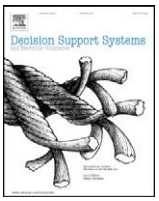
目录列表位于科学指导

决策支持系统

期刊主页：www.elsevier. com/locate/dss



通过物理化学性质的数据挖掘对葡萄酒偏好进行建模

保罗·科尔特斯[a,*]，安东尼奥·塞德拉（AntónioCerdeira）[b]，费尔南多·阿尔梅达[b]，Telmo Matos [b]，何塞·里斯（JoséReis）[a,b]

[a]Minho大学信息系统/研发中心Algoritmi，葡萄牙吉马良斯4800-058
[b]Vinho Verde地区葡萄栽培委员会（CVRVV），葡萄牙波尔图4050-501

article info

文章历史：
2008年7月28日收到
以修订版收到2009年5月22日接受2009年5月28日接受
在线可用2009年6月6日

关键字：
感官偏好回归
变量选择型号选择
支持向量机神经网络

摘 要

我们提出了一种数据挖掘方法来预测人的葡萄酒口味偏好，该方法基于在认证步骤中易于获得的分析测试。考虑了一个较大的数据集（与该领域的其他研究相比），带有白色和红色的vinho verde样本（来自葡萄牙）。在执行同时变量和模型选择的高效计算程序下，应用了三种回归技术。支持向量机取得了可喜的结果，优于多元回归和神经网络方法。这种模型对于支持酿酒师的品酒评估和提高葡萄酒产量很有用。此外，类似的技术可以通过对利基市场的消费者口味进行建模来帮助目标市场。

1. 介绍

过去，葡萄酒一度被视为奢侈品，但如今越来越受到越来越多消费者的青睐。葡萄牙是世界十大葡萄酒出口国，2005年市场份额为3.17%[11]. 从1997年到2007年，其Vinho Verde葡萄酒（来自西北地区）的出口增长了36%[8]. 为了支持其发展，葡萄酒行业正在投资用于酿酒和销售过程的新技术。葡萄酒认证和质量评估是这种情况下的关键要素。认证可防止葡萄酒被非法掺假（以维护人类健康），并确保葡萄酒市场的质量。质量评估通常是认证过程的一部分，可用于改善葡萄酒酿造（通过确定最有影响力的因素）以及对诸如高级品牌之类的葡萄酒进行分层（可用于确定价格）。

葡萄酒认证通常通过理化和感官测试来评估[10]. 常规用于表征葡萄酒特性的物理化学实验室测试包括密度，酒精或pH值的测定，而感官测试则主要依靠人类专家。应该强调的是，味觉是人类感官中最少了解的[25] 因此，葡萄酒的分类是一项艰巨的任务。此外，理化和感官分析之间的关系很复杂，但仍未完全了解[20].

信息技术的进步使得收集，存储和处理海量的，常常是高度复杂的数据集成为可能。所有这些数据包含趋势和模式等有价值的信息，可用于改进决策制定并优化成功机会[28]. 数据挖掘（DM）技术[33] 旨在从原始数据中提取高级知识。有几种DM算法，每种算法都有自己的优势。在对连续数据建模时，线性/多元回归（MR）是经典方法。反向传播算法于1974年首次引入[32] 后来在1986年普及[23]. 从那时起，神经网络（NNs）越来越多地被使用。最近，还提出了支持向量机（SVM）[4,26]. 由于其更高的灵活性和非线性学习能力，NN和SVM都在DM领域引起了关注，通常获得较高的预测性能[16,17]. SVM比NN具有理论上的优势，例如在学习阶段不存在局部最小值。实际上，SVM最近被认为是最有影响力的DM算法之一[34]. 尽管MR模型更易于解释，但仍然可以从神经网络和支持向量机中提取知识，具体取决于输入变量的重要性[18,7].

当应用这些DM方法时，变量和模型选择是关键问题。变量选择[14] 有助于丢弃不相关的输入，从而产生更易于解释且通常可提供更好性能的更简单模型。复杂的模型可能会过度拟合数据，从而失去泛化能力，而过于简单的模型将带来有限的学习能力。实际上，NN和SVM都具有需要调整的超参数[16]，例如NN隐藏节点的数量或SVM内核参数，以便获得良好的预测准确性（请参见第2. 3节）.

葡萄酒行业决策支持系统的使用主要集中在葡萄酒生产阶段[12]. 尽管DM技术有潜力根据理化数据预测葡萄酒质量，

◆ 通讯作者。电话：+351 253510313；传真：+351 253510300。
电子邮件地址：pcortez@dsi.uminho.pt (P. Cortez).

0167-9236 / $ –见前题土井：10.1016/j.dss.2009.05.016

它们的使用非常稀少，并且大多考虑使用小型数据集。例如，在1991年，“葡萄酒”数据集被捐赠到UCI存储库中[1]。数据包含178个实例，测量了13种化学成分（例如，酒精，镁），目标是对来自意大利的三个品种进行分类。该数据集非常容易区分，并且主要用作新DM分类器的基准。1997年[27]，使用神经网络输入15个输入变量（例如锌和镁的含量）来预测六个地理上的葡萄酒起源。数据包括来自德国的170个样本，报告的预测率为100%。2001年[30]，基于葡萄的成熟度水平和化学分析（例如可滴定的酸度），使用神经网络对加州葡萄酒的三个感官属性（例如甜度）进行分类。仅使用了36个示例，并且实现了6%的误差。几个理化参数（例如酒精，密度）用于[20]来表征56种意大利葡萄酒。然而，作者们认为，将这些参数与感官味觉小组进行映射是一项非常艰巨的任务，相反，他们使用了从电子舌中获取数据的神经网络。最近，使用矿物表征（例如锌和镁）将54个样品区分为两种红酒[21]。采用概率神经网络，达到95%的准确性。作为一种功能强大的学习工具，SVM在诸如预测肉类偏好等应用中的性能优于NN[7]。但是，在葡萄酒质量领域，仅报道了一种应用，其中成功地使用了147瓶的光谱测量结果来预测米酒的3种年龄[35]。

