文章编号： 1000-5013(2013)01-0030-06

一种基于最小二乘支持向量机的葡萄酒品质评判模型

吴瑞红，王亚丽，张环冲，王鲜芳

（河南师范大学 计算机与信息技术学院，河南 新乡 453003）

摘要： 对源自 UCI 数据库的葡萄酒数据进行预处理，选取径向基函数作为最小二乘支持向量机的核函数；然后，根据“一对一”算法设计出最小二乘支持向量机多元分类器，并应用交叉验证算法对参数寻优，建立葡萄酒质量评判模型. 同时，用 BP 神经网络、标准支持向量机分类器对葡萄酒进行训练. 对比实验结果表明：最小二乘支持向量机比 BP 神经网络、标准支持向量机的平均分类准确率高，最高分类准确率为 100%.

关键词： 最小二乘支持向量机；葡萄酒；多元分类器；交叉验证；品质评判

中图分类号： TS 262.6；TS 207.3；TP 183 **文献标志码：** A

葡萄酒具有特殊的营养价值和医疗保健作用，是食品工业的重要组成部分. 如何对葡萄酒科学分类，提高葡萄酒的质量，对促进行业健康发展具有重要的实际意义. 人们一直靠感官来判定葡萄酒质量的好坏，而感官鉴定受到多种因素的影响，其准确性难以得到保证. 国内外对葡萄酒质量评判的研究主要有遗传神经网络[1]、模糊神经网络[2]、数据挖掘[3]和贝叶斯[4]等算法. 本文针对通过感官鉴别葡萄酒质量的准确性难以保证的问题，建立一种基于最小二乘支持向量机学习算法的酒质量评判模型.

1 最小二乘支持向量多元分类器

1.1 最小二乘支持向量机原理

Suykens 等[5]提出的最小二乘支持向量机(LS-SVM)是标准支持向量机[6-8]的一种改进. 它将标准支持向量机中的不等式约束改为等式约束，且将误差平方和损失函数作为训练集的经验损失. 这样就把解二次规划问题转化为求解线性方程组问题，提高求解问题的速度和收敛精度.

设定训练集$\{x_i,y_i\}_{i=1}^N$，则最小二乘支持向量机优化问题表示为

$$\left.\begin{aligned}&\min_{\boldsymbol{w},b,\boldsymbol{\xi}} J(\boldsymbol{w},b,\boldsymbol{\xi})=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\gamma\frac{1}{2}\sum_{i=1}^{N}\xi_i^2, \qquad \gamma>0,\\&\text{s. t.}\quad y_i[\boldsymbol{w}^{\mathrm{T}}\cdot\varphi(x_i)+b]=-1-\xi_i, \qquad i=1,\cdots,N.\end{aligned}\right\} \tag{1}$$

式(1)中：$\xi_i\geqslant 0$ 是允许错分的松弛变量；γ 为错误惩罚分量.

通过式(1)的对偶形式可以求它的最优解，而对偶形式可以根据目标函数和约束条件建立拉格朗日函数. 即

$$L(\boldsymbol{w},b,\boldsymbol{\xi},\alpha)=J-\sum_{i=1}^{N}\alpha_i\{y_i[\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b]-1+\xi_i\}. \tag{2}$$

式(2)中：α_i 是 Lagrange 乘子.

按照 KKT(Karush-Kuhn-Tucker)条件[9]对式(2)进行优化，分别对 $\boldsymbol{w},b,\xi_i,\alpha_i$ 求导，即有

收稿日期： 2012-06-15

通信作者： 王鲜芳(1969-)，女，教授，主要从事复杂过程建模与优化控制的研究. E-mail：xfwang11@yahoo.com.cn.

