Hans Journal of Data Mining 数据挖掘, 2016, 6(1), 42-53
Published Online January 2016 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2016.61006

Hans汉斯

Research on Wine Quality Prediction Based on Support Vector Machine Method

Enlai Li

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: 825634143@qq.com

Received: Jan. 7th, 2016; accepted: Jan. 22nd, 2016; published: Jan. 27th, 2016

Abstract

With the continuous improvement of people's living standards, wine has become more and more popular among people. Wine production is growing. However, the quality of the wine is still only determined by wine tasters' grading, which is obviously difficult to meet the needs of today's market. Many scholars use data mining algorithms (such as Logistic multinomial model, artificial neural network, support vector machine, decision tree, Bagging, AdaBoost, nearest neighbor algorithm) to predict the wine quality. The results are not very good. The results are reliable and can be used to support vector machine (SVM), which can be used to predict the quality of wine. In this paper, we use the Quality Data Set UCI data to verify the proposed method and the data mining algorithm for comparison, using ten-fold cross validation method to determine the quality of the method.

Keywords

Wine Quality, Prediction, Data Mining, Support Vector Machine

基于支持向量机方法的葡萄酒质量预测研究

李恩来

云南财经大学统计与数学学院，云南 昆明
Email: 825634143@qq.com

收稿日期：2016年1月7日；录用日期：2016年1月22日；发布日期：2016年1月27日

文章引用：李恩来. 基于支持向量机方法的葡萄酒质量预测研究[J]. 数据挖掘, 2016, 6(1): 42-53.
http://dx.doi.org/10.12677/hjdm.2016.61006

摘　要

随着人们生活水平不断的提高，葡萄酒越来越受到人们的喜爱。葡萄酒的产量越来越大。然而葡萄酒质量鉴定手段还是仅靠品酒师的人工品尝打分来判定葡萄酒质量的好坏，显然这种鉴定方式难以满足当今市场的需求。现在有不少学者运用数据挖掘中的一些算法(比如Logistic多项模型，人工神经网络，支持向量机，决策树，Bagging，AdaBoost，最近邻方法等算法)来对葡萄酒质量进行预测研究，其结果并不是很好(误判率均在15%以上)，但相对于仅靠品酒师的人工品尝打分来判定，其结果还是较为可靠的，通过前人的研究可以知道仅仅简单使用支持向量机中的常见核函数，并不能很好的预测葡萄酒质量，因此本文基于支持向量机方法的核函数进行修改，主要将支持向量机方法中常见核函数进行线性组合而得到新的核函数。本文通过使用UCI数据库中的“Wine Quality Data Set”的数据来验证本文所提出的方法与数据挖掘常用的算法进行对比，通过十折交叉验证的方法来判断方法的好坏。

关键词

葡萄酒质量，预测，数据挖掘，支持向量机

1. 研究背景

葡萄酒是用葡萄果实或葡萄汁，经过发酵酿制而成的酒精饮料。在水果中，由于葡萄的葡萄糖含量较高，贮存一段时间就会发出酒味，因此常常以葡萄酿酒。葡萄酒是目前世界上产量最大、普及最广的单糖酿造酒。早在六千年以前，在盛产葡萄的地中海区域，两河流域的苏美尔人和尼罗河流域的古埃及人就会酿造葡萄酒。有趣的是，在舞蹈文化中，有一种葡萄酒舞是在酿酒用葡萄丰收时，庆祝的团体舞蹈。葡萄酒在基督教被视为耶稣基督宝血的象征物。随着科学研究，科学家们发现葡萄酒中含有许多的糖、氨基酸、维生素以及矿物质，都是人体必不可少的营养素，使得葡萄酒有滋补、消化、杀菌、利尿、延缓衰老、防治脑血栓、预防癌症、预防乳腺癌、促进新陈代谢、保护心肌、预防心脏病、软化血管、预防糖尿病、养肺保肺、适量喝红葡萄酒能减少患老年痴呆症危险、防治肾结石等功效与作用，而且葡萄酒有增进食欲、供给热量、缓解疲劳，减肥、美容养颜、防治视网膜变性、有助于提高记忆力以及防治感冒等好处。葡萄酒有这么多对人体有益的作用，葡萄酒在欧美等地区十分受人们的喜爱。近年来，随着我国的经济不断的发展，人们的生活水平不断地提高，我国消费者越来越喜欢葡萄酒，而且对葡萄酒品色和质量的要求不断提高。因此葡萄酒的质量越来越受到广大消费者的关注，而且我国的葡萄酒已经从初级阶段进入到一个发展阶段[1]。如何能够准确地鉴定葡萄酒的质量成为消费者和生产者最关心的问题?传统上，我们都是通过人工品尝打分来进行鉴定葡萄酒的质量，我们要求专业的品酒师通过自己的专业素质根据自己的感官体验(葡萄酒的香气，口感以及外观等指标)来评定葡萄酒的质量。然而这种人工品尝鉴定往往易受到品酒师个人的嗜好、经验以及习惯等因素的影响。这导致葡萄酒的质量会受到品酒师的主观性影响，使得葡萄酒的质量鉴定结果并不是十分可靠。然而不管是国外的葡萄酒市场还是国内的葡萄酒市场都没有一个十分完善的葡萄酒质量标准。葡萄酒的一些理化指标还是能够较为客观的反映葡萄酒质量的好坏。为了使葡萄酒质量鉴定更为科学可靠，不少的学者提出了许多关于葡萄酒质量评定的统计模型和数据挖掘方法[2] [3]。

