

# 基于 LIBSVM 的葡萄酒品质评判模型

高媛媛, 刘强国

(四川理工学院理学院, 四川 自贡 643000)

**摘要:**葡萄酒的成分复杂,是划分葡萄酒品质的重要依据。文章通过对 178 个葡萄酒样品化学分析数据进行分析处理,其中葡萄酒属性 13 个,建立基于支持向量机的葡萄酒品质评判模型,利用 LIBSVM 工具对高维复杂葡萄酒属性数据进行分析、处理、优化和解释,分类结果准确率高达 98%,从而对葡萄酒品质快速有效的评判提供了理论依据。

**关键词:**支持向量机;核函数;径向基函数;惩罚因子

**中图分类号:** TP183

**文献标识码:** A

## 引言

葡萄酒化学成分复杂,葡萄酒的质量是各种化学成分的综合反映。通常检测的方法有感官评定和常规的理化指标检测。品质检测主要方法是化学分析方法,国内外普遍采用近红外光谱和三维荧光光谱等技术,该技术需要复杂的化学计量学知识,解释困难,不仅繁琐、费时和费用高,而且不能对所有成分分析,检测结果不全面。感官评审法虽然在生产中也有较多的应用,但是评测周期长、影响因素多、主观性强和重复性差并且人的器官具有适应性,容易出现疲劳影响分析结果,无法实现快速检测。随着模式识别技术的发展,其应用领域越发广阔。模式识别技术在葡萄酒品质分类中得到了广泛的应用。

## 1 模型分析

### 1.1 数据

数据源来自 UCI 数据库中的 wine 数据包<sup>[1]</sup>,该数据包含意大利在不同地点所生产的三种葡萄酒的资料,特性如下:实例共 178 个,特征共 13 个,都是由化学分析所得到的数值,没有未知量。根据化学测试得到的输入变量包括 13 个,分别是 Alcohol(g/L), Malic acid(g/L), Ash(g/L), Alca linity of ash(g/L), Magnesium(g/L), Total

phenols(g/L), Flavonoids(g/L), Nonflavonoid phenols(g/L), Proanthocyanins(g/L), Cobr intensity(g/L), Hue(g/L), OD280/OD315 of diluted wines(g/L), Prolin(e)(g/L)。三个分类实例的数目分别为:类 1 有 59 个,类 2 有 71 个,类 3 有 48 个。

建立模型的目的是对 178 个葡萄酒样品化学分析数据进行分析,其中葡萄酒属性 13 个,从而对 3 种不同品质的葡萄酒进行分类。

### 1.2 支持向量机

支持向量机<sup>[2]</sup>(support vectomachines SVM)是在统计学习的基础上发展起来的一种新的机器学习方法,它是建立在统计学习理论的 VC 维理论和结构风险最小化原则上的,避免了局部极小点(支持向量机是一种凸二次优化问题,能够保证极值点是全局最优解),并能够解决过学习问题,具有良好的推广性能和较好的分类精确性(由有限训练样本得到的决策规则对独立的测试集仍能够得到小的误差)。支持向量机在解决小样本、非线性及高维模式识别问题中表现的许多特有的优势,使它成为一种优秀的机器学习算法,所以对于 178 个样本,每一类样本实例数目较少,属性较多,采用 VSM 方法适合<sup>[3]</sup>。

## 2 模型建立与优化

### 2.1 理论

收稿日期: 2010-07-07

基金项目: 四川理工学院人才引进科研启动项目(2008RCYJ09); 四川理工学院人才引进科研启动项目(07ZR41)

作者简介: 高媛媛(1981-)女,山东东营人,助教,硕士,主要从事模式识别与智能系统方面的研究。

SVM 的主要思想可以概括为两点: (1) 它是针对线性可分情况进行分析, 对于线性不可分的情况, 通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分, 从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能; (2) 它基于结构风险最小化理论之上在特征空间中建构最优分割超平面, 使得学习器得到全局最优化, 并且在整个样本空间的期望风险以某个概率满足一定上界。

支持向量机的非线性变换是通过定义适当的内积(核)函数实现的, 通过非线性变换将输入空间变换到一个高维空间, 然后在这个新空间中求最优线性分类超平面。

核函数方法具有以下特点: (1) 核函数的引入避免了“维数灾难”, 大大减小了计算量。而输入空间的维数  $n$  对核函数矩阵无影响, 因此, 核函数方法可以有效处理高维输入。(2) 无需知道非线性变换函数  $\Phi$  的形式和参数。(3) 核函数的形式和参数的变化会隐式地改变从输入空间到特征空间的映射, 进而对特征空间的性质产生影响, 最终改变各种核函数方法的性能。(4) 核函数方法可以和不同的算法相结合, 形成多种不同的基于核函数技术的方法, 且这两部分的设计可以单独进行, 并可以为不同的应用选择不同的核函数。所以核函数的选择是模型的核心, 直接影响分类的效率与效果。

