具有可信度分析的卷烟质量评估模型预测控制方法

曾建新[1]，宫会丽[2*]，石 硕[2]，杨 宁[3]

[1.红塔集团工程设备技术部信息网络管理科，云南 玉溪 653100；2.中国海洋大学信息科学与工程学院，青岛 266071；3.青岛海大新星计算机工程中心，青岛 266071]

摘 要：为了改善大多数已建模型在预测时出现盲目的、机械的预测错误情况，以不同产地烤烟和白肋烟数据作为实验样本，综合集成假设检验、凸壳构造与内点分析、序列随机性检验等理论和方法，在预测控制环节设计了具有拒绝识别和可信度分析特征的分类器预测控制算法。实验结果表明，分类器能有效地接受与训练数据相似的测试样本，并给出凸壳内点测试样本的预测值和可信度参考值，同时亦能准确拒绝识别与烤烟质量数据差异较大的白肋烟和特异香型烤烟样本。不同类型测试数据实验验证了该算法的可行性和有效性，尤其是对于以专家经验或领域知识为主的卷烟质量评价问题更加实用。

关键词：卷烟；感官评估；智能技术；支持向量机；可信度分析

中图分类号：S572.099 文章编号：1007-5119（2013）04-0067-05 DOI：10.3969/j.issn.1007-5119.2013.04.014

Predictive Control Method with Credibility in Cigarette Sensory Evaluation

ZENG Jianxin[1], GONG Huili[2*], SHI Shuo[2], YANG Ning[3]

(1. Information Networks Department of Hongta Tobacco Group, Yuxi, Yunnan 653100, China; 2. College of Information Science and Engineering, Ocean University of China, Qingdao 266071, China; 3. New Star Computer Engineering Center of Qingdao Ocean University, Qingdao 266071, China)

Abstract: In order to improve mechanical and blind prediction behavior of some built models, a classifier prediction control algorithm was designed with flue-cured tobacco and burley tobacco in different producing areas as experimental samples. It had the characteristic of rejecting recognition and credibility analysis through integrating several theories and methods including hypothesis testing, convex hull, interior point analysis and sequence random testing. The results demonstrated that classifier could effectively accept test sample set and give predictive values and reliability reference values of test data in convex hull. In the meanwhile, classifier could also accurately reject burley tobacco sample and special type flue-cured sample, which was different from flue-cured sample set. The feasibility and validity of classifier were verified through different type of testing data, especially the practicality of cigarette sensory evaluation was based on expert experience or domain knowledge.

Keywords: cigarette; sensory evaluation; intelligent technology; support vector machine; credibility analysis

计算机辅助卷烟质量评估已引起了国内外研究人员的关注，主要研究方法有遗传算法和 RBF 神经网络、BP 网络、Kohonen 网络、模糊逻辑、灰色聚类、支撑向量机等[1-5]，但大部分学者开展计算机质量评估研究目标在构建性能良好的分类器模型，但在进行工业领域的建模与预测应用时经常存在如下问题：模型的使用者不了解模型的构建背景，测试样本与训练样本存在着较大的统计分布特征差异，测试样本的类别未包括在训练样本中。通常大多数研究人员把精力放在了如何构建具有良好分类性能的学习模型上，忽视了预测环节对实际应用的影响和价值。面对上述问题，本研究拟通过对云南、福建、湖南、贵州、巴西等地烤烟和白肋烟的化学成分、感官品评数据进行支持向量机（Support Vector Machine，SVM）建模分析，并对已经构建好的分类器模型进行具有拒绝识别和可信度分析的预测控制[6-7]，旨在解决目前分类器在预测中所出现的一些盲目识别错误问题，提高模型预

作者简介：曾建新，男，高级工程师，主要研究方向为烟草信息化和商业智能。E-mail：zjx@hongta.com。*通信作者，E-mail：huiligong@163.com

收稿日期：2011-07-22 修回日期：2012-02-09

测准确度和实用性。

1 实验数据与建模方法

1.1 实验数据

收集 2004—2008 年的云南、福建、湖南、贵州以及山东 5 大产地代表等级为 B2F、C2L、C3F、X2F 的烤烟样品共 260 份，收集 2007—2009 年大理、巴西的白肋烟数据 10 份分别作为实验数据和测试数据，具体如表 1。