国外有许多的学者基于葡萄酒的理化指标(例如：酒精的浓度、pH 值、糖的含量、非挥发性酸含量、挥发性酸含量、柠檬酸含量等)以及结合现代的科学技术手段研究葡萄酒的质量。在 1997 年，L.Sun 等人基于数据挖掘中的人工神经网络方法来预测葡萄酒的质量。在 1999 年，Ebeler 等人基于葡萄酒的理化指

标评估以及感官测试来鉴定葡萄酒的质量[4]。而在 2001 年，S.Vlassides 等人使用 NNs 最近邻分类器来预测葡萄酒的质量。随着科技的不断发展，葡萄酒样本的数目不断增加，在 2009 年，Cortez 等人通过使用支持向量机方法来预测葡萄酒的质量，当时的样本数达到了 1000 多，使用的理化指标多达 10 个[5]。

而在国内，对于葡萄酒质量的研究相对于国外较晚，在 2009 年，李运，李记明等人通过了解葡萄酒中的成分与感官之间的相关性来研究葡萄酒质量问题，他们提出通径分析、相关性分析、变异系数分析、主成分分析来研究葡萄酒质量的预测，他们使用了将近 20 多个理化指标[6]。2012 年，张浩彬提出使用数据挖掘方法来预测葡萄酒质量，他主要使用 Logistic 多项式模型，Tan 神经网咯，带偏差项的 BP 神经网咯，决策树，Bagging，Adaboost，最近邻方法等数据挖掘方法。2013 年，裴慧宇提出基于理化指标的 Logit 模型来预测葡萄酒质量，她运用主成分分析，聚类分析对葡萄酒中的主要理化指标进行归类以及筛选，使用 Logistic 模型预测葡萄酒质量。2013 年，陈欣提出建立多元线性回归模型，从而通过回归方程得到葡萄酒质量预测质量得分，将其与实际评分进行误差分析。

国内外的学者提出的方法均能很好通过葡萄酒的理化指标以及葡萄酒的口感等因素预测葡萄酒的质量，但是美中不足的是，他们的预测精确度均在 85%以下，本文主要是基于支持向量机方法[7]-[9]的改进，使得预测精度提高。由于支持向量机核函数的线性组合还核函数，本文主要基于支持向量机常用核函数的线性组合，得到新的核函数，使得误判率最小。也就是预测精度最高。

2. 模型构建

支持向量机(support vector machine)是一种分类算法，通过寻求结构化风险最小来提高学习机泛化能力的机器学习方法，最早是由 Vapnik 在 20 世纪 90 年代所提出的。为了实现经验风险和置信范围的最小化，从而使得在统计样本量较少的情况下，亦能获得良好统计规律的目的。其基本思想是通过最大化分类之间的间隔来达到经验风险和置信范围的最小。通俗来讲，它是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。

已知 n 类数据样本训练集: $x_1^1,\cdots,x_{t_1}^1,\cdots,x_n^1,\cdots,x_{t_n}^1$ 上标代表类别数，t_i 代表第 i 类训练样本数，训练集样本总数为 $t_1+t_2+\cdots+t_n$，其中 $x_i\in R^d$，$y_i\in\{1,2,\cdots,M\}$，M 代表类别数。R^d 上的一个判别函数 $f(x)$，对于任一个输入 x 都有对应的 y 输出值。利用二值分类方法构造 n 类分类器的方法步骤：

1) 首先构造 n 个二值分类器，$f_k(x),k=1,\cdots,n$ 将第 k 类的训练样本和其他训练样本集分开。如果样本 x_i 属于第 k 类，则有 $\operatorname{sgn}\left[f_k(x_i)\right]=1$；否则 $\operatorname{sgn}\left[f_k(x_i)\right]=-1$。

2) 然后，寻找函数 $f_k(x_i)$，$k=1,\cdots,n$ 中最大值所对应的类别即为 x_i 的类别：

$$y_i=\arg\max\left\{f_1(x_i),f_2(x_i),\cdots,f_n(x_i)\right\}$$

对于回归问题，y 不仅仅可以是 –1 和 1。如果令 $\varphi(x)=\omega^{\mathrm{T}}x+b$，希望 y 与 $\varphi(x)$ 的离差越小越好，那么问题就可以归结于求一个 ω 使得 $\|\omega\|^2/2=\omega^{\mathrm{T}}\omega/2$ 取得最小值，但是其约束条件为 $\|y_i-\varphi(x_i)\|\le\varepsilon$，这里 ε 是某一个目标值，这里允许出现一些误差，这样就可以将上面的约束减弱为(对于大于 0 的 ξ_i,ξ_i^*) $y_i-\varphi(x_i)\le\varepsilon+\xi_i$ 和 $y_i-\varphi(x_i)\le\varepsilon+\xi_i^*$，就可以得出 Lagrange 函数：

$$\begin{aligned}L(\omega,b,\xi,\alpha,\eta)=&\frac{1}{2}\|\omega\|^2+C\sum_i\left(\xi_i+\xi_i^*\right)-\sum_i\left(\eta_i\xi_i+\eta_i^*\xi_i^*\right)\\&-\sum_i\alpha_i\left(\varepsilon+\xi_i-y_i+\omega^{\mathrm{T}}x_i+b\right)-\sum_i\alpha_i^*\left(\varepsilon+\xi_i^*-y_i+\omega^{\mathrm{T}}x_i+b\right)\end{aligned}$$