在本文中，我们提供了一个案例分析，该案例基于在葡萄酒认证步骤中容易获得的分析数据来对口味偏好进行建模。建立这样的模型不仅对认证实体有价值，而且对葡萄酒生产商甚至消费者都是有价值的。它可用于支持酿酒师的葡萄酒评估，从而有可能提高其决策的质量和速度。而且，测量理化测试对最终葡萄酒质量的影响对于改善生产过程很有用。此外，它可以帮助目标营销[24]，即通过将类似技术应用于模拟消费者对利基市场和/或有利可图市场的偏好。

这项工作的主要贡献是：

- 我们提出了一种新颖的方法，可以同时对NN和SVM技术进行变量和模型选择。变量选择基于敏感性分析[18]，这是一种计算有效的方法，可测量输入相关性并指导变量选择过程。此外，我们提出了一种简约搜索方法，以较低的计算量来选择最佳的SVM内核参数。
- 我们在实际应用中测试这种方法，vinho verde葡萄酒（来自葡萄牙Minho地区）具有口味偏好，显示了其在这一领域的影响。与以前的研究相比，考虑了一个大型数据集，总共有4898个白色样本和1599个红色样本。葡萄酒的偏好是在回归方法下建模的，该方法保留了等级的顺序，并且我们展示了容差概念的定义如何用于访问不同的性能水平。我们认为，这种集成方法对于支持需要排名感官偏好的应用（例如在葡萄酒或肉类质量保证中）的支持非常有价值。

本文的结构如下：第二节介绍葡萄酒数据，DM模型和变量选择方法；在第三节描述了实验设计并分析了获得的结果；最后得出结论第4节.

2. 材料和方法

2.1. 葡萄酒数据

这项研究将考虑vinho verde，这是葡萄牙的Minho（西北）地区。酒精含量中等吗特别是由于它的新鲜感（特别是在夏天）。这种葡萄酒占葡萄牙总产量的15%[8]，其中约10%出口，主要是白葡萄酒。在这项工作中，我们将分析来自Vinho Verde划界区域的两种最常见的变体，白色和红色（也产生玫瑰红）。数据收集自2004年5月/ 2004年2月/ 2007年，仅使用经过官方认证机构（CVRVV）测试的受保护原始样品名称。CVRVV是一个跨行业的组织，旨在提高Vinho Verde的质量和营销。数据由计算机系统（iLab）记录，该系统自动管理从生产者的要求到实验室和感官分析的葡萄酒样品测试过程。每个条目表示一个给定的测试（分析或感官），并且最终数据库被导出到单个工作表（.csv）中。

在预处理阶段，对数据库进行了转换，以便每行包含一个不同的葡萄酒样品（包括所有测试）。为了避免丢弃示例，仅选择了最常见的理化测试。由于红色和白色的口味差异很大，因此将分别进行分析，因此使用1599个红色和4898个白色示例构建了两个数据集[1]。表格1显示每个数据集的理化统计数据。关于偏好，每个样品至少由三名感官评估者（使用盲品）进行评估，这些评估者将葡萄酒的等级分为0（非常差）至10（优秀）。最终的感官评分由这些评估的中位数给出。图。1绘制目标变量的直方图，表示典型的正态分布（即，具有比极端更高的正态等级）。

2.2. 数据挖掘方法和评估

我们将采用回归方法，以保留首选项的顺序。例如，如果真实等级为3，则预测4的模型要好于预测7的模型。回归数据集D由k个{1，…，N}个示例组成，每个示例都将输入向量映射为I输入变量（x_1，…，x_I）到给定目标y。回归性能通常通过误差度量来衡量，例如平均绝对偏差（MAD）[33]：

$$MAD = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \quad (1)$$

其中$\hat{y}_i$是第k个输入模式的预测值。回归误差特征曲线[2]也用于比较回归模型，理想模型的面积为1.0。该曲线绘制了绝对误差公差T（x轴）相对于公差（y轴）内正确预测的点的百分比（精度）的曲线图。

混淆矩阵通常用于分类分析，其中通过将预测值（以列为单位）与所需的类（以行为单位）匹配来创建C×C矩阵（C是类的数量）。对于有序输出，如果$|y_i - y'_i|$，则预测类由$p_i = y'_i$给出。≤T，否则$p_i = y_i$，其中y'_i表示与y_i最接近的类$y'_i \neq y_i$。从矩阵中，可以使用几个指标来访问总体分类性能，例如准确度和精确度（即预测的列精度）[33].

保持验证通常用于估计模型的泛化能力[19]。该方法将数据随机分为训练和测试子集。前一个子集用于拟合模型（通常使用2/3的数据），而后一个子集（其余1/3）用于计算估计值。一种更强大的估计程序是k倍交叉验证[9]，其中数据分为大小相等的k个分区。每次测试一个子集，其余数据用于拟合模型。顺序重复此过程，直到所有子集都经过测试。因此，

[1] 数据集位于：http://www3.dsi.uminho.pt/pcortez/wine/.