在线性不可分情况下, 引入松弛项  $\xi$

$$y_i [f(w \cdot x_i) + b] - 1 + \xi \geq 0 \quad i = 1, 2, \dots, n \quad (1)$$

使分类间隔最大, 建立约束条件

$$\|w\|^2 \leq c \quad (2)$$

对于足够小的  $c > 0$  在约束条件 (1) 式和 (2) 式下

只要使  $F^\delta(\xi) = \sum_{i=1}^n \xi_i^\delta$  最小就可以使错分样本数最小。

为计算方便, 设  $\delta = 1$ , 广义最优超平面进一步演化为在条件 (1) 式的约束下求函数最小值。

$$\phi(w, \xi) = \frac{1}{2} (w \cdot w) + C \left( \sum_{i=1}^n \xi_i \right) \quad (3)$$

其中,  $\|w\|^2$  起最大化分类间隔的作用,  $\sum_{i=1}^n \xi_i$  起最小分类错误的作用,  $C$  为某个指定的常数, 它实际上起控制对错分样本惩罚的程度的作用, 实现在错分样本的比例与算法复杂度之间的折中。 $C$  越大, 表示主要把重点放在减少分类错误上,  $C$  值越小, 表示主要把重点放在分离超平面上, 避免过学习问题。通过调整训练集和交叉验证来取得合适的  $C$  值。

支持向量机的多类分类问题采用一对一方法<sup>[4]</sup>。在这种方法中, 要为  $K$  类训练样本构造所有可能的两类

分类器, 每类仅仅在  $K$  类中的两类训练样本上训练, 共有  $K(K-1)/2$  个 SVM 两类分类器。该方法速度快, 鲁棒性强。

$$\min_{w^j, b^j, \xi^j} \frac{1}{2} (w^j)^T w^j + C \sum_{i=1}^n \xi_i^j \quad (4)$$

$$\begin{cases} (w^j)^T \phi(x_i) + b^j \geq 1 - \xi_i^j & y_i = i \\ (w^j)^T \phi(x_i) + b^j \geq -1 + \xi_i^j & y_i \neq i \end{cases} \quad (5)$$

$$\begin{cases} \xi_i^j \geq 0 & j = 1, 2, \dots, n \end{cases} \quad (6)$$

$$\text{最小化 } \frac{1}{2} (w^j)^T w^j \text{ 意味着最大化 } \frac{2}{\|w^j\|} \text{ 也就是最大}$$

化分类间隔。当训练样本不是线性可分的情况下, 增加惩罚项  $C \sum_{i=1}^n \xi_i^j$  来减少训练误差。

组合这些两类分类器用投票法, 如果得到的结果  $\text{sgn}(w^j)^T (\phi(x_i) + b^j)$  显示  $x$  属于第  $i$  类, 那么就给属于第  $i$  类的一方加一票, 否则, 得到的结果显示  $x$  属于第  $j$  类, 那么就给属于第  $j$  类的一方加一票, 得票最多的即为测试样本  $x$  所属类别。如果两个类具有相同的票数, 则简单的选择小索引的类别为测试样本  $x$  所属的类别。

## 2.2 实验

### (1) 收集和整理样本, 并进行标准化

按照 LIBSVM 软件包所要求的格式准备数据集<sup>[5]</sup>, 并对数据进行简单的缩放操作, 将数据分为 traindata 和 testdata 对整个样本数据集, 任选约占其总数 2/3 的数据组成训练集以建立模型, 其余约 1/3 的数据组成测试数据集以测试模型性能。建立 wine scale 分为 wine train 和 wine test 两部分。

### (2) 选择或构造核函数

模型选择是 SVM 算法的一个难点, 目前还没有从理论上很好地解决这个问题。在实际工程中, 研究者主要依靠经验来选取模型。常用的核函数包括线性核、多项式核、径向基核和 sigmoid 核等, 核函数中的核函数参数应该正确设置。径向基核函数参数少, 分类效果较优<sup>[6]</sup>。RBF 核函数  $K(x, y) = e^{-\gamma \|x-y\|^2}$ 。分类方法采用一对一的方法, 该方法较一对多的方法分类速度较快, 鲁棒性较强。

### (3) 对整个训练集进行训练获取支持向量机模型

SVM 核心问题是惩罚因子和核函数参数。惩罚因子控制对大间隔和最小训练错误率之间的平衡, 用于核空间上非线性可分的数据。基于交叉验证和网格搜索对 SVM 进行参数选择  $C$  与  $\gamma$  采用最佳参数  $C$  与  $\gamma$  对整个训练集进行训练获取支持向量机模型<sup>[7]</sup>, 得到 wine train.txt model