需要在约束的条件下 $\alpha,\eta>0$ 下，解 $\min_{\omega,b,\xi}\left\{\max_{\alpha,\eta}L(\omega,b,\xi,\alpha,\eta)\right\}$ 问题。

当样本是线性可分时，支持向量机(SVM)算法可以通过构造线性的最优分类超平面 $w\times x+b=0$，能

够将两个类别的样本十分准确地分开。而当样本不是线性可分时，支持向量机(SVM)算法可以通过非线性的映射，将原样本映射到高维特征空间，其映射为：$\phi: x \in R^m \to \phi(x) \in R^n \ (n > m)$。

一个支持向量机(SVM)具有良好的性能，关键是核函数的构造，核函数的构造主要包括两个部分的工作：一是核函数的类型构造，二是在确定好核函数的类型后一些相关参数的选择，因此如何根据具体的数据来构造合适的核函数，是支持向量机(SVM)应用领域遇到的一个重大难题，这也成为了研究人员所关注的焦点，即便如此，现在依然没有具体的理论和方法来指导构造合适的核函数。

其核函数的定义并不困难，只需要根据泛函的有关理论，如果有一种函数 $K(x_i, x_j)$ 能够满足 Mercer 条件，那么它就对应某一变换空间的内积。目前根据 Mercer 定理可以得到以下几种常用的核函数类型：

(1) 线性核函数： $K_1 = K(x, x_i) = x \cdot x_i$；

(2) 多项式核： $k_2 = k(x, x_i) = ((x \cdot x_i) + 1)^d$；

(3) 径向基核(RBF)： $K_3 = K(x, x_i) = \exp\left(-\frac{|x - x_i|}{\sigma^2}\right)$ 高斯径向基函数则是局部性强的核函数，其外推能力随着参数 σ 的增大而减弱。多项式形式的核函数具有良好的全局性质。局部性较差；

(4) 傅里叶核： $K_4 = K(x, x_i) = \frac{1 - q^2}{2(1 - 2q\cos(x - x_2) + q^2)}$。

在本文根据核函数的一些性质：

性质 1：$k: x \times x \to R$ 是核函数，当且仅当它是正定的。

性质 2：若 $K_1, K_2, \cdots$ 是核函数，则 $K_1 + K_2$ 是核函数；αK_1 是核函数；$K_1 K_2$ 是核函数。

对训练集 $\{(x_1, y_1), \cdots, (x_l, y_l)\}$，拟合函数 $f(x) = \omega^{\mathrm{T}} \varphi(x) + b$ 可以通过求解下面的最优问题得到下式。

$$\min_{\omega, \varsigma, b, \xi} \frac{1}{2}\omega^{\mathrm{T}}\omega + C\sum_{i=1}^{l}\varsigma_i + C\sum_{i=1}^{l}\xi_i$$

$$s.t. \quad y_i - \omega^{\mathrm{T}}\varphi(x_i) - b \le \varepsilon + \varsigma_i$$

$$\omega^{\mathrm{T}}\varphi(x_i) + b - y_i \le \varepsilon + \xi_i$$

$$\varsigma_i, \xi_i \ge 0, \ i = 1, \cdots, l$$

其中，l 为训练集的样本数，C 为惩罚因子，主要调节回归函数的准确率，值越小准确率越高，如果过小的话，会出现过拟合现象。而 $\varsigma = (\varsigma_i, \cdots, \varsigma_l) \in \mathcal{R}^l$ 和 $\xi = (\xi_1, \cdots, \xi_l) \in \mathcal{R}^l$ 为松弛因子。将上面的式子进行对偶化，并将其转化成半正定规划的标准形式，这样方便用 Matlab 求解，上式可转化成下面式子

$$\min_{\alpha} 2h^{\mathrm{T}}\alpha - \alpha^{\mathrm{T}}Q(K)\alpha$$

$$s.t. \quad y^{\mathrm{T}}\alpha = 0 \quad \alpha_i \in [0, C], i = 1, \cdots, 2l$$

其中，$h = \begin{pmatrix} -\varepsilon e + z \\ -\varepsilon e - z \end{pmatrix} \in \mathcal{R}^{2l}$，$e$ 为单位向量，$z = (z_1, \cdots, z_l)^{\mathrm{T}}$，$\varepsilon$ 为非负的不敏感损失函数的参数。$Q(K) = \begin{pmatrix} K & -K \\ -K & K \end{pmatrix} \in \mathcal{R}^{2l}$，$y = (1, -1) \otimes 1_l$。由性质 2 可以知道，核函数的线性组合仍为核函数，因此可以通过上面介绍 4 种的常用的核函数来构造出对应的实际数据最优核函数。先对核函数矩阵进行归一化，也就是对原核做一个特征映射：$\overline{K(x, y)} = \frac{K(x, y)}{\sqrt{K(x, x), K(y, y)}}$，归一化后的核矩阵线性组合，要满足条件：

$$K = \sum_{i-1}^{n}(\alpha_i K_i), \ K \ge 0, \ trace(K) \le 0$$

其中，$\{K_1,\cdots,k_n\}$为初始核函数矩阵集，α为目标向量，可以得到目标函数为：

$$\min_{K\in\kappa}\max_{\alpha} 2h^{\mathrm{T}}\alpha-\alpha^{\mathrm{T}}Q(k)\alpha$$

$$s.t.\ \ y^{\mathrm{T}}\alpha=0$$

$$\alpha_i\in[0,C],\ i=1,\cdots,2l$$

$$trace(K)=c$$

其中，c为约束参数，主要目的是约束K的迹来提高搜索效率，防止出现过拟合现象。将拉格朗日乘数应用于目标函数，那么将有约束条件的最优化问题转化成 SDP 的标准形式代入 libsvm 软件包内所提供的网格搜寻算法搜寻上面 4 种常用的核函数的最优参数。在求解凸二次规划时使用 yalmip 软件包。