表格1
每种葡萄酒的理化数据统计。

属性（单位）	红酒			白色 葡萄酒		
	敏	最高	意思	敏	最高	意思
固定酸度（g（酒石酸）/ dm³）	4.6	15.9	8.3	3.8	14.2	6.9
挥发性酸度（g（乙酸）/ dm³）	0.1	1.6	0.5	0.1	1.1	0.3
柠檬酸（g / dm³）	0.0	1.0	0.3	0.0	1.7	0.3
残糖（g / dm³）	0.9	15.5	2.5	0.6	65.8	6.4
氯化物（g（氯化钠）/ dm³）	0.01	0.61	0.08	0.01	0.35	0.05
游离二氧化硫（mg / dm³）	1	72	14	2	289	35
二氧化硫总量（mg / dm³）	6	289	46	9	440	138
密度（g / cm³）	0.990	1.004	0.996	0.987	1.039	0.994
pH值	2.7	4.0	3.3	2.7	3.8	3.1
硫酸盐（g（硫酸钾）/ dm³）	0.3	2.0	0.7	0.2	1.1	0.5
酒精含量	8.4	14.9	10.4	8.0	14.2	10.4

在此方案下，所有数据均用于培训和测试。但是，由于拟合了k个模型，该方法需要大约k倍的计算量。

2.3. 数据挖掘方法

我们将采用最常见的NN类型，即多层感知器，其中神经元被分组并通过前馈链接进行连接[3]. 对于回归任务，此NN体系结构通常基于具有逻辑激活的H个隐藏节点的一个隐藏层和具有线性函数的一个输出节点。[16]:

$$\hat{y} = w_{o;0} + \sum_{i=1}^{o-1} \left(\frac{1}{1 + \exp\left(-\left(w_{j;0} + \sum_{j=1}^{I} x_j w_{j;i} \right) \right)} \cdot w_{o;i} \right) \quad (2)$$

其中$w_{i,j}$表示从节点j到i的连接权重，以及
o输出节点。性能对拓扑选择（H）敏感。H = 0的NN等效于MR模型。通过增加H，可以执行更复杂的映射，但是H的过多值将使数据过拟合，从而导致泛化损失。设置H的一种有效计算方法是搜索范围{0，1，2，3，…，$H_{最大值}$}（即，从最简单的NN到更复杂的NN）。对于每个H值，对NN进行训练并测量其泛化估计（例如，在验证样本上）。当泛化降低或H达到最大值（$H_{最大值}$）时，该过程停止。

在SVM回归中[26]，输入xaR^I^通过使用不需要明确知道但取决于内核函数（K）的非线性映射（ϕ）转换为高m维特征空间。SVM的目的是找到最佳的线性分离超平面，在特征空间中拟合数据时容许小的误差（ε）：

$$\hat{y} = w_0 + \sum_{i=1}^{M} w_i \varphi_i(x) \quad (3)$$

ε不敏感损失函数在残差周围设置了不敏感管，并且管中的微小误差被丢弃（图2).

我们将采用流行的高斯核，该核提供的参数比其他核（例如多项式）少[31]：K（x，x'）= exp（−γ|| x −x'||²），γ NO。在这种设置下，SVM性能受三个参数影响：γ，ε和C（交易试穿间错误和映射的平坦度）。为了减少搜索空间，前两个值将使用启发式方法进行设置[5]：C = 3（对于标准化输出）和e = σ̂ = $\sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{N}}$，其中σ= 1.5 / N，ŷ是通过三最近邻算法预测的值。的内核参数（γ）对SVM性能的影响最大，其值太大或太小都会导致预测不佳。设置γ的一种实用方法是从一个极端开始搜索，然后在预测估计值增加的同时向范围的中间搜索[31].

2.4. 变量和模型选择

敏感性分析[18]这是一个简单的过程，将在训练阶段之后应用，并在更改输入时分析模型响应。最初是针对NN提出的，这种敏感性方法也可以应用于其他算法，例如SVM[7]. 令y_a表示通过将除x_a以外的所有输入变量保持在平均值处而获得的输出，x_a在整个范围内随j∈{1，…，L}而变化
水平。如果给定的输入变量（x_a∈{x_1，…，x_l}）是相关的，则它

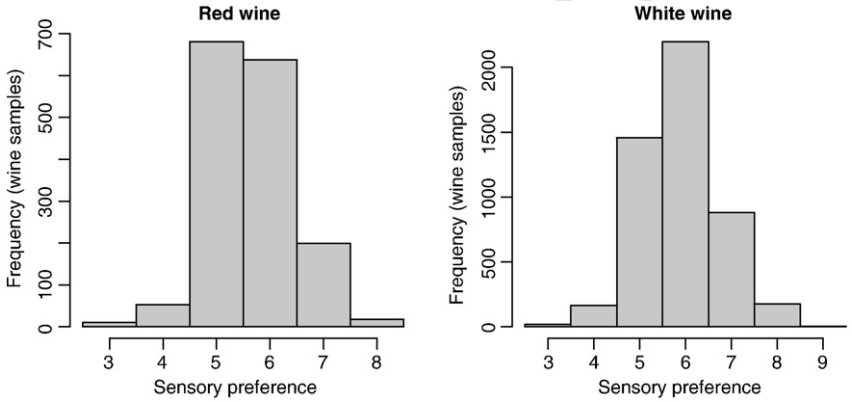


图1. 红色和白色感官偏好的直方图。

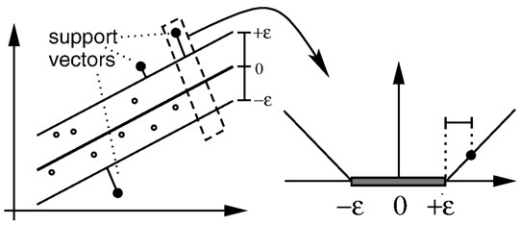


图2. 线性SVM回归和ε不敏感损失函数的示例（改编自[26]）.

