基于高光谱成像的香肠菌落总数回归预测及数据可视化

董小栋，郭培源，徐盼

（北京工商大学计算机与信息工程学院，北京 100048）

摘要：香肠的好坏有很多种评价指标，菌落总数（TVC）是其中的一种。高光谱成像技术已经成为一种快速、无损检测食品品质的有效方法。本文利用高光谱成像技术对香肠的菌落总数进行了定量分析，对数据进行了主成分分析（PCA），研究发现数据集中前四个主成分累计贡献率已达 97.65%，已经可以反映出香肠所包含的绝大部分信息。对前四个主成分对应的优化区间采用高斯核函数的 SVM 回归模型进行预测，并为了提高回归预测模型的精确度，对模型的 *c*，*g* 参数，进行了遗传算法（GA）、网格搜索算法和粒子群算法（PSO）寻优对比，其中 PSO 寻优可使回归预测值和真实值的相关系数为 0.9777，交互验证均方根误差为 0.0823，能够准确快速的实现香肠菌落总数的预测。除此之外，利用 python 对回归预测的数据进行可视化，更加直观的显示菌落总数变化，且可以达到实时观看的效果。

关键词：香肠；菌落总数；高光谱成像；SVM；可视化

文章篇号：1673-9078(2017)7-308-314 DOI: 10.13982/j.mfst.1673-9078.2017.7.043

The Prediction of the Total Viable Count on Sausage Based with Hyperspectral Imaging Technique and Data Visualization

DONG Xiao-dong, GUO Pei-yuan, XU Pan

(Beijing Technology and Business University School of Computer and Information Engineering, Beijing 100048, China)

Abstract: There are a lot of evaluation standard of the quality for sausage, one of which is the total viable count(TVC). Hyperspectral imaging technique has become an effective method to detect food rapidly and nondestructively. In this paper, Hyperspectral imaging technique has carried on the quantitative analysis to the total viable count (TVC) on the sausage. The data sets of sausage were assessed using the PCA method, and then the study found that the contribution rate of the first four principal component reaches 97.65% which can reflect the most information of the sausage. The SVM regression model based on Gaussian kernel function and the optimal interval the first four principal components is used to forecast TVC. In order to improve the accuracy of the regression model, the genetic algorithm (GA), grid search algorithm and particle swarm optimization (PSO) are compared to get the c and g parameters of the model. The correlation coefficient of regression prediction value and real value is 0.9777, and the root mean square error of interactive verification of PSO is 0.0823, which can accurately and quickly predict the TVC. Besides, Use python to realize visualization of regression prediction data which can show the change of TVC more intuitively and can achieve real-time watching.

Key words: Sausages; total viable count; Hyperspectral Imaging Technique; SVM; data visualization

随着社会的不断进步，人们对食品安全的要求不断提高。

众所周知，谈菌色变，主要是因为人们对于香肠中的菌落总数没有直观地了解。在中国，香肠文化历史悠久，以其独特的色、香、味，闻名于世[1]。然而，在香肠制作的过程中，因为其生产环境条件不合格，会导致香肠的微生物大量残留，以及外来菌的污染，致使香肠品质根本无法满足国家标准。细菌总数是反映肉品被污染和腐败状况的重要指标。目前的检测方法主要是依靠对香肠进行理化实验和辅以感官等传统的检测方法，效率低，花费大量时间，不仅成本高，还无法达到食品无损检测的目的。因而寻找快速、无损的肉制品品质检测方法变得非常迫切[2,3]。