基金项目： 国家自然科学基金资助项目(61173071)；河南省科技攻关计划项目(112102210412)；河南省基础与前沿技术研究计划项目(112300410254)；河南省高校创新人才支持计划项目(2012HASTIT011)

$$\left.\begin{aligned}
&\frac{\partial L}{\partial w}=0\rightarrow \boldsymbol{w}=\sum_{i=1}^{N}\alpha_i y_i\varphi(x_i),\\
&\frac{\partial L}{\partial b}=0\rightarrow \sum_{i=1}^{N}\alpha_i y_i=0,\\
&\frac{\partial L}{\partial \xi_i}=0\rightarrow \alpha_i=\gamma\xi_i,\\
&\frac{\partial L}{\partial \alpha_i}=0\rightarrow y_i[\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b]-1+\xi_i=0,\qquad i=1,\cdots,N.
\end{aligned}\right\}\tag{3}$$

式(3)能被直接表示为

$$\begin{bmatrix}\boldsymbol{I}&\boldsymbol{0}&\boldsymbol{0}&-\boldsymbol{Z}^{\mathrm{T}}\\ \boldsymbol{0}&\boldsymbol{0}&\boldsymbol{0}&-\boldsymbol{Y}^{\mathrm{T}}\\ \boldsymbol{0}&\boldsymbol{0}&\gamma\boldsymbol{I}&-\boldsymbol{I}\\ \boldsymbol{Z}&\boldsymbol{Y}&\boldsymbol{I}&\boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{w}\\ b\\ \boldsymbol{\xi}\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\\ \boldsymbol{0}\\ \boldsymbol{1}\end{bmatrix}.\tag{4}$$

式(4)中：$\boldsymbol{Y}=[y_1,\cdots,y_N]$，$\boldsymbol{\xi}=[\xi_1,\cdots,\xi_N]$，$\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_N]$，$\boldsymbol{Z}=[\varphi(x_1)^{\mathrm{T}}y_1,\cdots,\varphi(x_N)^{\mathrm{T}}y_N]$，$\boldsymbol{1}=[1,\cdots,1]$，$\boldsymbol{I}$ 是单位矩阵. $\boldsymbol{w}$ 和 $\boldsymbol{\xi}$ 的值可以从式(3)得出，那么式(4)可以表示为

$$\begin{bmatrix}\boldsymbol{0}&\boldsymbol{Y}^{\mathrm{T}}\\ \boldsymbol{Y}&\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}+\gamma^{-1}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{1}\end{bmatrix}.\tag{5}$$

根据文献[9]中的 Mercer 定理，可以实用核函数 $K(\cdot,\cdot)$，即

$$\boldsymbol{\Omega}_{i,j}=y_iy_j\varphi(x_i)^{\mathrm{T}}\varphi(x_j)=y_iy_jK(x_i,x_j).\tag{6}$$

常用的核函数有多项式核、高斯核、感知器核和 B 样条核[10]. 文中建模选用的是高斯核函数，即径向基核函数，其形式为

$$K(x,x_i)=\exp(-\parallel x-x_i\parallel/2\sigma^2).\tag{7}$$

式(7)中：σ 为核宽度参数.

通过式(5)，(6)就可以得到分类器，避免了标准的 SVM 中相对复杂的二次规划问题. 所求出的 $\boldsymbol{\alpha}$，b 可以得到训练集的分类决策函数，其表达式为

$$y(x)=\mathrm{sgn}(\sum_{i=1}^{N}\alpha_iy_iK(x,x_i)+b)=\mathrm{sgn}(\sum_{i=1}^{N}\alpha_iy_i\exp(-\parallel x-x_i\parallel/2\sigma^2)+b).\tag{8}$$

1.2　多元分类最小二乘支持向量机的构建

多分类支持向量机常用的方法有 4 种：一对多、一对一、导向无环图、二叉树. “一对一”的分类方法虽计算复杂，但精度高，适合类别数目较少的情形[11]. 因此，基于文献[11,12-16]的思想，设计了基于最小二乘支持向量机(LS-SVM)的多元分类器. 文中采用“一对一”(OAO)的方法进行多元分类，该方法是基于两类问题的分类方法，但两类问题是从原来的多类问题中抽取的. 应用该方法需要构建 $k(k-1)/2$ 个二元分类器，也就是需要构建 $k(k-1)/2$ 个决策函数，k 为所需分类问题的种类数.