表 1 实验数据

Table 1 The test data

类型	省或国家	市（州）	代表等级	数量
烤烟	云南	楚雄、大理、玉溪	B2F、C2L	56
烤烟	湖南	长沙、郴州、永州	C3F、X2F	90
烤烟	贵州	贵阳、同仁、遵义	B2F、X2F	60
烤烟	山东	临沂、潍坊	C1L/S、BC2F	54
白肋烟	云南	大理	C1L/S、BC2F	5
白肋烟	巴西			5
总计				260

白肋烟样品化学成分均值和取值范围与烤烟有很大差异，如烤烟总糖均值为 22 mg/g 左右、烟碱均值在 3.7 mg/g；而白肋烟总糖均值在 0.5 mg/g 左右，烟碱为 2.6 mg/g。烤烟和白肋烟的化学成分统计特征差异很大。

1.2 化学成分检测

采用近红外光谱（NIR）定量分析技术对烟叶样品的主要化学成分进行分析，包括总烟碱、总糖、还原糖、总氮、钾、氯等含量，计算蛋白质、施木克值、钾氯比、糖碱比等。

1.3 感官质量评价

卷烟样品的感官质量评价按照香型、劲头、浓度、香气质、香气量、余味、杂气、刺激性等项目进行评价与打分。香型指标按照企业制定的卷烟感官质量评价标准将其划分为 1~8 分制等级，依次为特异香型（1）、清香型（2）、清偏中（3）、中偏清（4）、中间型（5）、中偏浓（6）、浓偏中（7）、浓香型（8）。表 2 中的 10 条特异香型数据则为香型指标分值为 1 的样本数据。

表 2 训练与测试数据

Table 2 Training and testing data

用途	类别	说明	样本数
训练	烤烟	清香型、中间香型烟叶	240
测试 a	白肋烟	与烤烟质量数据差异非常大	10
测试 b	特异香型烤烟	与烤烟数据的统计特征略有差异	10
测试 c	烤烟	与训练样本具有相似的统计特征	10

注：（1）化学成分包括总糖、总氮、总烟碱、钾、氯等 10 项化学成分；（2）建模输出中的香型指标可量化为 8 档：清香、清偏中、中偏清、中间香、中偏浓、浓偏中、浓香、特异香型；（3）参与测试所有数据均不包含在训练样本数据中。

1.4 SVM 建模方法

针对企业数据采集的难度和费用较大，样本数据量相对较小的情况，本研究采用回归函数估计的 SVM 方法[8]解决感官质量评估的多类模式识别问题。通过对结构风险最小化原理来提高泛化能力，实现实数样本函数逼近，构造具有优良推广性能、抗噪性、鲁棒性的学习模型。建模时，从表 2 中的 260 条烤烟数据中随机选取其中具有清香型、中间香型以及浓香型特征的 240 条样本作为训练数据，无需任何数据预处理，直接输入 SVM 分类器进行训练，并保存已建好的香型模型，其余 20 条作为测试样本。

2 具有可信度分析的模型预测控制方法

2.1 假设检验

首先，需要对测试样本和训练样本中每个属性进行独立同分布假设检验，以此判断 2 个样本是否来自于同一个总体，或是来自于同一个总体且具有相近的统计特征。对于卷烟质量数据，每个指标均进行正态分布检验，以总糖指标为例进行正态分布分析，图 1 卷烟质量总糖指标的正态分布特征分析，各数据点基本与对角线（理论值）重合；图 2 为残差分布图或称去势正态分布 P-P 图，各实际值与理论值之差都在 0.05 以内，这也说明了烟叶质量数据中总糖指标的正态分布特性，其它指标亦如此。对于服从正态分布烟叶数据特征，故采用均值的双样本 t 检验，若有任何一个变量拒绝原假设，则认为不是来自同一个总体，分类器建议拒绝识别测试样本。如果测试样本通过假设检验，则进行 2.2 中的