收稿日期：2016-10-11

基金项目：国家自然科学基金项目（61473009）；北京市自然科学基金项目（4122020）

作者简介：董小栋（1991-），男，硕士，主要从事高光谱图像与机器学习等研究

通讯作者：郭培源（1958-），男，博士，教授，主要从事高光谱成像与食品检测等研究

高光谱图像集光电子学、计算机科学于一身，真正实现了图谱合一，既包含物质内部的光谱信息，也包含物质外部的图像信息，具有更高的精确度[4~6]。利用高光谱成像技术对食品进行无损检测已经成为一种新型的方式，并且符合实际所需。田潇瑜研究对比了不同预处理方法对各指标的预测效果，并通过建立全波段偏最小二乘回归(PLSR)以及联合区间偏最小二乘回归(si-PLSR)、遗传算法-偏最小二乘回归(GA-PLSR)模型，最后得到剪切力值、a*的最佳预测模型为 si-PLSR 模型，预测相关系数和标准差分别达到 0.9085、0.9027 和 7.5212、1.4878，模型 RPD 值为 2.16 和 2.65[7]。王伟等在综合比较偏最小二乘回归(PLSR)、人工神经网络(ANNs)和最小二乘的支持向量机(LS-SVM)三种建模方法的基础上，验证了高光谱成像技术结合 LS-SVM 预测建模方法可作为快速、非破坏预测生鲜猪肉 TVC 的有效手段[8]。

数据可视化技术的基本思想是，数据集中的每一个数据按照特定的设置都可以成为一个图元原素，数据集中的数据存在着联系，而数据与图元元素也存在着某一映射关系，这样由大量图元元素构成一幅数据图像，其主要是通过图像含义向用户清晰、高效地传递数据中包含的信息，在进行可视化设计时需要优先考虑的是实现信息传递。周志光研究了在最大标量差累积法的启发下，提出一种有效展示隐藏特征的直接体绘制方法[9]。

本试验以香肠为研究对象，不仅对 SVM 回归预测模型方法的三种参数优化方法进行了比较，还对菌落总数回归预测的值实现了可视化。在高光谱成像系统 400 nm~1000 nm 波长范围进行，采集样品的图像和光谱信息，使用 PCA 确定最佳建模特征波长，并利用 SVM 建立预测模型加以分析比较，提出了一种高光谱成像技术结合数据可视化对香肠的菌落总数进行预测研究方法，实现香肠菌落总数的质量安全检测。

1 数据样本

以广味香肠为实验对象，设置样本数为 70 个，所有的样本均处于同一环境条件，即温度为 40 ℃、空气混浊。每隔 8 h 随机的取 2 个样本进行高光谱数据的采集，两个样本分为训练集样本和预测及样本，直到训练集样本为 40 组，预测及样本为 30 组为止。

1.1 仪器

高光谱成像数据采用北京安洲科技有限公司的 SOC710VP 便携高光谱成像光谱仪采集，实验仪器参数设置如表 1 所示。

表 1 SOC710VP 高光谱分析仪系统参数设置

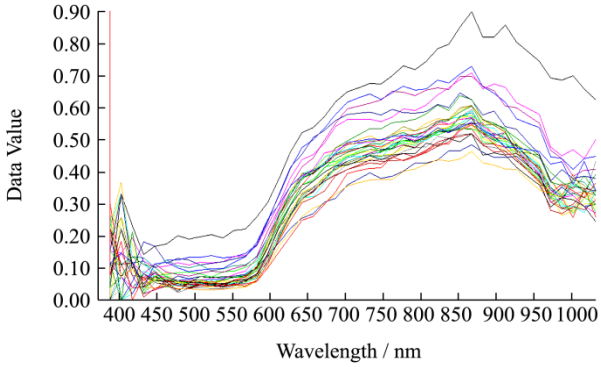
Table 1 the parameter settings of spectrum analyzer system of SOC710VP

名称	数量	说明
SOC710VP 高光谱分析仪	1 台	技术参数：1.光谱扫描范围：400~1000 nm；2.光谱分辨率：4.6875 nm；3.采样间隔：2 nm；4.光谱通道数：128；5.测定速度：30 行/s 主要特点：1.快速准确；2.不需要任何化学试剂及特殊设备

1.2 光谱值的获取和菌落总数的检测

1.2.1 高光谱图像采集

在进行高光谱图像获取之前，需要对高光谱图像进行黑白板校正。



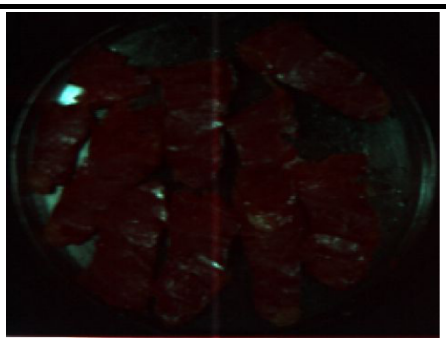


图 1 香肠样品原始光谱图

Fig.1 The original spectral on bacon

校正后，高光谱图像数据分析采用 ENVI 软件平台。由于高光谱数据量庞大，在选取数据时，应该选取表面一个感兴趣区域（region of interest，ROI）并计算平均反射光谱，样本香肠的形状为椭圆形，所以在