LS-SVM$_{st}$ 是训练类标签为 c_s 和类标签为 c_t 的分类器，测试集的类标签是由所有决策值和投票策略决定的. 如果对类 c_s 和 c_t 分类的决策函数 $y^{s-t}(x)$ 的决策值显示样本 x 的预测类标签为 c_s，那么类 c_s 的投票为 $\arg\max\limits_{s=1,\cdots,k}\mathrm{vote}_x(s)$. 其中：$v_x(s,t)=\begin{cases}1,\text{if } y^{s-t}(x)\text{ 得到 }x\text{ 是类 }c_s,\\ 0,\text{其他；}\end{cases}$ $\mathrm{vote}_x(s)=\sum\limits_{t=1,t!=s}^{k}v_x(s,t)$；

$y^{s-t}(x)=\mathrm{sgn}\sum\limits_{i=1}^{N}\alpha_i^{s-t}y_i\exp(-\parallel x-x_i\parallel/2\sigma^2)+b^{s-t}),s,t=1,\cdots,k;t!=s.$

2　仿真实验分析

2.1　葡萄酒品质评判模型

为了验证构建的多元分类器的预测准确率，对 UCI 机器学习数据库(http://archive.ics.uci.edu/ml/datasets/Wine)中的葡萄酒数据集进行仿真实验. 该数据集共包括 178 个样本，分成 3 类，第 1 类的样本有 59 个，第 2 类的样本有 71 个，第 3 类的样本有 48 个. 每个样本含有 13 个特征分量，分别是 Al-

cohol, Malic acid, Ash, AlKalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline, 这些特征分量作为分类器的输入数据 X. 图 1 为葡萄酒数据可视化图.

