volatile.acidity	0.080 0	0.210 0	0.260 0	0.270 8	0.320 0	1.100 0
citric.acid	0	0.27	0.32	0.33	0.38	1.66
residual.sugar	0.600	1.700	5.200	6.372	9.800	65.800
chlorides	0.009 00	0.036 00	0.043 00	0.045 66	0.050 00	0.346 00
free.sulfur.dioxide	2.00	23.00	34.00	35.15	46.00	131.00
total.sulfur.dioxide	10	108	134	138.1	167	344
density	0.987 1	0.991 7	0.993 6	0.993 9	0.995 9	1.003 0
pH	2.7	3.0	3.1	3.1	3.2	3.8
sulphates	0.220 0	0.410 0	0.470 0	0.489 7	0.550 0	1.080 0
alcohol	8.0	9.5	10.4	10.5	11.4	14.2

Table 3　Forecast results

表 3　预测结果

模型	预测精度			总体预测精度
	劣质	普通	优质	
SVM	0.041 096	0.95 0139	0.375 796	0.782 312 9

Table 4　Commonly used kernel functions

表 4　常用核函数

名称	表达式	参数
线性核	$k\left(x_i, x_j\right) = {x_i}^T x_i$	/
多项式核	$k\left(x_i, x_j\right) = \left({x_i}^T x_i\right)^d$	$d \geqslant 1$ 为多项式的次数
高斯核	$k\left(x_i, x_j\right) = exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)$	$\sigma > 0$ 为高斯核的带宽(width)
Sigmoid 核	$k\left(x_i, x_j\right) = tanh\left(\beta {x_i}^T x_i - \theta\right)$	tanh 为双曲正切函数,$\beta > 0, \theta > 0$

使用不同核函数作为模型参数并进行预测精度比较,结果如表5所示。可以看出,高斯核作为核函数时模型拟合效果最优。

Table 5　Prediction results of different kernel functions

表 5　不同核函数预测结果

核函数	预测精度			总体预测精度
	劣质	普通	优质	
高斯核	0.041 096	0.950 139	0.375 796	0.782 312 9
线性核	0	100	0	0.736 734 7
多项式核	0.082 192	0.963 989	0.792 994	0.737 415
Sigmoid 核	0.027 397	0.817 175	0.592 357	0.689 795 9

3.2　模型参数优化

SVM算法中有很多控制支持向量的参数,通过交叉验证等方法可以选择合适的参数值以提高算法性能[19]。当使用高斯核的SVM拟合模型时,核函数中控制分割超平面形状的参数 gamma 和控制犯错成本的参数 cost 可能会影响最终结果。当成本参数较大时,模型对误差的惩罚也更大,从而生成一个更复杂的分类边界,对应训练集中的误差会更小,但会增加模型过拟合的可能,即训练出的模型不能很好地适用于新样本,模型泛化能力较弱;当成本参数较小时,生成的分类边界会更平滑,但模型欠拟合的可能性会增大[20]。

本文使用不同的 gamma 和 cost 参数拟合高斯核的SVM模型,共尝试6个 gamma 值(0.000 04~10)和20个 cost 值(0.000 000 000 1~100 000 000 000),拟合了120个模型,得出10折交叉验证误差最小的模型所对应的参数为 gamma=1 ,cost=100 。基于该组参数再次对训练数据进行拟合,并进行预测。模型混淆矩阵和优化后的预测结果分别如表6和表7所示。可以看出,经过数据预处理以及参数优化后的葡萄酒质量预测模型精度达到96.46%。

Table 6　Confusion matrix of the model

表 6　模型混淆矩阵

真实值	预测值		
	劣质	普通	优质
劣质	68	4	1
普通	2	1 078	3
优质	0	42	272

Table 7　Prediction results of optimized model

表 7　优化后模型预测结果

模型	预测精度			总体预测精度
	劣质	普通	优质	
SVM	0.931 506 8	0.995 383 2	0.866 242	0.964 625 9

4　结语

为实现对葡萄酒质量的快速、批量检测,本文利用描述化统计葡萄酒各指标与质量的关系以及数据的分布状况,采用Box-plot法对分类后的数据进行异常值处理,同时利用交叉验证方法选出最优核函数和参数建立SVM模型,得到一个精度较高的葡萄酒质量检测模型。然而从结果上看模型成本参数较高,可能出现过拟合现象,需要更多数据进行实验以检验模型是否能很好地应对样本以外的数据。

参考文献:

[1] MU W S, WU X Q, QI J F, et al. Analysis of the development situation and market demand characteristics of China's wine industry [J]. Sino-Overseas Grapevine & Wine, 2022(4): 81-89.
穆维松,吴晓倩,齐建芳,等. 中国葡萄酒产业发展形势及市场需求特征分析[J]. 中外葡萄与葡萄酒,2022(4):81-89.