选取 ROI 时选择自定义功能，根据样本的形状选择区域，避免数据的缺失。获得样本的全波段原始反射光谱曲线如图 1 所示。

1.2.2 样本菌落总数测定

上述每次获得高光谱数据后立即进行样本的菌落总数的测定，执行食品安全国家标准食品微生物学检验菌落总数测定（GB 4789.2），对样本经过一定的处理，在一定条件下培养后活得 1 mL/g 检样，即单位面积样品中所含菌落的总数，每个样本为 30 g。

2 支持向量机

2.1 SVM 概述

支持向量机(SVM)是 Vapnik 等人提出的一种新型的机器学习的方法。它能解决神经网络不能解决的过学习问题，能较好的解决小样本高位数、局部最小点和非线性等实际问题。基本思想是通过一个非线性映射ø，将数据 x 映射到高维特征空间 F，并在这个空间进行线性回归。

已知一个训练数据集，

$\{(x_i, y_i)|x_i \in R^n, y_i \in R, i=1,2\ldots l\}$ （1）

以及假设函数集，

$F=\{f|f=\omega^*x+b\}$ （2）

其中权值 $\omega \in R^n$，而 $b \in R$ 为阈值。回归支持向量机一般可以表示为下面的规划问题：

$$\min = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}[c(\xi_i)+c(\xi_i^*)] \quad (3)$$

其中约束条件为：

$W^*xi+b-y_i \le \varepsilon+\xi_i$ （4）

$$y_i - w^*x_i - b \le \varepsilon + \xi_i^* \quad (5)$$

$$\xi_i \ge 0,\ \xi_i^* \ge 0,\ i=1,2...l \quad (6)$$

其中，ξ_i 和 ξ_i^* 为松弛变量，c（ξ）是损失函数，C 为惩罚系数。本文使用的核函数如下：

高斯核：$k=(x,\ x_i)=\exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)$，$\sigma$为宽度函数

将采集的40组高光谱数据作为SVM的输入样本，已检测得到的菌落总数为 SVM 进行回归预测的输出样本。

3 主成分分析(PCA)

3.1 主成分分析法概述

PCA 是一种简化数据集的技术。不仅仅是对高维数据进行降维，更重要的是经过降维去除了噪声，发现了数据中的模式。PCA 把原先的 n 个特征用数目更少的 m 个特征替代，新特征是旧特征的线性组合[10]，这些线性组合最大化样本方差，尽量使新的 m 个特征互不相关。从旧特征到新特征的映射捕获数据中的固有变异性。

3.2 主成分分析法的设计步骤

（1）对数据阵进行标准；

（2）求出相关矩阵；

（3）求出 R 的特征知系特征向量；

（4）求出主成分；

（5）将求出的特征值按大小依次排列，设置主成分比例，并依次排列特征向量，就可以得出我们需要的主成分。

3.3 PCA 和 SVM 的高光谱预测模型

本研究然后对香肠建立 PCA-SVM 回归模型：

（1）将载入 70 组的香肠光谱信息进行多远散射矫正后利用 PCA 算法对其进行特征提取及降维处理。

（2）将上述处理后的 40 组特征光谱，也就是训练集样本，输入到 SVM 回归预测模型，进行 model 模型训练。

（3）采用 30 组测试集样本对 model 进行测试，并采用相关系数 R 和均方根误差 MSE 作为模型的 2 个评价指标，R 值越接近 1，MSE 值越小，所建模型性能就越好。