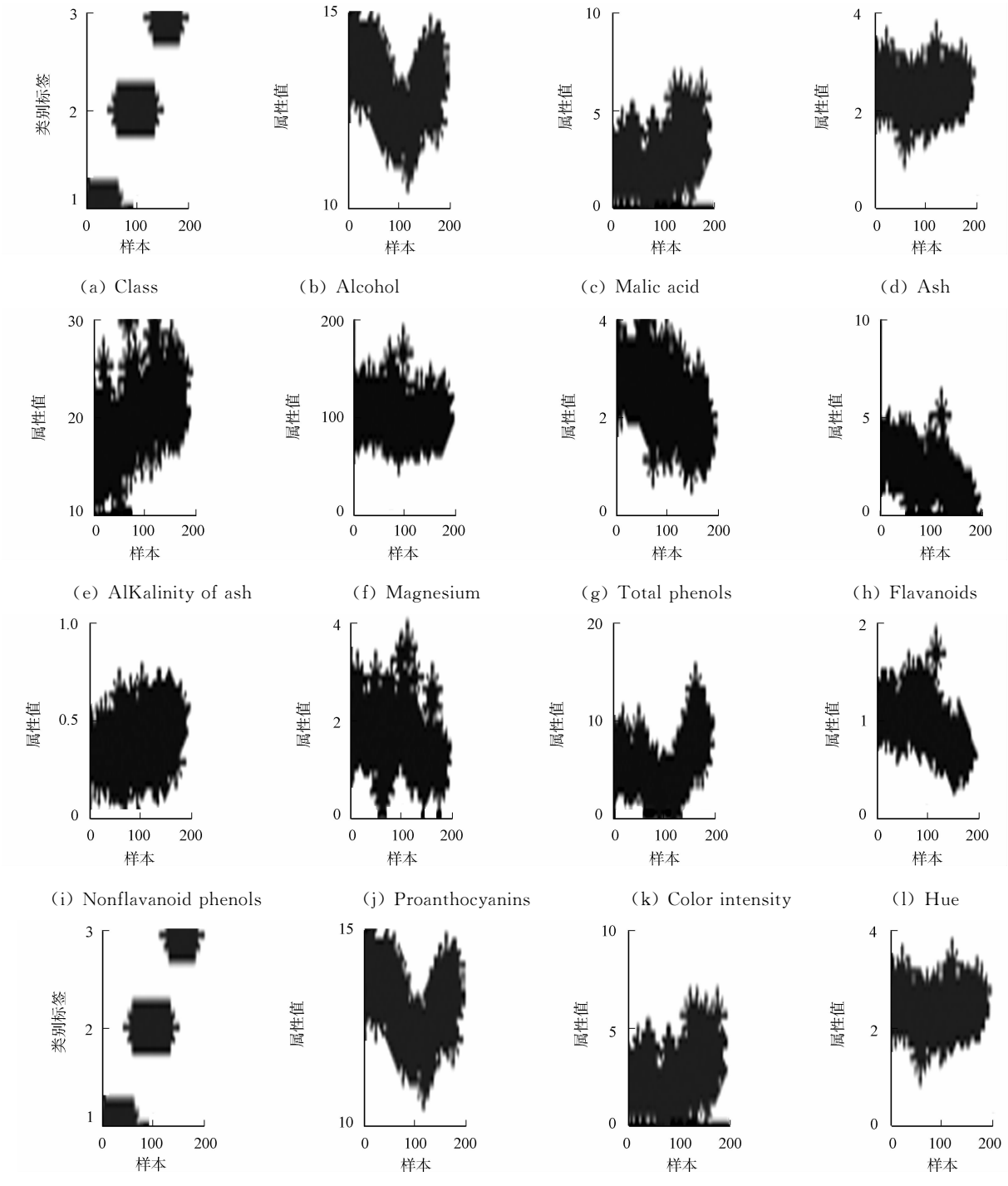

图 1 葡萄酒数据的分维可视化图

Fig. 1 Fractal dimension visual figure of wine data

所研究的葡萄酒品质有 3 类,基本的 LS-SVM 是基于两类的,必须应用多元分类 LS-SVM,即采用 OAO LS-SVM 方法对葡萄酒进行分类. 在 OAO LS-SVM 分类方案中,要想对葡萄酒进行分类,需要构建 3 个二元分类器. 二元分类决策函数表达式为式(10),再结合式(9),可以构建出基于 LS-SVM 的葡萄酒品质多元分类器.

为了提高分类准确率,对数据集进行归一化预处理,采用的归一化映射为

$$f: x \to y = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{11}$$

式(11)中：$x, y \in \mathbf{R}^n$；$x_{\min} = \min(x)$；$x_{\max} = \max(x)$归一化的效果是原始数据被规整到[0,1]范围内，即$y_i = [0,1], i = 1,2,\cdots,n$.

2.2　分类结果

仿真运用的平台为 Windows 7,4 G 内存，软件为 MATLAB(R2010b). 为了验证模型的健壮性和适应性，对预处理过的 178 个葡萄酒样本采取随机采样的方法选取训练集，每次从全体数据中随机的选择 1/2 作为训练集，其余的数据作为测试集，即实验时训练集样本为 89，测试集样本为 89.

利用式(9)，(10)构建的多元分类器，以及文献[7]中的交叉验证方法得到相关参数. LS-SVM 对训练集训练时，多维无约束非线性优化问题采用 Nelder-Mead 单纯形算法[18]，即为 Simplex. 最好分类结果如图 2 所示. LS-SVM 对 4 次随机采样的训练结果，如表 1 所示. 表 1 中：N 为运行次数；γ 为惩罚系数；σ 为核参数；t 为运行时间；φ 为准确度 .

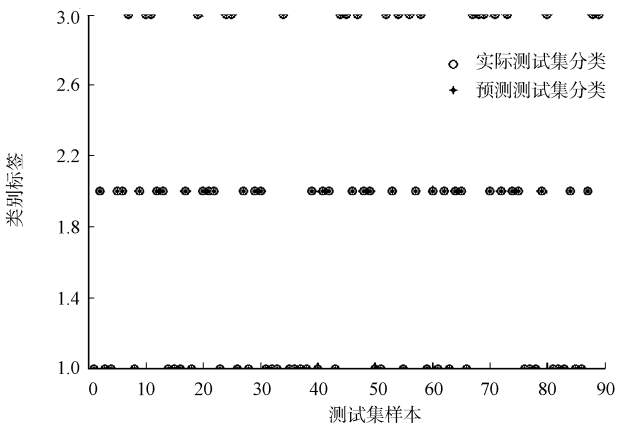

图 2　LS-SVM 多元分类器分类结果可视化图

Fig. 2　Visual figure of classified results with LS-SVM multi-classifier

为了证明所构筑分类器的分类性能，在同样的输入数据和测试数据的条件下，分别又构筑了 SVM 和 BP 神经网络多元分类器[17]. SVM 同样采用[17]中的交叉验证得到相关参数，取值和分类结果如表 2 所示.