应该产生高方差（V_a）。因此，其相对重要性（R_a）可以由下式给出：

$$V_a = \sum_{j=1}^{p} (\hat{y}_j - \hat{y}_{j,-a})^2 \quad (4)$$

$$R_a = V_a / \sum_{i=1}^{p} V_i \times 100\% \quad (5)$$

在这项工作中，R_a值将用于测量输入的重要性，并丢弃不相关的输入，从而指导变量选择算法。我们将采用流行的向后选择，该选择从所有变量开始，并反复删除一个输入，直到满足停止条件为止[14].但是，我们通过敏感性分析来指导变量删除（在每个步骤中），这种变体允许将计算工作量减少I倍（与标准后向程序相比），而[18] 优于其他方法（例如反向算法和遗传算法）。类似于[36]，变量和模型选择将同时执行，即，在每个向后迭代中，将搜索几个模型，其中一个模型将显示最佳的泛化估计。对于给定的DM方法，整个过程如下所示：

(1) 从所有$F = \{x_1, \ldots, x_I\}$输入变量开始。
(2) 如果存在要调整的超参数$P \in \{P_1, \ldots, P_k\}$（例如NN或SVM），则从$P_1$开始并遍历其余范围，直到泛化估计降低为止。计算一般-使用内部验证方法对模型进行规模化估计。例如，如果使用保持方法，则将可用数据进一步分为训练（以拟合模型）和验证集（以获取预测性估计）。
(3) 拟合模型后，计算所有$x_i \in F$变量的相对重要性（R_i），并从F中删除最不相关的输入。如果满足停止条件，则转到步骤4，否则返回步骤2。
(4) 选择最佳的F（在NN或SVM的情况下为P）值，即提供最佳预测估计的输入变量和模型。最后，使用所有可用数据重新训练此配置。

3. 实验结果

R环境[22] 是一种用于统计和数据分析的开源，多平台（例如Windows，Linux）和高级矩阵编程语言。这项工作中报告的所有实验都是用R编写的，并在带有Intel双核处理器的Linux服务器中进行。特别是，我们采用了RMiner[6]，这是R工具的库，可促进在分类和回归任务中使用DM技术。

在拟合模型之前，首先将数据标准化为零均值和一个标准差[16].RMiner使用高效的BFGS算法训练NN（nnet R程序包），而SVM拟合则基于LIBSVM（kernlab程序包）提供的顺序最小优化实现。我们采用了默认的R建议[29].唯一的例外是超参数（H和γ），它们将使用上一部分中所述的过程进行设置，并且搜索范围为$H \in \{0, 1, \ldots, 11\}$[36] 和$\gamma \in \{2^3, 2^1, \ldots, 2^{-15}\}$[31].而
搜索的最大数量是12/10，实际上是简约
方法（第2步，第4节）将大大减少该数字。关于变量选择，我们将估算
指标设置为
MAD值（等式（1）），如建议[31].为了减少计算量，我们采用了更简单的2/3和1/3保留拆分作为内部验证方法。灵敏度分析参数设置为$L = 5$，即$x_a \in \{-1.0, -0.5, \ldots, 1.0\}$用于标准化输入。在寻求简化模型和增加压力之间的合理平衡
在计算搜索中，停止标准设置为2次迭代，没有任何改进，或者只有一个输入可用。

为了评估所选模型，我们采用了20次更可靠的5倍交叉验证，每种测试配置总共进行了20×5 = 100次实验。统计置信度将由t-学生测试在95%置信度下给出[13].结果总结在表2.测试集误差以平均值和置信区间表示。存在三个度量标准：MAD，不同公差的分类精度（即T = 0.25、0.5和1.0）
和Kappa（T = 0.5）。在平均输入数（I）和超参数值（H或γ）的t_erms_中描述了选择sel_的模型。的
最后一行显示所需的总计算时间（以秒为单位）。

对于任务和所有错误指标，SVM是最佳选择。对于小公差，尤其是白葡萄酒，差异更大（例如，对于T = 0.25，与其他方法相比，SVM精度几乎提高了两倍）。绘制完整的REC曲线时，这种效果清晰可见（图3).Kappa统计资料[33] 与随机分类器（其Kappa值为0%）相比，可测量准确性。统计数据越高，结果越准确。最实用的公差值为T = 0.5和T = 1.0。前者的公差将回归响应四舍五入为最接近的类别，而后者接受两个最接近的类别之一中正确的响应（例如，3.1值可以解释为3或4级，但不能解释为2或5级）。对于T = 0.5，红酒的SVM精度提高为3.3 pp（Kappa为6.2 pp），白色任务的SVM精度提高到12.0 pp（Kappa为20.4 pp）。