4 RBF 核函数的 SVM 模型 c 和 g 参数寻优

4.1 网格寻优

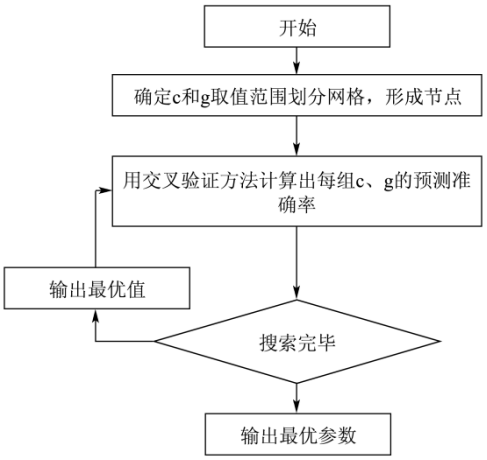


图 2 网格寻优算法

Fig.2 The flow chart of grid search algorithm

所谓的网格搜索就是尝试各种可能的(c，g)对值，然后进行交叉验证，找出使交叉验证精确度最高的(c，g)对。设计思路如图 2 所示。

4.2 GA 寻优

虽然采用网格搜索能够找到在 CV 意义下的最高的分类准确率，即全局最优解，但有时候如果想在更大的范围内寻找最佳的参数 c 和 g 会很费时，采用启发式算法就可以不必遍历网格内的所有的参数点，也能找到全局最优解。设计思路如图 3 所示。

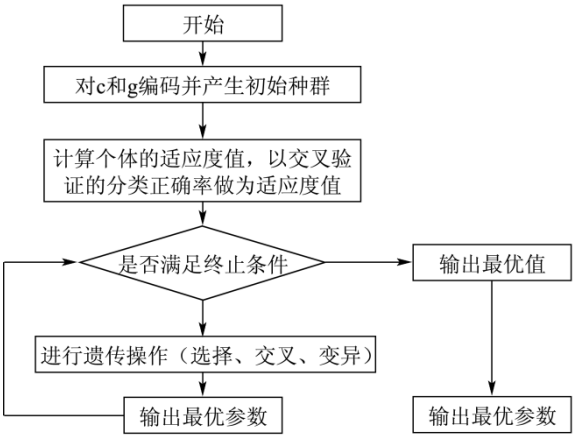


图 3 GA 寻优算法

Fig.3 The flow chart of genetic algorithm

4.3 PSO 寻优

PSO 不必遍历解空间的所有点，算法运行速度快、效率高，而且样本数据为平稳、单峰序列时算法的优化精度也非常高。目前已广泛应用到函数优化，神经网络训练。模糊系统控制领域。设计思路如图 4 所示。

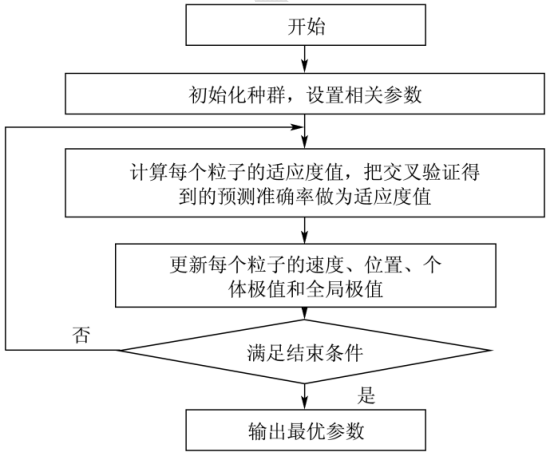


图 4 GA 寻优算法

Fig.4 The flow chart of particle swarm optimization

5 结果与分析

5.1 主成分分析

若使用全波段进行 SVM 回归预测建模，因为数据的高维度和复杂度，不仅计算量多而且冗余的信息也比较多。采用 PCA 不仅可以实现数据的降维，还能保证原始信息的完整性，可以有效的去掉信息中相关性较高的信息。如图 5 所示是经过 PCA 主成分分析后的结果图。

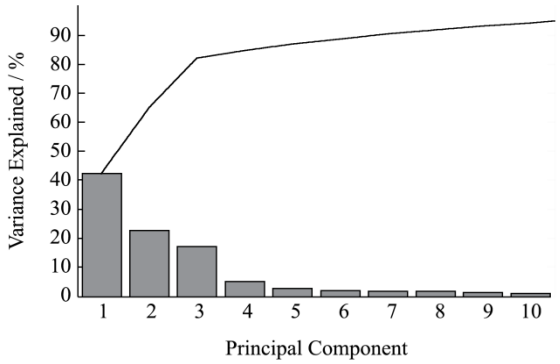


图 5 PCA 主成分分析

Fig.5 principal components analysis

经过分析，前四个主成分累计贡献率已达 97.65%，已经可以反映出香肠的绝大部分所包含信息，所以在接下来的 SVM 回归预测模型中会选择前四个 PCA 主成分作为特征波段区域来建模。