表 1　LS-SVM 多元分类器的分类结果

Tab. 1　Classified results with LS-SVM multi-classifier

N	γ	σ	t/s	φ/%
1	7.914 100	1.795 9	1.060 807	97.75
2	4.000 600	2.562 0	1.092 007	96.63
3	3.272 518	382.335 6	1.528 810	98.88
4	0.263 061	26.356 4	1.419 609	100.00

表 2　SVM 多元分类器的分类结果

Tab. 2　Classified results with SVM multi-classifier

N	γ	σ	t/s	φ/%
1	5.656 850	0.062 500	2.964 019	98.876 4
2	0.250 000	0.176 777	3.744 024	97.752 8
3	5.656 850	0.062 500	3.120 020	98.876 4
4	0.707 107	0.500 000	3.120 020	97.752 8

在 BP 神经网络多元分类器中，输入层取 13 个节点，隐含层取 8 个节点，输出层取 3 个节点. 其最佳隐含层节点数选择参考如下公式，即

$$l < \sqrt{(m+n)} + a. \tag{12}$$

式(12)中：n 为输入层节点数；l 为隐含层节点数；m 为输入出层节点数；a 为 0～10 之间的常数. 学习速率为 0.1，惯性系数为 0.01. 表 3 为 BP 分类器的分类结果. BP，SVM，LS-SVM 分类方法的比较结

果,如表4所示.表4中的平均运行时间(t_{av})、平均分类准确率(φ_{av})分别从表1~3中计算得出.

表3 BP多元分类器分类结果

Tab.3 Classified results with BP multi-classifier

N	φ/%				t/s
	第一类	第二类	第三类	整体	
1	0.914 3	0.944 4	1.000 0	94.380 0	1.138 807
2	1.000 0	0.885 7	1.000 0	95.510 0	1.248 008
3	1.000 0	0.966 7	0.965 5	97.750 0	1.154 407
4	1.000 0	0.931 0	1.000 0	97.750 0	1.248 008

从表4可见:最小二乘支持向量机其健壮性和适应性最好,其平均分类准确率最高为98.315%,且收敛速度最快.标准支持向量机的分类准确率次之为98.3146%,收敛速度较慢.这是因为最小二乘支持向量机把解二次规划问题转化为求解线性方程组问题,提高求解问题的速度和收敛精度,且所需计算资源少.LS-SVM最高分类准确率达到100%,表明当训练参数和训练样本选择最佳情况时,LS-SVM分类准确率也将达到最佳.与BP神经网络分类方法所得结果比较发现:SVM和LS-SVM分类器均具有较高的准确率.这说明支持向量机能较好地解决小样本、非线性等实际问题,具有很强的泛化能力.

表4 不同分类方法的结果对比

Tab.4 Comparison of different classified methods′ results

分类方法	t_{av}	φ_{av}/%	φ_{max}/%
BP	1.197 308	96.347 5	97.750 0
SVM	3.237 021	98.314 6	98.876 4
LS-SVM	1.275 308	98.315 0	100.000 0

模型的参数最终取使得训练集验证分类准确率最高的那组γ和σ做为最佳参数,如有多组γ和σ对应于最高的验证分类准确率,则取能够达到最高分类准确率中参数γ最小的那组γ和σ做为最佳的参数,如果对应最小的γ有多组σ,就选取搜索到的第1组γ和σ做为最佳参数.

3 结论

针对通过感官鉴别葡萄酒质量的准确性难以保证的问题,建立了一种基于最小二乘支持向量机学习算法的葡萄酒质量评判模型.同时,用BP神经网络、标准支持向量机分类器对葡萄酒进行训练.

从BP神经网络、标准支持向量机和最小二乘支持向量机3种分类准确率及运行时间对比,最小二乘支持向量机平均分类准确率最高,所能达到的最高分类准确率为100%.由此可见,最小二乘支持向量机在模式分类问题上能提供好的泛化性能,求解速度快,求解所需的计算资源较少.

运用交叉验证方法选取惩罚系数和核参数来训练分类器是有效的."一对一"算法为多元分类器实现提供了很好的方法,虽其计算复杂,但精度高,适合类别数目较少的情形.最小二乘支持向量机能较好地解决小样本、非线性等实际问题,在葡萄酒品质评判中具有很大应用潜力.