3. 葡萄酒数据初步分析

3.1. 数据来源

本文实证分析所选用的葡萄酒数据来源于University of Minho提供的Wine Quality Data Set的数据(该数据集由 P.Cortez 等三人在 2009 年向 UCI 机器学习网站捐赠的数据，数据来源网址：http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/)，该数据集共计 1599 个样本，包含了 11 个葡萄酒的理化指标和化学指标，还有一个表示葡萄酒质量的指标。本文根据前人的一些研究，将葡萄酒质量的指标分成三个质量等级，分别为高等、中等以及低等，其分别对应的样本数为 217、1319、63。本文使用 libsvm 软件包、yalmip 软件包以及 R 软件对数据进行分析。

3.2. 数据变量说明

fixed acidity (非挥发性酸)：fixed acidity 的含量会影响葡萄酒的口感。fixed acidity 是指酒石酸的含量，这种酸在葡萄酒加热时不会挥发出去。fixed acidity 是连续变量。

volatile acidity (挥发性酸)：volatile acidity 是指醋酸的含量，这种酸在葡萄酒加热时会挥发出去。volatile acidity 是连续变量。

citric acid (柠檬酸)：citric acid 主要用于抑制有害细菌的发育和去除葡萄酒中多余的铁以及铜。citric acid 是连续变量。

residual sugar (残余糖分含量)：residual sugar 主要反映葡萄酒的甜度，其会让葡萄酒的酒味更加的柔和。residual sugar 是连续变量。

chlorides (氯化钠)：chloride 主要影响葡萄酒的适口性。chloride 是连续变量

free sulfur dioxide (游离二氧化硫)：free sulfur dioxide 在葡萄酒中主要有杀菌、澄清、抗氧化、增酸，其能够使葡萄酒酒的风味变好等作用。free sulfur dioxide 的含量不能过高，由于其对人体有害。free sulfur dioxide 是连续变量。

total sulfur dioxide (总二氧化硫)：total sulfur dioxide 是指葡萄酒中的游离二氧化硫和绑定二氧化硫的总含量。total sulfur dioxide 是连续变量。

Density (密度)：Density 是指葡萄酒的密度，如果葡萄酒的密度越大，那么其口感就会越丰郁。Density 是连续变量。

pH (酸碱度)：pH 主要影响葡萄酒的风味和颜色。pH 是一个极易管理和控制的质量参数，是连续变量。

sulphates (硫酸盐)：sulphates 主要影响葡萄酒的香气。Sulphates 是连续变量。

Alcohol (酒精含量)：Alcohol 能够使得葡萄酒更加具有甜润感，而且 Alcohol 也能反映葡萄酒的浓厚度。Alcohol 是连续变量。

quality (质量)：quality 是指葡萄酒质量的一个指标。本文将 quality 分成三个质量等级，编码成 1、2、3。其中 1 表示葡萄酒质量属于低等，2 表示葡萄酒质量属于中等，3 表示葡萄酒质量属于高等。这也本文的目标变量。quality 是分类变量。

3.3. 数据初步分析

为了对数据有一个初步的了解，首先对数据进行标准化处理，防止一些变量的量纲对预测结果有干扰。需要知道自变量之间的相关性如何，那么就要对数据进行相关性分析，图 1 给出了自变量之间相关性的散点图，表 1 给出了自变量之间的相关性矩阵。

通过各变量之间的散点图我们可以观察到非挥发性酸与柠檬酸成线性关系而且斜率为正，则说明非挥发性酸与柠檬酸高度正相关。非挥发性酸与密度同样成线性关系而且斜率为正，则说明非挥发性酸与密度高度正相关。非挥发性酸与 pH 成线性关系但是其斜率为负值，则说明非挥发性酸与密度高度负相关。透过葡萄酒就各变量之间的相关性矩阵表，可以观察到非挥发性酸与密度，酒石酸与密度，非挥发性酸与柠檬酸，酒石酸与 pH 的相关系数均在 0.60 以上。除了非挥发性酸与 pH，挥发性酸与柠檬酸，柠檬酸与 pH 的相关系数在−0.50 以下，其他变量之间的相关系数均较小。说明了葡萄酒中的一些理化指标和化学成分存在着高度的相关性。