### (4) 利用获取的模型进行测试与预测

将  $w_{inetrain}$  代入  $w_{inetrain} \text{ tx model}$  输出  $w_{inetrain} \text{ txt out}$ 。从  $w_{inetrain} \text{ txt out}$  文本可以对比分类数据和实际数据。60个测试样本中, 其中类一 20个, 类二 20个, 类三 20个, 类二 1个测试样本错判为类三, 错误率 1/60。

### 2.3 结果分析

从图 1 可以看到不同颜色等值线, 由于  $C$  和  $g$  的取值不同, 表现出不同分类效果, 如绿色线条  $a$  曲线趋于复杂, 拟合效果好。黄色线条  $f$  曲线趋于平滑, 拟合效果略差。当核函数参数  $g$  取不同的值, 惩罚因子  $C$  需配以不同的值, 拟合效果才好, 因此存在核函数参数数和惩罚因子的最佳匹配问题。当  $C$  和  $g$  取最佳参数时, 准确率高达 98.8%, 实验表明该方法能够对葡萄酒的品质进行快速有效的评判。

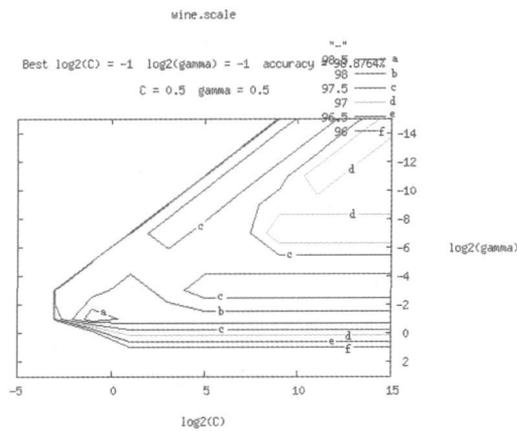


图 1 分类结果

### 3 结束语

基于 SVM 的葡萄酒品质评判模型不受样本空间维数的限制, 取决于支持向量的数目, 只由少数的支持向量所确定, 避免了“维数灾难”, 而且具有较好的“鲁棒”

性, 增、删非支持向量样本对模型没有影响, 支持向量样本集具有一定的鲁棒性。该模型为客观数据驱动评判方法提供了理论依据, 并为理化数据测试方法<sup>[8]</sup>的研究与应用奠定了理论基础。

### 参考文献:

- [1] Frank A, Asuncion A. UCI Machine Learning Repository [DB/OL]. University of California, Irvine, School of Information and Computer Sciences <http://archive.ics.uci.edu/ml/datasets/Wine>
- [2] 肖健华. 智能模式识别方法 [M]. 广州: 华南理工大学出版社, 2006
- [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): 547-553
- [4] Frank A, Asuncion A. UCI Machine Learning Repository [DB/OL]. University of California, Irvine, School of Information and Computer Sciences <http://www.csie.ntu.edu.tw/~cjlin/libsvmto6/datasets/multiclass.htm#wine>
- [5] 黄亮综. 分类器之实作与比较 [R/OL]. <http://www.mdu.edu.tw/~dim/doc/2005052016%20A4%20C0%20C3%20FE%20BE%20B9%20A4%20A7%20B9%20EA%20A7%20BB%20A4%20F%20B8%20FB.pdf>
- [6] 田景文, 高美娟. 人工神经网络算法研究及应用 [M]. 北京: 北京理工大学出版社, 2006
- [7] Hsu Chih-Wei, Chang Chih-Chung, Lin Chih-Jen. A Practical Guide to Support Vector Classification [K/OL]. <http://www.csie.ntu.edu.tw/~cjlin>
- [8] 王金甲, 尹涛, 李静, 等. 基于物理化学性质的葡萄酒质量可视化评价研究 [J]. 燕山大学学报, 2010, 34(2): 133-136

## Model of Wine Quality Identification Based on LBSVM

GAO Yuan-yuan, LIU Qiang-guo

(School of Science, Sichuan University of Science & Engineering, Zigong 643000, China)

**Abstract** Numerous and complex ingredient of wine is an important basis for the quality of wine. After processing and analyzing chemical analysis data of 178 wine samples which contains 13 properties, the model of wine quality identification based on support vector machine is proposed in this article. By means of LIBSVM, the complex high dimension wine property data is analyzed, processed, optimized and interpreted. The precision of result about classifying the wine is 98%, thereby it offers a theoretical basis for identifying the wine quality rapidly and efficiently.

**Key words** support vector machine; demel function; radial basis function; penalty cost