5.2 光谱数据预处理

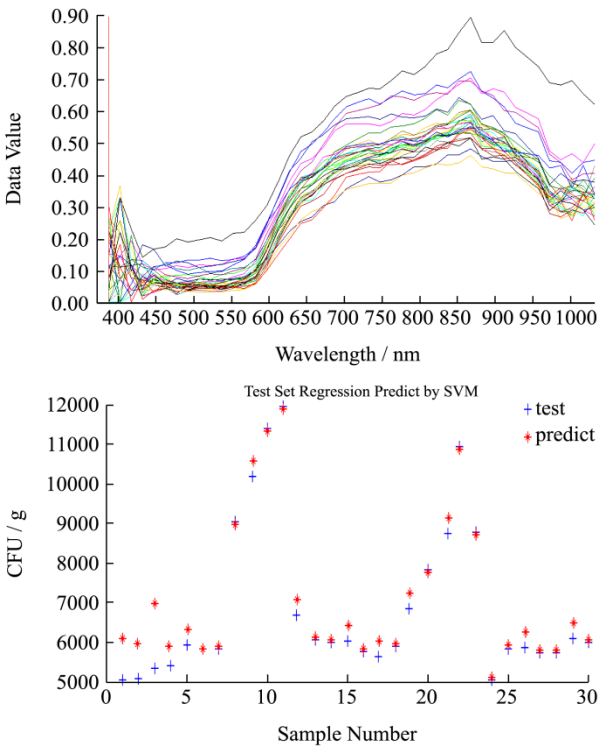


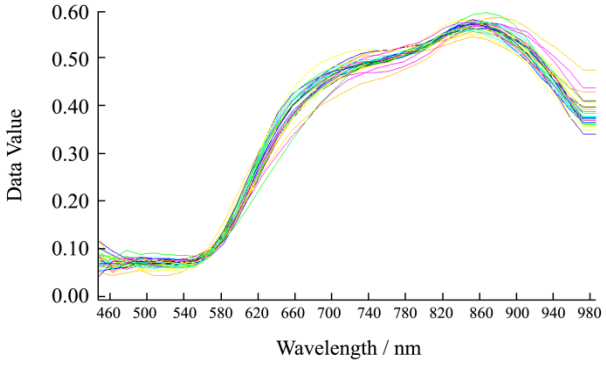
图 6 原高光谱数据图和预测结果图

Fig.6 The original hyperspectral data and predicted results

高光谱采集到的香肠样品原始光谱虽然含有丰富的待测成分的有用信息，但是波段数目多，各波段间具有较强的相关性，因此通过主成分分析(PCA)方法对高光谱数据进行预处理，既能很好的达到降维的目的，同时也去除了噪声波段，这些噪声不仅会对光谱信息产生干扰，还可能会导致光谱曲线发生基线漂移，从而影响到模型的预测有效性。结合光谱预处理地方法有力保证了样品成分的精确分析度和模型的预测能力。

对于样本表面颗粒分布不均匀或者大小不一等情况造成高光谱数据不同程度散射的情况，采用多远散射校正有可有效地消除散射的影响。图 6 是采用原光谱进行建模的结果。

采用上述预处理方法对采集到的香肠校正集样品的高光谱数据进行预处理，其光谱图如图 7 所示。



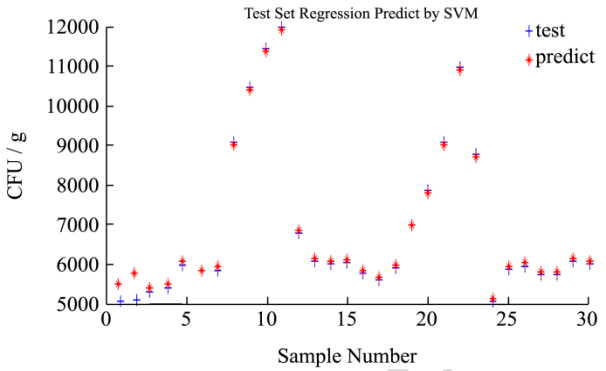


图 7 预处理后的高光谱图

Fig.7 The pretreatment of spectrum

由上述分析可知，经过预处理后可以提高模型预测的精度。

5.3 参数 c、g 优化

本文选用 RBF 核函数，所以在核函数参数的选择上有惩罚项常数 c 和 RBF 的参数 g。目前参数寻优方法有网格寻优[11]，GA 寻优[12]，PSO 寻优[13]等，其中最简单最有效的方法是网格寻优。如表 2 所示是不同参数寻优方法中参数 c，g 的数值以及，经过 PCA+SVM 回归预测模型后的 *MSE* 和 *R* 值的结果如表 2 所示。