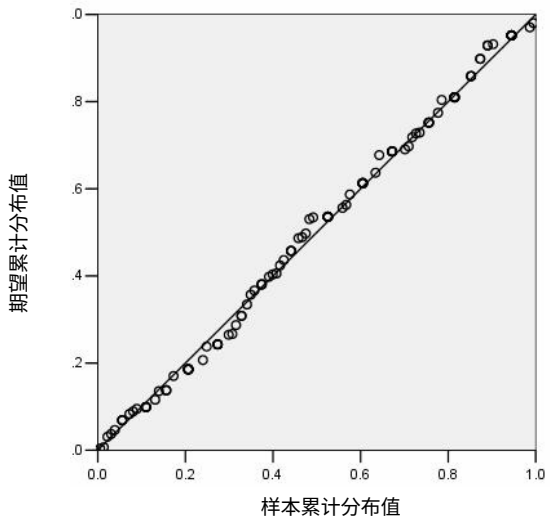

图 1　总糖正态 P-P 分析图

Fig. 1　Normal P-P diagram of total sugar

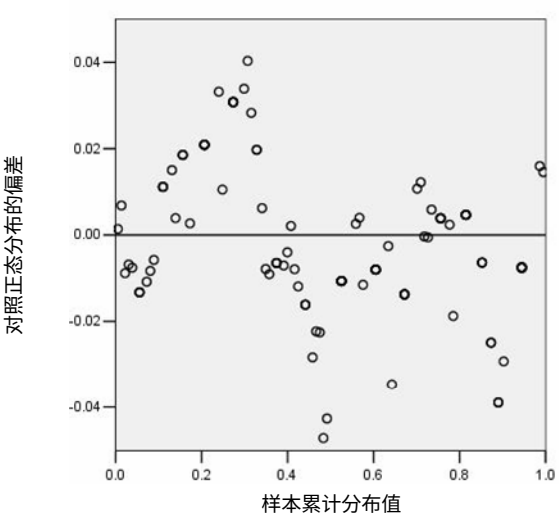

图 2　总糖正态分布残差图

Fig. 2　Total sugar normal distribution of residual plot

凸壳构造和可视化过程。基于近红外检测数据质量相对较差的现状，该检验的显著性水平设为α＝0.1。

2.2　凸壳构造与可视化

凸壳是计算几何中最普遍、最基本的一种结构，自身有许多优良特性[9]，已在模式识别、图像处理和人工智能领域得到了比较广泛的应用，许多复杂问题可归纳成凸壳问题求解。本研究运用凸壳构造与凸壳内点判别，将训练样本数据以直观、简单、可视化方式展示出来，并将专家经验和领域知识有效融入分类器预测中。该方法对解决卷烟质量预测这类工程问题具有很高的应用价值。但卷烟质量数据的高维、非线性特性，导致在高维空间进行可视化难以实现，需要采用流形学习方法进行低维流形变换[10-11]，将样本输入空间压缩到 2 维空间，从而构造训练样本的各个凸壳并可视化。整个可视化过程比传统“黑箱子”操作更易于让技术人员接受与理解。

如果测试样本通过假设检验，则对训练样本按其类别进行多类别凸壳构造，形成凸壳集 $\{CH(\omega_1),CH(\omega_2),...,CH(\omega_c)\}$，同时对于所有训练样本构造一个最外层凸壳 $CH(\omega_{c+1})$，$\omega_i, c=1,2,...,c$，c 为训练集中类别个数。当测试样本点 x_i 进入分类器时，采用 Quickhull 算法[12]执行凸壳内点判定。

（1）如果 x_i 不是 $CH(\omega_{c+1})$ 凸壳的内点，则拒绝识别。

（2）如果 x_i 是凸壳 $CH(\omega_{c+1})$ 的内点但不是 $\{CH(\omega_1),CH(\omega_2),...,CH(\omega_c)\}$ 中任意凸壳的内点，则交由分类器预测获得预测类别 y_{ω_c}，分类器此时输出预测结果为“邻近 y_{ω_c} 类别”。

（3）如果 x_i 是凸壳集 $\{CH(\omega_1),CH(\omega_2),...,CH(\omega_c)\}$ 中某一凸壳内点，则由分类器预测输出，同时进行下面的可信度分析。