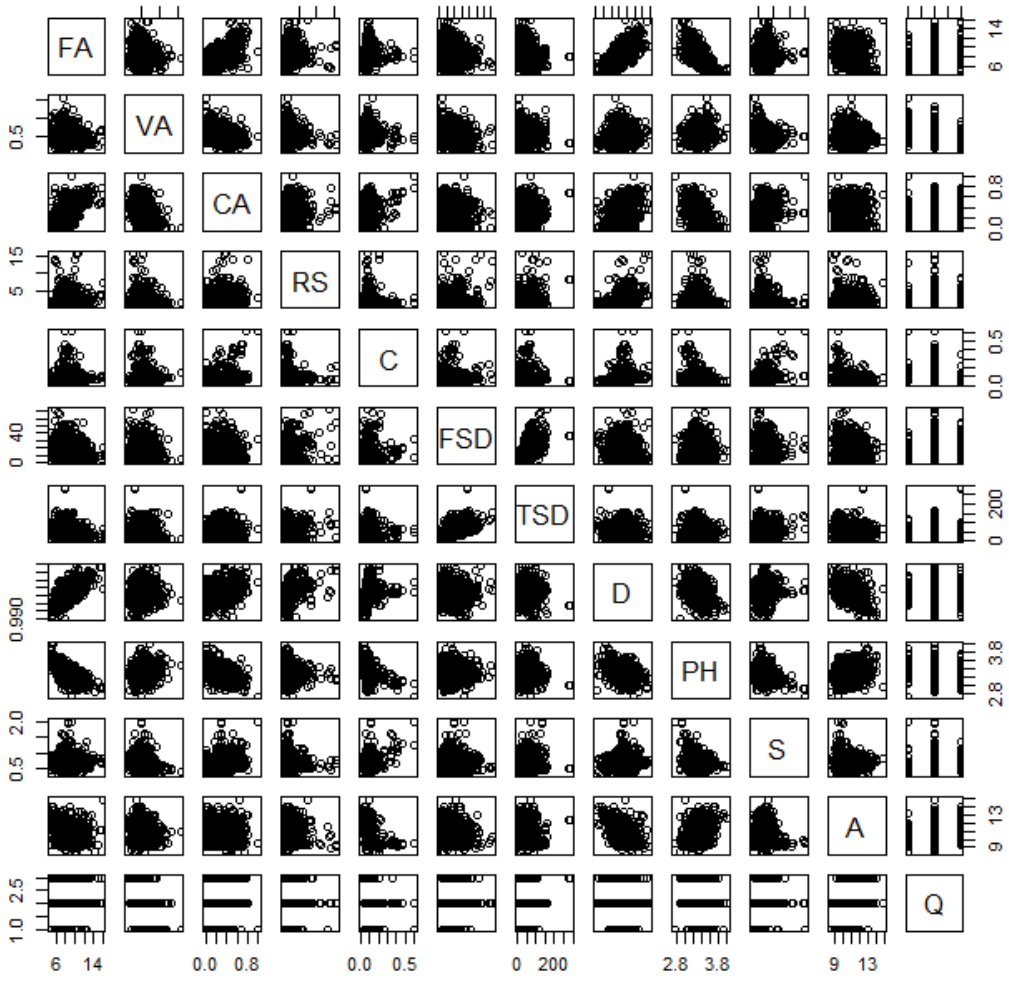

Figure 1. Scatter plots among variables
图 1. 各变量之间的散点图

Table 1. Correlation matrix between variables
表 1. 各变量之间的相关性矩阵

	FA	VA	CA	RS	C	FSD	TSD	D	PH	S	A
FA	1.00	−0.26	0.67	0.11	0.09	−0.15	−0.11	0.67	−0.68	0.18	−0.06
VA	−0.26	1.00	−0.55	0.00	0.06	−0.01	0.08	0.02	0.23	−0.26	−0.20
CA	0.67	−0.55	1.00	0.14	0.20	−0.06	0.04	0.36	−0.54	0.31	0.11
RS	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	−0.09	0.01	0.04
C	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	−0.27	0.37	−0.22
FSD	−0.15	−0.01	−0.06	0.19	0.01	1.00	0.67	−0.02	0.07	0.05	−0.07
TSD	−0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	−0.07	0.04	−0.21
D	0.67	0.02	0.36	0.36	0.20	−0.02	0.07	1.00	−0.34	0.15	−0.50
PH	−0.68	0.23	−0.54	−0.09	−0.27	0.07	−0.07	−0.34	1.00	−0.20	0.21
S	0.18	−0.26	0.31	0.01	0.37	0.05	0.04	0.15	−0.20	1.00	0.09
A	−0.06	−0.20	0.11	0.04	−0.22	−0.07	−0.21	−0.50	0.21	0.09	1.00

4. 葡萄酒数据实验结果

本次实验的样本数据共计 1599 个，所选用的试验方法分别为决策树[10]，Bagging，Adaboost，人工神经网络[11]，最近邻方法，随机森林[12]，常用核函数的两个核函数(高斯和线性)的支持向量机，组合核函数的支持向量机(本文所提出的方法)来分别拟合数据，使用十折交叉验证的方法来验证本文所提出的方法是否比其他的方法效果要好(误判率是否是最低的一个)。下面分别介绍各种方法拟合葡萄酒数据的结果以及十折交叉验证的结果。

4.1. 决策树方法

决策树是我们用于拟合分类数据的方法，而且其效果往往优于经典统计的方法。本文首先使用决策树拟合葡萄酒数据，输出结果如图 2、图 3 以及表 2 所示。从图 2 中，我们可以知道决策树在对葡萄酒进行分类时各个理化指标以及化学成分的分节点在何处，这对我们了解决策树是如何生成一棵树很有帮助。图 3，这是葡萄酒数据生成的一棵决策树。最关键还是要看表 3，因为表 3 为我们提供决策树的分类效果。从表 3 中，我们可以知道决策树将第一类(低等)葡萄酒全部被错分给第二类(中等)葡萄酒，第二类葡萄酒中有 108 个葡萄酒数据样本被错分给第三类(高等)葡萄酒，并且第三类葡萄酒中有 107 个葡萄酒数据样本被错分给第二类葡萄酒。计算其误判率为 17.39%。虽然其误判率低于 20%，但是如今的葡萄酒市场越来越大，而且随着人们生活水平的不断提高，人们对于物质的要求越来越高，那么决策树对葡萄酒质量预测的精确度就显得不够好。这就需要我们寻找对葡萄酒质量预测的精确度更高的方法。