由表 2 可得知，采用 PSO 粒子群算法对香肠菌落总数分析模型最优，得到测试集样本的香肠含量的真实值与预测值的散点信息如图 8 所示。

表 2 不同参数寻优方法的结果

Table 2 The results of different parameters optimization method

寻优方法	Best c	Best g	MSE	R
网格寻优	0.0825	337.7940	2.1650	0.9632
GA 遗传算法	1.1635	43.0110	0.0856	0.9653
PSO 粒子群算法	2.4107	60.2924	0.0823	0.9777

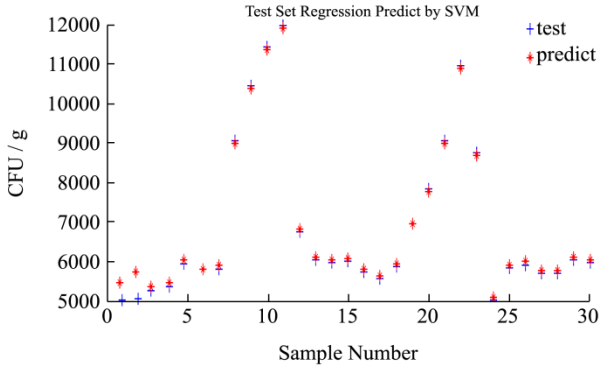


图 8 香肠菌落总数实际值与预测值相关性散点图

Fig.8 The scatter plot of the actual and predicted value on TVC

由模型的预测结果可得，用 PSO 参数寻优的 PCA-SVM 建模方法建立的香肠表面菌落总数含量检测模型进行检测的结果与真实结果的相关系数为 0.9777，交互验证均方根误差为 0.0823，能够准确快速的实现香肠菌落总数的预测。

5.4 数据可视化

可视化技术能够将人的大脑和计算机两个强大的信息处理结合起来，而生成有效的可视界面更能方便我们与之交互，便于我们观察、理解、研究，探索大规模信息数据，可以使我们有效地发现隐藏在抽象信息内部的某些特征和规律[14]。

虽然回归预测得到了菌落总数，但是人们对菌落总数并没有一个特别清晰的概念，这就要涉及到对菌落总数的渲染问题。本文借助 python 数据可视化[15]，实现人与数据之间的图像通信，可以更好的表示菌落总数的变化，将预测得到的菌落总数进行时间序列的

可视化显示。

Numpy 是 python 中的一款高性能科学计算和数据分析的基础包。这种工具可用来存储和处理大型矩阵。Matplotlib 是 python 著名的绘图库，它提供了一整套和 MATLAB 相似的命令 API，十分适合交互式地进行制图，经常运用到数据可视化中。

首先获取一块香肠的图片，并对其进行反色处理，如图 9 所示。

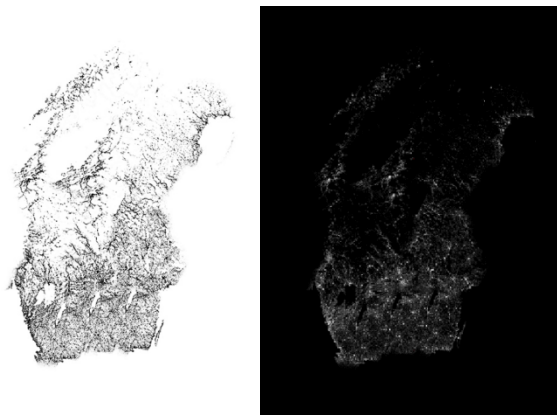


图 9 香肠图片及反色处理后的图片

Fig.9 sausage images and color images after processing

获取图像的像素大小，也就是图像的点，样本为二维 Numpy 阵列，以此创建一个离散密度图，在每个点图上的点应放在与它的颜色对应的值，使用颜色映射表决定每个像素的颜色。通过 SVM 预测后通过 model 中的参数可以获得决策函数：