通过可视化方法可将上述凸壳集合直接绘制出来。当新来一个测试样本点时，可首先在凸壳集上进行描点，直观展示该测试样本点是否被包含在训练样本集所形成的凸壳中。

2.3　可信度分析

本文采用非一致性度量函数和序列随机性检验 p 值函数对分类器预测输出结果进行可信度分析。当一条测试样本 x_{n+1} 附加上某一个类标签 y_{w_c} 后，与训练样本共同构成检验样本序列 $Z^{(n+1)}=\{z_1,z_2,...,z_n,z_{n+1}\}$，对 z_{n+1} 在 $Z^{(n+1)}$ 检验序列中的随机性 p 值进行计算，p 值的计算如下：

$$p=\frac{\left|(i=1,\ldots,n+1:a_i\geq a_{n+1}\right|}{n+1}$$

上式中 i 指在非一致性度量值序列 $a_1,a_2,\cdots a_n,a_{n+1}$ 中大于 a_{n+1} 的度量值个数，每个样本的非一致性度量值 a_i 是通过构造样本的非一致性度量函数来获得，计算公式如下：

$$a_i=\frac{\sum_{j=1}^{k}d_{ij}^{+}}{\sum_{j=1}^{k}d_{ij}^{-}},i=1,2,\ldots,n+1$$

其中 $\sum_{j=1}^{k}d_{ij}^{+}$ 表示与 x_i 样本同类别的 K 个近邻距离的和，$\sum_{j=1}^{k}d_{ij}^{-}$ 表示与 x_i 样本不同类别的 K 个近邻距离的和。该 p 值描述了当前检验样本序列符合独立同分布的相似程度，相似度越高则说明该测试样本隶属于 y_{w_c} 类别的可能性越大。可信度定义为：C = p×100（%）。

3 结 果

将测试数据 a、b、c 分别进入 1.4 中所建立的香型模型，并运用具有拒绝识别和可信度分析方法对分类器进行预测，实验结果如下。

3.1 对于测试数据 a

对于香型指标，分类器全部拒绝识别 10 条白肋烟测试数据。测试样本数据中 10 个测试变量有 9 个 *P* 值（显著性水平）<0.05，可认为测试样本与训练样本的均值或分布存在显著性差异，分类器拒绝识别该批测试数据。该预测结果验证了行业专家知识：白肋烟和烤烟在化学成分、评吸质量上有非常明显差异。

3.2 对于测试数据 b

对于香型指标，测试数据所采用的 10 条特异香型样本数据在假设检验中未发现两样本总体分布存在显著性差异，因此进入预测过程的凸壳内点分析环节。在凸壳内点分析时发现，有 8 个测试样本点为凸壳 $CH(\omega_{c+1})$ 的外点，因此拒绝识别 8 条测试样本。

下面图 3 是训练样本的凸壳构造情况及测试样本的内点判别。

在图 3 中，黑色圆点为测试样本，2 个为凸壳 $\mathrm{CH}(\omega_{c+1})$ 的外点，3 个为 $\mathrm{CH}(\omega_{c+1})$ 的内点，其中 1 个为所有类别凸壳的外点。

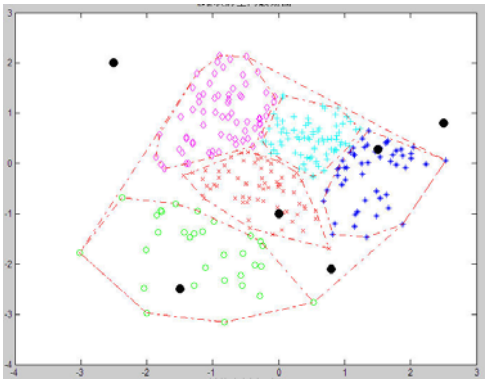

图 3 凸壳内点判别示意图

Fig. 3 The convex hull graph of internal point determination

3.3 对于测试数据 c

对于香型指标，分类器做出测试数据 c 的预测，并给出每个测试样本点的预测可信度参考值。10 条测试数据中，有 1 个测试样本被判为外点，拒绝识别。有 3 个测试样本点未被判断为 5 个类别凸壳的内点，由分类器预测输出类别。其他 6 个测试样本点预测输出的类别，并计算测试样本点的可信度值。3 条外部测试点在分类器预测输出结果见表 3。从表 3 可看出，被类别凸壳集判定为外点的测试样本，SVM 大多也没有正确分类，只有 1 个样本是预测准确的。