4.2. Bagging 方法

Bagging 方法是数据挖掘常用的方法，其分类效果往往优于经典统计的效果。本文使用 Bagging 方法拟合葡萄酒数据的结果如图 4 以及表 3 所示。从图 4 中，我们可以看出，Bagging 方法在葡萄酒数据分类时各个理化指标以及化学成分的重要性是不一样的，非挥发性酸、硫酸盐以及酒精含量在 Bagging 方法在葡萄酒数据分类时，重要性要大一些，其他变量重要性差不多。从最为关键的表 3 中，我们可以看出，我们可以知道决策树将第一类(低等)葡萄酒全部被错分给第二类(中等)葡萄酒，第二类葡萄酒中有 38 个葡萄酒数据样本被错分给第三类(高等)葡萄酒，并且第三类葡萄酒中有 154 个葡萄酒数据样本被错分给第二类葡萄酒。计算其误判率为 15.94%。相比决策树效果好一些，但是其误判率还是相对有些高。

```
n= 1599

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 1599 280 2 (0.039399625 0.824890557 0.135709819)
   2) A< 11.55 1349 167 2 (0.044477391 0.876204596 0.079318013)
     4) VA>=0.375 1111 104 2 (0.051305131 0.906390639 0.042304230) *
     5) VA< 0.375 238  63 2 (0.012605042 0.735294118 0.252100840)
      10) A< 10.45 120  13 2 (0.016666667 0.891666667 0.091666667) *
      11) A>=10.45 118  50 2 (0.008474576 0.576271186 0.415254237)
        22) PH>=3.265 68  18 2 (0.014705882 0.735294118 0.250000000) *
        23) PH< 3.265 50  18 3 (0.000000000 0.360000000 0.640000000)
          46) C>=0.0885 10   3 2 (0.000000000 0.700000000 0.300000000) *
          47) C< 0.0885 40  11 3 (0.000000000 0.275000000 0.725000000) *
   3) A>=11.55 250 113 2 (0.012000000 0.548000000 0.440000000)
     6) S< 0.685 136  38 2 (0.022058824 0.720588235 0.257352941)
      12) TSD>=15.5 101  18 2 (0.019801980 0.821782178 0.158415842)
        24) FSD< 31.5 91  11 2 (0.021978022 0.879120879 0.098901099) *
        25) FSD>=31.5 10   3 3 (0.000000000 0.300000000 0.700000000) *
      13) TSD< 15.5 35  16 3 (0.028571429 0.428571429 0.542857143)
        26) S< 0.585 19   6 2 (0.052631579 0.684210526 0.263157895) *
        27) S>=0.585 16   2 3 (0.000000000 0.125000000 0.875000000) *
     7) S>=0.685 114  39 3 (0.000000000 0.342105263 0.657894737)
      14) FSD>=18.5 39  16 2 (0.000000000 0.589743590 0.410256410)
        28) FSD< 27.5 22   5 2 (0.000000000 0.772727273 0.227272727) *
        29) FSD>=27.5 17   6 3 (0.000000000 0.352941176 0.647058824) *
      15) FSD< 18.5 75  16 3 (0.000000000 0.213333333 0.786666667) *
```