表2
葡萄酒建模结果（测试设置错误和所选模型；最佳值以粗体显示）。

	红酒			白酒		
	先生	NN	虚拟机	先生	NN	虚拟机
狂	0.50 ± 0.00	0.51 ± 0.00	0.46 ± 0.00[a]	0.59 ± 0.00	0.58 ± 0.00	0.45±0.00[a]
精度$_{T=0.25}$ (%)	31.2 ± 0.2	31.1 ± 0.7	43.2 ± 0.6[a]	25.6 ± 0.1	26.5 ± 0.3	50.3 ± 1.1[a]
精度$_{T=0.50}$ (%)	59.1 ± 0.1	59.1 ± 0.3	62.4 ± 0.4[a]	51.7 ± 0.1	52.6 ± 0.3	64.6 ± 0.4[a]
精度$_{T=1.00}$ (%)	88.6 ± 0.1	88.8 ± 0.2	89.0 ± 0.2[b]	84.3 ± 0.1	84.7 ± 0.1	86.8 ± 0.2[a]
卡伯$_{\kappa=0.5}$ (%)	32.2 ± 0.3	32.5 ± 0.6	38.7 ± 0.7[a]	20.9 ± 0.1	23.5 ± 0.6	43.9 ± 0.4[a]
输入（I）	9.2	9.3	9.8	9.6	9.3	10.1
模型	-	$\bar{H}=1$	$\bar{\gamma}=2^{0.19}$	-	$\bar{H}=2.1$	$\bar{\gamma}=2^{1.55}$
时间（秒）	518	847	5589	551	1339	30674

[a]与MR和NN成对比较时具有统计意义。
[b]与MR成对比较时具有统计意义。

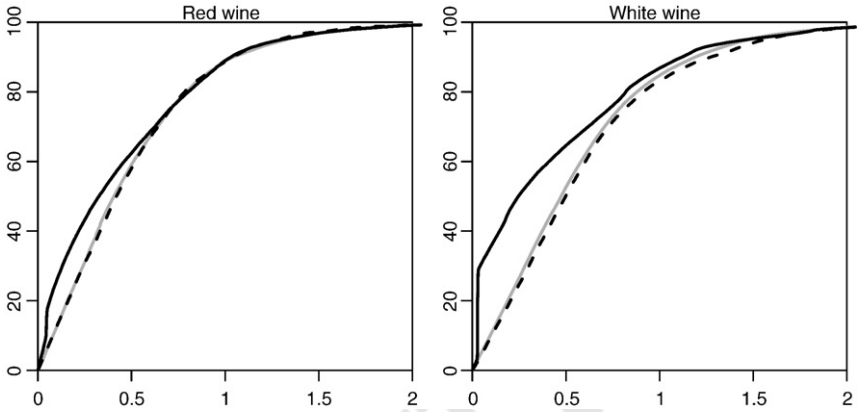


图3. 红色（左侧）和白色（右侧）葡萄酒平均测试集REC曲线（SVM-实线，NN-灰色线和MR-虚线）。

在红酒建模中，NN与MR非常相似，因此相似
性能 实现。对于白色数据，选择了更为复杂的NN模型（H = 2.1），其性能稍好于MR结果。
关于变量选择，已删除输入的平均数量在0.9到1.8的范围内，这表明使用的大多数理化测试都是相关的。就计算工作而言，SVM是最昂贵的方法，尤其是对于较大的白色数据集而言。

对于T = 0.5的平均混淆矩阵，提供了对SVM分类结果的详细分析。表3).为了简化可视化，省略了3级和9级预测，因为这些预测始终为空。大多数值都接近对角线（粗体），表示模型很合适。每个等级的真实预测准确度由精度指标给出（例如，对于4级和白葡萄酒，$precision_{T=0.5}$= 19 / (19 + 7 + 4) = 63.3%）。该统计数据在实践中很重要，因为在实际部署设置中，实际值是未知的，并且给定列中的所有预测都将被视为相同。对于0.5的公差，中级（5至7）的SVM红酒准确度约为57.7至67.5%，而极端级（3、8和4）的SVM红酒准确度则非常低（0%/20%）。减少频率（图。1).通常，白色数据的结果更好：6和4级为60.3 / 63.3%，7和5级为67.8 / 72.6%，8级为令人惊讶的85.5%（例外是3和9极限为0 %，未在表格中显示）。当公差提高（T = 1.0）时，葡萄酒类型和4至8类的准确度将达到81.9至100%。

分析测试的平均SVM相对重要性图（R_a值）显示在图4.应该注意的是整个
显示了11个输入，因为在每个模拟中可以选择不同的变量集。在某些情况下，所得结果证实了酿酒学理论。例如，酒精含量的增加（第4和第二最相关的因素）往往会导致更高质量的葡萄酒。此外，每种葡萄酒的排名也不相同。例如，柠檬酸和残留糖含量在白葡萄酒中更为重要，在白葡萄酒中，新鲜度和甜味之间的平衡更为可取。此外，挥发性酸度有负面影响，因为乙酸是醋中的关键成分。最有趣的结果是硫酸盐的高度重要性，在两种情况下均排名第一。从生物学上来看，这一结果可能非常有趣。硫酸盐的增加可能与发酵营养有关，这对于改善葡萄酒的香气非常重要。