6 个内点输出结果及可信度参考值见表 4，在表 4 中，其中 4 个测试样本点预测准确，正确分类率为 66.67%，对于卷烟香型这一比较难预测的指标在 10 项化学成分作为输入的情况达到如此性能已经算是理想结果。这种性能的提升不仅是分类器的功劳，也是预测控制环节减少了错误样本输入带来的误差。

本研究提出的具有拒绝识别和可信度分析的

表 3　三条外部测试点在分类器预测输出结果

Table 3　Output results of 3 exterior test points in SVM classifier

序号	SVM 预测类别	实际类别	误差	预测输出
1	5	3	2	邻近第 5 类
2	3	2	1	邻近第 3 类
3	6	6	0	邻近第 6 类

表 4　SVM 分类器预测 6 个内点输出结果及可信度参考值

Table 4　Output results of six internal points in SVM classifier

序号	预测类别	实际类别	误差	p 值	可信度参考值/%
1	5	5	0	0.9537	95.37
2	5	5	0	0.8750	87.50
3	4	5	1	0.6204	62.04
4	3	3	0	0.7231	72.31
5	5	5	0	0.5972	59.72
6	4	2	2	0.7222	72.22

预测控制算法使得分类器能够对与训练样本特征差异较大的测试样本实现拒绝识别，避免了分类器在预测环节出现的一些低级错误。同时，对预测输出类别所做的可信度分析，也为用户理解和把握输出结果提供了一定的参考。

4　小　结

针对当前主流分类器预测时出现盲目的、机械的预测错误情况，本研究综合集成假设检验、凸壳构造与内点分析、序列随机性检验等理论和方法，设计了具有拒绝识别和可信度分析特征的分类器预测控制算法，并在已建立的香型模型上运用不同类型测试数据实验验证了该方法对降低预测错误率的实际效果。研究表明，有效、可控的模型预测控制方法，对解决卷烟或者其它行业感官评估类工程问题具有很高的应用价值。

参考文献

[1] 殷勇，吴守一. 基于遗传算法的卷烟质量评定神经网络模型[J]. 农业机械学报，1999. 30（3）：71-75.

[2] 李东亮，许自成. 基于化学成分的烟草质量评判方法研究与应用[D]. 郑州：河南农业大学，2008.

[3] 冯天瑾，丁香乾，杨宁，等. 计算智能与科学配方[M]. 北京：科学出版社，2008：58-116.

[4] 丁香乾，曹均阔，贺英. BP 网络与 Kohonen 网络的集成应用与研究[J]. 中国海洋大学学报，2003（4）：615-620.

[5] 冯天瑾，林丽莉，周文晖，等. 基于神经－模糊方法的单料烟感官质量评价专家系统[J]. 青岛海洋大学学报：自然科学版，2001，31(06)：931-936.

[6] Vapnik V. The nature of statistical learning theory[M]. New York: springer-verlag, 1995.

[7] Vapnik V. An overview of statistical learning theory [J]. IEEE Trans on Neural Network, 1999, 10(5): 988-999

[8] 杨宁. 支持向量机在感官评估中的应用研究[D]. 青岛：中国海洋大学，2004.

[9] 周培德. 计算几何—算法分析与设计[M]. 北京：清华大学出版社，2000：57-87.

[10] Bregler C, Omohundro S M. Nonlinear manifold learning for visual speech recognition[C]//In Proc. of 5th International Conference on Computer Vision, 1995：494-499.

[11] Jenkins C, Mataric M J. Deriving action and behavior primitives from human motion data[C]//In the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2002), Switzerland, 2002: 2551-2556.

[12] Bradford Barber, David P D, Huhdanpaa H. The quickhull algorithm for convex hulls[J]. ACM, Transactions on Mathematical Software, 1996. 22(4): 469-483.