Figure 2. Decision tree output results for wine data
图 2. 葡萄酒数据的决策树输出结果

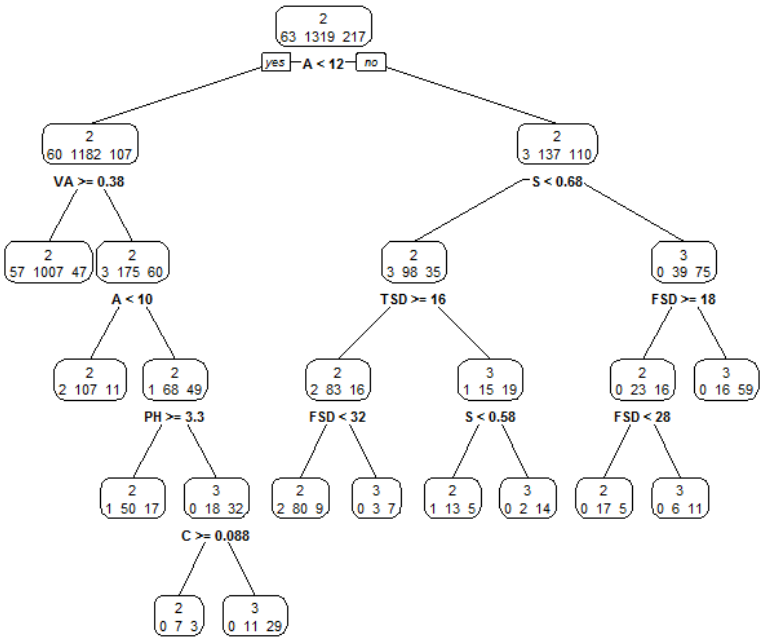

Figure 3. Decision tree for wine data
图 3. 葡萄酒数据的决策树

Table 2. Decision tree classification results on wine data
表 2. 决策树对葡萄酒数据的分类结果

	1	2	3
1	0	63	0
2	0	1211	108
3	0	107	110

Table 3. Bagging classification results of wine data
表 3. Bagging 对葡萄酒数据的分类结果

	1	2	3
1	0	63	0
2	0	1281	38
3	0	154	63

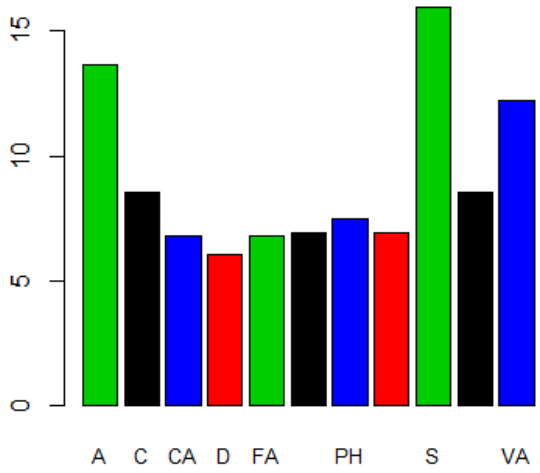

Figure 4. Bagging the variables of the wine data
图 4. Bagging 拟合葡萄酒数据时的变量重要性图

4.3. 随机森林

随机森林方法是数据挖掘常用的方法，其分类效果往往优于经典统计的效果。本文使用随机森林方法拟合葡萄酒数据的结果如图 5 以及表 4 所示。从最为关键的表 4 中，我们可以看出，我们可以知道决策树将第一类(低等)葡萄酒中有 50 个葡萄酒数据样本被错分给第二类(中等)葡萄酒，第二类葡萄酒中有 59 个葡萄酒数据样本被错分给第三类(高等)葡萄酒，并且第三类葡萄酒中有 143 个葡萄酒数据样本被错分给第二类葡萄酒。计算其误判率为 15.76%。相比决策树和 Bagging 方法效果好一些，但是其误判率还是相对有些高。

4.4. 人工神经网络

人工神经网络方法是数据挖掘常用的方法，其分类效果往往优于经典统计的效果。本文使用人工神经网络方法拟合葡萄酒数据的结果表 5 所示。从最为关键的表 5 中，我们可以看出，我们可以知道决策树将第一类(低等)葡萄酒中有 62 个葡萄酒数据样本被错分给第二类(中等)葡萄酒，第一类葡萄酒中有 1 个葡萄酒数据样本被错分给第三类(高等)葡萄酒，第二类葡萄酒中有 60 个葡萄酒数据样本被错分给第三类(高等)葡萄酒，并且第三类葡萄酒中有 139 个葡萄酒数据样本被错分给第二类葡萄酒。计算其误判率为 16.32%。相比决策树效果好一些，但是其误判率相比于随机森林和 Bagging 方法略高些。

4.5. 支持向量机

支持向量机方法是数据挖掘常用的方法，其分类效果往往优于经典统计的效果。本文使用支持向量机(线性核函数)方法拟合葡萄酒数据的结果如图 5 以及表 4 所示。从最为关键的表 4 中，我们可以看出决策树将第一类(低等)葡萄酒全部被错分给第二类(中等)葡萄酒，第二类葡萄酒中有 76 个葡萄酒数据样本

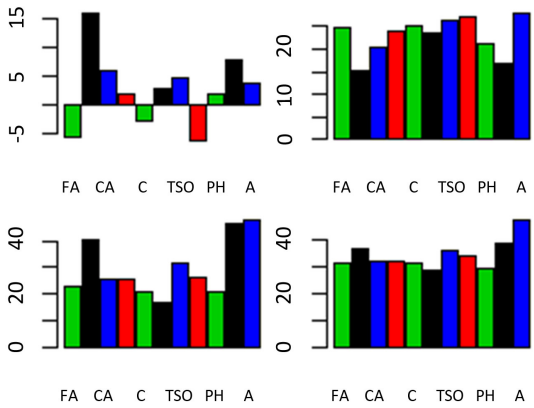

Figure 5. The importance of the variables in the data of a random forest
图 5. 随机森林拟合葡萄酒数据时的变量重要性图

Table 4. Classification of wine data by random forest
表 4. 随机森林对葡萄酒数据的分类结果

	1	2	3
1	13	50	0
2	0	1260	59
3	0	143	74

Table 5. Classification results of grape wine by artificial neural network
表 5. 人工神经网络对葡萄酒数据的分类结果

	1	2	3
1	1	61	1
2	0	1259	60
3	0	139	78

被错分给第三类(高等)葡萄酒，并且第三类葡萄酒中有 120 个葡萄酒数据样本被错分给第二类葡萄酒。计算其误判率为 16.20%。相比决策树和 Bagging 方法效果好一些，但是其误判率还是相对有些高。支持向量机(线性核函数)对葡萄酒数据的分类结果如表 6 所示。

4.6. 支持向量机(组合核函数)