表3
SVM模型的平均混淆矩阵（T = 0.5）和精度值（T = 0.5和1.0）（粗体值表示准确的预测）。

实际班	红色	葡萄酒的预测				白色	葡萄酒 预测			
	4	5	6	7	8	4	5	6	7	8
3	1	7	2	0	0	0	2	17	0	0
4	1	36	15	1	0	19	55	88	1	0
5	3	514	159	5	0	7	833	598	19	0
6	0	194	400	44	0	4	235	1812	144	3
7	0	10	107	82	1	0	18	414	441	7
8	0	0	10	8	0	0	3	71	43	59
9						0	1	3	2	0
精度$_{T=0.5}$ (%)	20.0	67.5	57.7	58.6	0.0	63.3	72.6	60.3	67.8	85.5
精度$_{T=1.0}$ (%)	93.8	90.9	86.6	90.2	100	90.0	93.3	81.9	90.3	96.2

4. 结论与启示

近年来，对葡萄酒的兴趣增加了，从而导致了葡萄酒行业的发展。因此，公司正在投资新技术以改善葡萄酒的生产和销售。质量认证对于这两个过程都是至关重要的一步，目前很大程度上取决于人类专家的品酒。这项工作旨在通过认证步骤中可用的客观分析测试来预测葡萄酒的喜好。考虑了一个大型数据集（包含4898白色和1599红色条目），包括来自葡萄牙西北地区的vinho verde样本。此案例研究通过两个回归任务解决，其中每种葡萄酒类型的偏好都以连续的比例建模，范围从0（非常差）到10（优秀）。这种方法保留了类的顺序，从而可以根据接受的容错度（T）评估不同的准确性。

由于数据挖掘（DM）领域的进步，有可能从原始数据中提取知识。确实，诸如神经网络（NN）和最近的支持向量机（SVM）等强大的技术正在兴起。虽然是更灵活的模型（即没有施加先验限制），但性能取决于超参数的正确设置（例如NN体系结构的隐藏节点数或SVM内核参数）。另一方面，多元回归（MR）比NN / SVM更容易解释，并且大多数NN / SVM应用程序将其模型视为黑盒。另一个相关方面是变量选择，它可以简化模型，同时通常可以提高预测性能。在这项研究中，我们提出了一种集成的，计算效率高的方法来处理这些问题。根据输入的相对重要性，使用灵敏度分析从NN / SVM模型中提取知识。还提出了同时变量和模型选择方案，其中变量选择由敏感性分析指导，并且模型选择基于简约搜索，该简约搜索从合理的值开始，并在泛化估计减少时停止。

通过SVM模型提供的最佳性能，取得了令人鼓舞的结果，胜过了NN和MR技术，

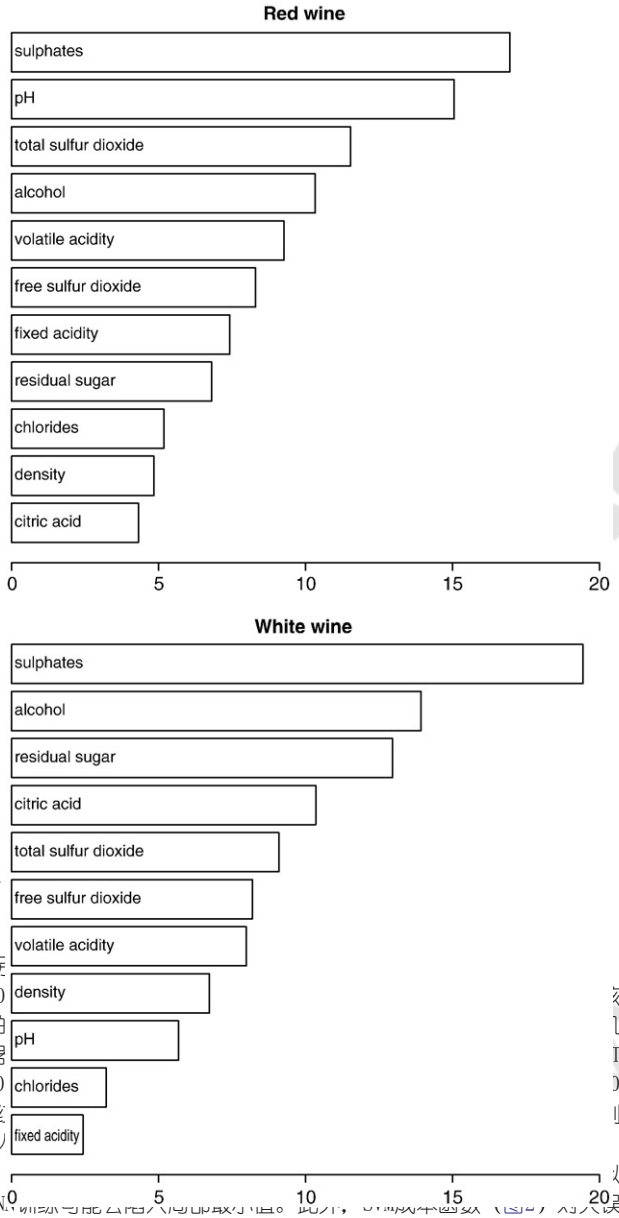


图4.