[2] YU J. Research on quality identification methods of wine and liquor[D]. Beijing: China Agricultural University, 2018.
于静. 葡萄酒和白酒质量识别方法的研究[D]. 北京:中国农业大学,2018.

[3] KRUZLICOVA D, FIKET Ž, KNIEWALD G. Classification of Croatian wine varieties using multivariate analysis of data obtained by high resolution ICP-MS analysis [J]. Food Research International, 2013, 54(1): 621-626.

[4] GEANA E I, MARINESCU A, IORDACHE A M, et al. Differentiation of

Romanian wines on geographical origin and wine variety by elemental composition and phenolic components[J]. Food Analytical Methods, 2014, 7(10):2064-2074.

[5] ROBERTO R, U W C, PAOLO B, et al. Discrimination between Shiraz wines from different Australian regions: the role of spectroscopy and chemometrics[J]. Journal of Agricultural and Food Chemistry, 2011, 59(18): 10356-10360.

[6] TANG K, MA L, HAN Y H, et al. Comparison and chemometric analysis of the phenolic compounds and organic acids composition of chinese wines [J]. Journal of Food Science, 2015, 80(1): C20-8.

[7] XIA M Z, SHI C P, LIU Z Y. Wine quality prediction based on support vector machine[J]. Manufacturing Automation, 2020, 42(5): 57-60.
夏铭泽, 石春鹏, 刘征宇. 基于支持向量机的葡萄酒质量预测[J]. 制造业自动化, 2020, 42(5): 57-60.

[8] CORTEZ P, CERDEIRA A, ALMEIDA F, et al. Modeling wine preferences by data mining from physicochemical properties[J]. Decision Support Systems, 2009, 47(4): 547-553.

[9] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[10] LI H. Statistical learning methods [M]. Beijing: Tsinghua University Press, 2019: 111-152.
李航. 统计学习方法[M]. 北京: 清华大学出版社, 2019: 111-152.

[11] WANG X, DONG Y Q, YU Q, et al. Review of structural support vector machines [J]. Computer Engineering and Applications, 2020, 56(17): 24-32.
王霞, 董永权, 于巧, 等. 结构化支持向量机研究综述[J]. 计算机工程与应用, 2020, 56(17): 24-32.

[12] WANG H Y, LI J H, YANG F L. Overview of support vector machine analysis and algorithm [J]. Application Research of Computers, 2014, 31(5): 1281-1286.
汪海燕, 黎建辉, 杨风雷. 支持向量机理论及算法研究综述[J]. 计算机应用研究, 2014, 31(5): 1281-1286.

[13] ZHANG G Y, TUO X G, ZENG X L, et al. Visual analysis of wine sensory evaluation based on high-dimensional multivariate data [J]. Science and Technology of Food Industry, 2021, 42(9): 78-84.
张贵宇, 庹先国, 曾祥林, 等. 基于高维多元数据的酒体感官评价可视分析[J]. 食品工业科技, 2021, 42(9): 78-84.

[14] CAO W Y, LU W P, SHU N, et al. Research progress on wine flavor compounds and their influencing factors [J]. China Brewing, 2022, 41(5): 1-7.
曹炜玉, 路文鹏, 舒楠, 等. 葡萄酒风味物质及其影响因素研究进展[J]. 中国酿造, 2022, 41(5): 1-7.

[15] GU G Q, LI X H. Exponential weighted smoothing prediction model based on box graph anomaly detection [J]. Computer and Modernization, 2021(1): 28-33.
顾国庆, 李晓辉. 基于箱线图异常检测的指数加权平滑预测模型[J]. 计算机与现代化, 2021(1): 28-33.

[16] CHE J X. Variable selection and prediction methods for complex data [D]. Xi'an: Xidian University, 2019.
车金星. 复杂数据的变量选择与预测方法研究[D]. 西安: 西安电子科技大学, 2019.

[17] SONG Y S, LIU G Y, ZHU L, et al. Application of improved GWO algorithm in SVM parameter optimization [J]. Transducer and Microsystem Technologies, 2022, 41(9): 151-155.
宋玉生, 刘光宇, 朱凌, 等. 改进的灰狼优化算法在SVM参数优化中的应用[J]. 传感器与微系统, 2022, 41(9): 151-155.

[18] ZHOU Z H. Machine learning [M]. Beijing: Tsinghua University Press, 2016: 122-140.
周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 122-140.

[19] ZHAI Y Y, ZUO L, ZHANG E D. RBF neural network structure design algorithm based on parameter optimization[J]. Journal of Northeast Normal University(Natural Science Edition), 2020, 41(2): 176-181, 187.
翟莹莹, 左丽, 张恩德. 基于参数优化的RBF神经网络结构设计算法[J]. 东北大学学报(自然科学版), 2020, 41(2): 176-181, 187.

[20] KABACOFFROBERT I. R in action[M]. Westampton: Manning Publications, 2015: 370-376.

(责任编辑: 尹晨茹)