支持向量机(组合核函数)是本文基于常用支持向量机的核函数线性组合得到新的支持向量机方法。首先本文通过参数网格搜索，将惩罚因子 C 定位 48，新核的迹约束 c 定为 $2l$，得到特征向量 $\alpha=(0.8,2.1,0,0)$，则其新的核函数为 $K=0.8K_1+2.1K_2$，本文使用支持向量机(组合核函数)方法拟合葡萄酒数据的结果，如表 7 所示。从最为关键的表 4 中，我们可以看出，我们可以知道决策树将第一类(低等)葡萄酒中有 30 个葡萄酒数据样本被错分给第二类(中等)葡萄酒，第二类葡萄酒中有 18 个葡萄酒数据样本被错分给第三类(高等)葡萄酒，并且第三类葡萄酒中有 79 个葡萄酒数据样本被错分给第二类葡萄酒。计算其误判率为 7.94%。相比其他数据挖掘方法都要好一些。

4.7. 十折交叉验证结果

下面我们使用十折交叉验证的方法来验证究竟是那种方法最优的。十折交叉验证的输出结果如表 8

Table 6. Support vector machine (linear kernel function) classification results on wine data
表 6. 支持向量机(线性核函数)对葡萄酒数据的分类结果

	1	2	3
1	0	63	0
2	0	1243	76
3	0	120	97

Table 7. Support vector machine (combined kernel function) classification results on wine data
表 7. 支持向量机(组合核函数)对葡萄酒数据的分类结果

	1	2	3
1	33	30	0
2	0	1301	18
3	0	79	138

Table 8. The various methods of wine classification data validation ten-fold cross-validation rate of false positives
表 8. 各种方法对葡萄酒数据分类的十折交叉验证验证的误判率

分类方法	误判率
决策树	0.1575832
Bagging	0.1518992
Adaboost	0.1524813
随机森林	0.1488992
人工神经网络	0.1593449
支持向量机(线性核函数)	0.1625522
支持向量机(高斯核函数)	0.1553949
最近邻方法	0.1519273
支持向量机(组合核函数)	0.1097768

所示，可以看出支持向量机(组合核函数)是这些方法中误判率最小的一个。说明支持向量机(组合核函数)的方法在对葡萄酒数据拟合效果要优于其他的常用数据挖掘方法。

5. 结论

本文首先回顾了前人对葡萄酒质量预测的一些研究，发现使用常用数据挖掘方法对葡萄酒质量预测效果太多误判率在 15%以上，虽然常用数据挖掘方法的误判率都低于 20%，但是如今的葡萄酒市场越来越大，而且随着人们生活水平的不断提高，人们对于物质的要求越来越高，那么决策树对葡萄酒质量预测的精确度就显得不够好。这就需要我们寻找对葡萄酒质量预测的精确度更高的方法。因此本文就提出基于支持向量机方法的改进，使得预测结果更加的精确。从前人的研究中发现，支持向量机方法的核心是核函数，只要确定了核函数，那么确定一种可用的支持向量机的方法了，而且通过前人对核函数的研究可以发现核函数的线性组合还是核函数。本文就是基于常用核函数的线性组合得到新的核函数，使得它对葡萄酒质量预测效果最好(支持向量机的核函数线性组合效果最好的一个)。通过实验发现，它不仅是支持向量机中最好的一个，也是常用数据挖掘方法中最好一个，它的误判率仅为 10%左右。相对其他的

方法要优秀一些。然而，本文只是简单进行线性组合，我们还可以进一步优化支持向量机的核函数使得其结果进一步好。还可以优化人工神经网络的残差项，使得人工神经网络的方法拟合数据效果更加的好。

参考文献 (References)

[1] 张建生. 中国葡萄酒市场白皮书(2007.2008) [EB/OL]. http://www.redwinelife.com, 2009-11.

[2] 林翠香. 基于数据挖掘的葡萄酒质量识别[D]: [硕士学位论文]. 长沙: 中南大学, 2010.

[3] Han, J.W. and Kamber, M. 数据挖掘: 概念与技术[M]. 第 3 版. 北京: 机械工业出版社, 2012.

[4] Ebeler, S.E. (1999) Linking Flavor Chemistry to Sensory Analysis of Wine. In: *Flavor Chemistry—Thirty Years of Progress*, Kluwer Academic Publishers, Dordrecht, 409-422. http://dx.doi.org/10.1007/978-1-4615-4693-1_35

[5] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009) Modeling Wine Preferences by Data Mining from Physicochemical Properties. *Decision Support System*, **47**, 533-547. http://dx.doi.org/10.1016/j.dss.2009.05.016

[6] 李运, 李记明. 统计分析在葡萄酒质量评价中的应用[J]. 酿酒科技, 2009(4): 79-80.

[7] 徐海涛. 改进的近似支持向量机在葡萄酒质量鉴定中的应用[J]. 安徽农业科学, 2010, 38(29): 16105-16106.

[8] 许志卿, 苏喜友, 张顾. 基于支持向量机方法的森林火险预测研究[J]. 中国农学通报, 2012, 28(13): 126-131.

[9] Cherkassy, V. and Ma, Y. (2004) Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. *Neural Networks*, **17**, 113-126. http://dx.doi.org/10.1016/S0893-6080(03)00169-2

[10] 李强. 创建决策树算法的比较研究——ID3, C4. 5, C5. 0 算法的比较[J]. 甘肃科学学报, 2007, 18(4): 84-87.

[11] 刘延玲. 新的 Hopfield 神经网络分类器在葡萄酒质量评价中的应用[J]. 价值工程, 2012(2): 181-182.

[12] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. http://dx.doi.org/10.1023/A:1010933404324