特定
= 0
意的
类器
1.0
（红
个人

而NN训练可能会陷入局部最小值。此外，SVM成本函数（图2）对大误差给出线性惩罚。相反，NN算法使平方误差之和最小。因此，预计SVM对异常值的敏感性较低，并且这种效果可为较低的错误容忍度带来更高的准确性。正如在[15]，很难以公平的方式比较DM方法，而数据分析师倾向于喜欢他们更了解的模型。我们采用了R工具的默认建议[29]，超参数除外（使用网格搜索设置的超参数除外）。由于默认设置更为常用，因此这似乎是比较合理的假设。然而，如果使用不同的隐藏节点和/或最小化成本函数，则可以实现不同的NN结果。在经过测试的设置下，SVM算法可提供最佳结果，同时需要更多计算。但是，仍然可以在合理的时间内使用当前处理器实现SVM拟合。例如，一次五折交叉验证测试需要大约较大的白色数据集需要26分钟，涵盖了三年的收集期。

这项工作的结果对葡萄酒行业很重要。在认证阶段和根据葡萄牙法律，感官分析必须由人类品尝者进行。但是，评估是基于专家的经验和知识，这容易受到主观因素的影响。所提出的数据驱动方法是基于客观测试的，因此可以将其集成到决策支持系统中，以协助酿酒师的工作速度和质量。例如，仅当专家的评分与DM模型所预测的评分相距甚远时，专家才能重复品尝。实际上，在该范围内，T = 1.0的距离被认为是一种良好的质量控制过程，并且如本研究所示，该公差已达到很高的精度。该模型还可以用于改进对酿酒学学生的培训。此外，输入的相对重要性带来了有关分析测试影响的有趣见解。由于可以在生产过程中控制某些变量，因此可以使用此信息来提高葡萄酒的质量。例如，可以通过在收获前监测葡萄糖浓度来增加或减少酒精浓度。同样，可以通过中止由酵母进行的糖发酵来增加葡萄酒中的残留糖。此外，红酒在苹果酸乳酸发酵过程中产生的挥发性酸取决于乳酸菌的控制活性。另一个有趣的应用是目标营销[24].来自利基市场和/或有利可图的市场（例如，针对特定国家/地区）的特定消费者偏好可以在促销活动期间（例如，在超市进行免费品酒会）进行衡量，并使用类似的DM技术建模，以设计符合这些市场需求的品牌。

致谢

我们要感谢克里斯蒂娜·拉吉多（Cristina Lagido）和匿名审阅者的宝贵意见。FCT项目PTDC / EIA / 64541/2006支持P. Cortez的工作。

参考文献

[1] A. Asuncion，D。Newman，UCI机器学习存储库，加利福尼亚大学尔湾分校，2007年 http://www.ics.uci.edu/~mlearn/MLRepository.html.
[2] J. Bi，K. Bennett，回归误差特征曲线，第20届国际会议论文集。Conf.2003年，美国华盛顿特区，学习机器学习（ICML）。
[3] C. Bishop，“用于模式识别的神经网络”，牛津大学出版社，1995年。
[4] B. Boser，I. Guyon，V. Vapnik，最优边际分类器的训练算法，COLT '92：第五届年度计算学习理论研讨会论文集，美国纽约ACM，1992年，第144-152页。
[5] V. Cherkassy，Y. Ma，“支持向量机参数的实际选择和用于支持向量机回归的噪声估计”，神经网络17（1）（2004）113-126。
[6] P.科尔特斯。RMiner：使用R.的神经网络和支持向量机进行数据挖掘。R。Rajesh（编），高级科学软件和工具箱简介，印刷中。
[7] P. Cortez，M。Portelinha，S。Rodrigues，V。Cadavez，A。Teixeira，通过支持向量机进行羔羊肉质量评估，《神经加工快报》24（1）（2006）41–51。
[8] CVRVV.葡萄牙葡萄酒-Vinho Verde。维戈斯维德斯州葡萄酒委员会（CVRVV），http://www.vinhoverde.pt，2008年7月。
[9] T. Dietterich，用于比较监督分类学习算法的近似统计检验，神经计算10（7）（1998）1895–1923。
[10] S. Ebeler，《风味化学— 30年的发展》，Kluwer学术出版社，1999年，第409-422页，将风味化学与葡萄酒的感官分析联系起来。