参考文献:

[1] 殷勇,邱明,刘云宏,等.基于遗传神经网络的酒类鉴别技术[J].农业机械学报,2003,34(6):104-106.

[2] RAPTIS C G,SIETTOS C I,KIRANOUDIS C T,et al. Classification of aged wine distillates using fuzzy and neural network systems[J]. Journal of Food Engineering,2000,46(4):267-275.

[3] CORTEZ P,CERDEIRA A,ALMEIDA F,et al. Modeling wine preferences by data mining from physicochemical properties[J]. Decision Support Systems,2009,47(4):547-557.

[4] BELTRÁN N H,DUARTE-MERMOUD M A,et al. Feature extraction and classiication of Chilean wines[J]. Journal of Food Engineering,2006,75(1):1-10.

[5] SUYKENS J K,VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letter,1999,9(3):293-300.

[6] VAPNIK V N. The nature of statistical learning theory[M]. New York:Pringer-Verlag,1995.

[7] VAPNIK V N. Statistical learning theory[M]. New York:Pringer-Verlag,1998.
[8] HE Xi-sheng,ZHE Wang,Cheng Jin,et al. A simplified multi-class support vector machine with reduced dual optimization[J]. Pattern Recognition Letters,2012,33(1):71-82.
[9] CRISTIANINI N,SHAWE-TAYLOR J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge:The Press Syndicate of Cambridge University,2000.
[10] 潘立登,李大字,马俊英. 软测量技术原理与应用[M]. 北京:中国电力出版社,2009.
[11] WANG Tai-yue,CHIANG Huei-min. One-against-one fuzzy support vector machine classifier: An approach to text categorization[J]. Expert Systems with Applications,2009,36(6):10030-10034.
[12] 朱家元,郭基联,张恒喜,等. 多元分类 LS-SVM 设计与装备保障性评估[J]. 装备指挥技术学院学报,2003,14(3):12-15.
[13] VEENMAN C J,BOLCK A. A sparse nearest mean classifier for high dimensional multi-class problems[J]. Pattern Recognition Letters,2011,32(6):854-859.
[14] 陈志刚,连香姣,于会媛,等. 多元支持向量机在压缩机故障诊断中的应用[J]. 石油机械,2009,37(11):63-65.
[15] LI Xiao-li,NIE Peng-cheng,QIU Zheng-jun,et al. Using wavelet transform and multi-class least square support vector machine in multi-spectral imaging classification of Chinese famous tea[J]. Expert Systems with Applications,2011,38(9):11149-11159.
[16] FU Jui-hsi,LEE Sing-ling. A muti-class SVM classification system based on learning methods from indistinguishable chinese official documents[J]. Expert Systems with Applications,2012,39(7):3127-3134.
[17] 史峰,王小川,郁磊,等. MATLAB 神经网络 30 个案例分析[M]. 北京:北京航空航天大学出版社,2010.
[18] CHELOUAH R,SIARRY P. A hybrid method combining continuous tabu search and Nelder-Mead simplex algorithms for the global optimization of multiminima functions[J]. European Journal of Operational Research,2005,161(3):636-654.

An Evaluation Model of Wine Quality Based on Least Square Support Vector Machine

WU Rui-hong, WANG Ya-li,
ZHANG Huan-chong, WANG Xian-fang

(School of Computer and Information Technology, Henan Normal University, Xinxiang 453007, China)

Abstract: In this paper, the wine dataset from UCI databases is preprocessed and radial basis function is adopted as the kernel function of least square support vector machine (LS-SVM). And then a multi-classifier is designed from LS-SVM according to one-against-one algorithm. In addition, the cross-validation method is used to optimize parameters and the wine quality evaluation model is built. Meanwhile, LS-SVM is used in the wine quality evaluation and compared with the evaluation methodology based BP (back propagation) neural network and standard support vector machine. Simulation results show that the LS-SVM can achieve higher accuracy than BP neural network and standard support vector machine, with a highest 100% rate.

Keywords: least square support vector machine; wine; multiple classifier; cross validation; quality evaluation

（**责任编辑**：钱筠　　**英文审校**：吴逢铁）