[11] 粮农组织，粮农组织统计数据库-粮食及农业组织农业贸易领域统计，2008年7月 http://faostat.fao.org/site/535/DesktopDefault.aspx?PageID=535.
[12] J. Ferrer，A。MacCawley，S。Maturana，S。Toloza，J。Vera，《葡萄酒采摘作业安排的优化方法》，《国际生产经济学杂志》112（2）（2008）985–999。
[13] A. Flexer，《神经网络实验的统计评估：最低要求和当前实践》，第13届欧洲控制论与系统研究会议论文集，第2卷，1996年，奥地利控制论学会，维也纳，奥地利，第1005至1008页。
[14] I. Guyon，A。Elisseeff，变量和特征选择简介，《机器学习研究杂志》3（7–8）（2003）1157–1182。
[15] D.手，分类器技术和进步的幻觉，统计科学21（1）（2006）1-15。
[16] T. Hastie，R。Tibshirani，J。Friedman，《统计学习的要素：数据挖掘，推理和预测》，美国纽约州Springer-Verlag，2001年。
[17] Z. Huang，H。Chen，C。Hsu，W。Chen，S。Wu，支持向量机和神经网络的信用评级分析：市场比较研究，决策支持系统37（4）（2004）543-558。
[18] R. Kewley，M。Embrechts，C。Breneman，“使用神经网络对药品进行虚拟设计的数据条挖掘”，IEEE Transactions on Neural Networks 11（3）（2000年5月）668-679。
[19] M. Kiang，分类方法的比较评估，决策支持系统35（4）（2003）441-454。
[20] A. Legin，A.Rudnitskaya，L.Luvova，Y.Vlasov，C.Natale，A.D'Amico，通过电子舌对意大利葡萄酒的评价：识别，定量分析以及与人类感官知觉的关联，Analytica Chimica Acta 484（1）（2003）33-34。
[21] I. Moreno，D. González-Weller，V. Gutierrez，M. Marino，A. Cameán，a.González，A. Hardisson，使用Probabilistic Neural Networks通过电感耦合等离子体发射光谱法和石墨炉原子吸收光谱法根据其金属含量区分两种Canary DO红酒，Talanta 72（1）（2007）263–268。
[22] R开发核心小组，R：统计计算的语言和环境，R统计计算基金会，维也纳，奥地利，ISBN：3-900051-00-3，2008，http://www.R-project.org.
[23] D. Rumelhart，G。Hinton，R。Williams，在：D. Rulmelhart，J。McClelland（编），通过错误传播学习内部表示，并行分布式处理：认知微观结构的探索，第1卷，麻省理工学院出版社，剑桥，1986年，第318-362页。
[24] M. Shaw，C. Subramaniam，G。Tan，M。Welge，知识管理和市场营销数据挖掘，决策支持系统31（1）（2001）127–137。
[25] D. Smith，R. Margolskee，“品味感”，《科学美国人》，特刊16（3）（2006）84-92。
[26] A. Smola，B。Schölkopf，关于支持向量回归的教程，《统计与计算》14（2004）199–222。
[27] L. Sun，K. Danzer，G. Thiel，通过人工神经网络和判别分析方法对葡萄酒样品进行分类，费森尤斯分析化学杂志359（2）（1997）143–149。
[28] E. Turban，R。Sharda，J。Aronson，D。King，《商业智能，一种管理方法》，Prentice-Hall，2007年。
[29] W. Venables，B。Ripley，《现代应用统计》第S4版，Springer，2003年。
[30] S. Vlassides，J. Ferrier，D. Block，使用历史数据进行生物过程优化：使用人工神经网络和存档的过程信息对葡萄酒特征进行建模，Biotechnology and Bioengineering 73（1）（2001）。
[31] W. Wang，Z。Xu，W。Lu，X。Zhang，高斯核中扩散参数的确定，用于分类和回归，Neurocomputing 55（3）（2003）643-663。
[32] P.韦尔博斯。回归之外：行为科学中用于预测和分析的新工具。博士学位论文，哈佛大学，马萨诸塞州剑桥，1974年。
[33] IH Witten，E。Frank，《数据挖掘：具有Java实现的实用机器学习工具和技术》第二版，Morgan Kaufmann，加利福尼亚州旧金山，2005年。
[34] X.Wu，V.Kumar，J.Quinlan，J.Gosh，Q.Yang，H.Motoda，G.MacLachlan，A.Ng，B.L.P.Yu，Z.Zhou，M.Steinbach，D。Hand，D. Steinberg，数据挖掘的十大算法，知识和信息系统14（1）（2008）1–37。
[35] 于华，林海，徐海，英颖，李宝龙，潘晓波，运用最小二乘支持向量机和近红外光谱技术对黄酒年龄进行酶学参数预测和判别，农业与食品化学56（2）（2008）307-313。
[36] M. Yu，M. Shanker，G. Zhang，M. Hung，“使用神经网络对长距离通信的消费者情境选择进行建模”，决策支持系统44（4）（2008）899-908。

Paulo Cortez拥有Minho大学计算机科学博士学位（2002年）。他是同一所大学信息系统系的讲师，也是Algoritmi中心的研究员，感兴趣的领域包括：商业智能，数据挖掘，神经网络，进化计算和预测。目前，他是《神经处理快报》杂志的副主编，并且参与了7个R&D项目（主要研究人员为2个）。他的研究发表在《启发式杂志》，《决策系统杂志》，《医学人工智能》，《神经计算》，《神经处理快报》等杂志上（见http://www.dsi.uminho.pt/~pcortez).

AntónioCerdeira于1995年毕业于Trás-os-Montese Alto Douro大学，获得酿酒学学位。目前，他负责Vinho Verde地区（CVRVV）葡萄栽培委员会的化学实验室和葡萄学实验。自1997年以来，他是OIV（国际葡萄与葡萄酒组织）葡萄牙酿酒学小组的成员，并且自2000年以来，他是ALABE（葡萄牙酿酒实验室协会）的主席。

Fernando Almeida拥有Minho大学的生物工程学位（2003年）。在2003年至2004年之间，他参加了同一大学生物工程中心的一项物理化学和微生物分析研发项目。自2004年以来，他是CVRVV的感官分析小组的成员，并一直致力于感官测试的认证。

Telmo Matos拥有波尔图大学的应用数学学位（2006年）。他目前在CVRVV的信息系统部门工作。

JoséReis获得了Portucalense大学的信息系统理学硕士学位（2000年），他目前是CVRVV信息系统系主任，并且是IPAM和ISMAI研究所的讲师。他还是Minho大学信息系统系的博士学位学生，研究领域涉及个性化信息系统，市场营销信息系统和数据挖掘。他是《个性化营销和信息技术》一书的作者。