对于给定的样本数据集$\{(x_i,y_i)|i=1,2,3\ldots,k\}$，其中$x_i$为数据的输入值$y_i$为数据的输出值，

$$g(x，\ z_i)=\exp(-\frac{\|x-z_i\|^2}{2\sigma^2}) \quad (7)$$

其中 g(x，z_i)为 SVM 回归预测用到的核函数，即 RBF 函数。

$$f(x，\ y)=\sum_{i=1}^{n} w_i g(x，\ z)+b \quad (8)$$

其中 f(x，y)作为颜色表的关系。

将样本置于温度为 40 ℃、空气潮湿的环境下，每隔 8 h 后，使用高光谱成像仪获得该样本的光谱信息，并使用 PCA+SVM 回归预测模型进行菌落总数的预测，将预测的结果存入矩阵，创建颜色图表，得到菌落总数的动态变化效果，如下图是部分时间段的菌落总数图。如图 10 所示。

从图 10 中可以看到，使用回归预测的菌落总数值经计算机重构成像，足够的数据量可以非常直观、清晰的在任意时间段，看到香肠菌落总数的变化情况，将人眼无法观测的的菌落总数通过数据的可视化实现菌落的可视。

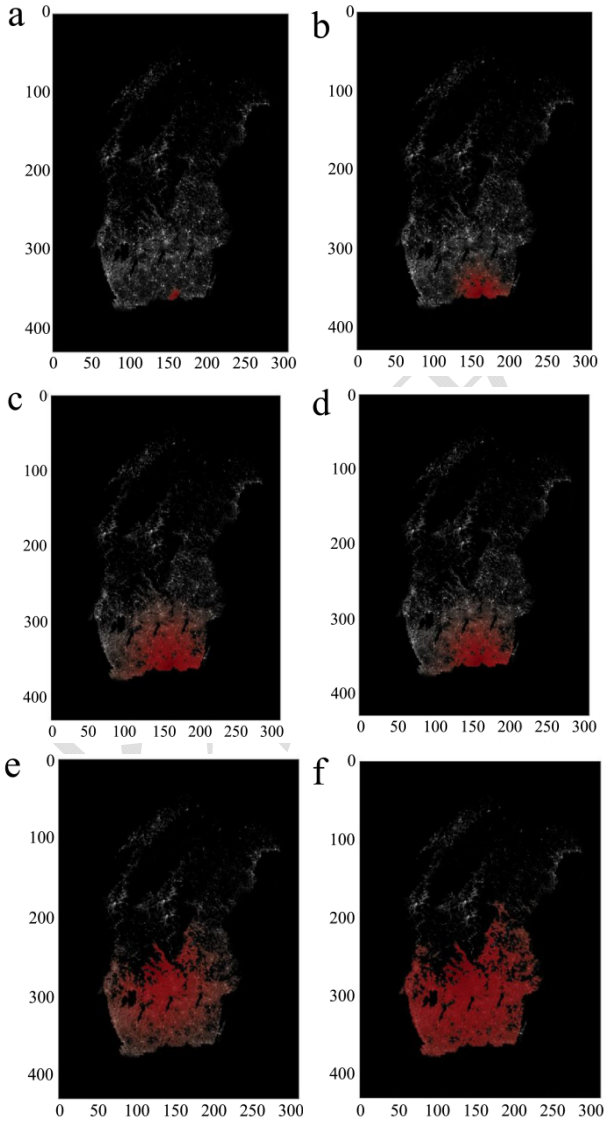


图 10 香肠菌落总数预测值可视化

Fig.10 predicted visualization sausage

注：a，初始时刻；b，8 h 后的菌落总数；c，16 h 后的菌落总数；d，24 h 后的菌落总数；e，30 h 后的菌落总数；f，40 h 后的菌落总数。

6 结论

6.1 本论文对高维的高光谱数据进行 MSC 预处理[16]，PCA 处理后利用最优区间进行建模，大大提高回归预测模型的运行时间，采用 SVM 进行回归预测，能够较好的达到预期的效果，还克服了传统检测方法的低效率等缺点。不仅实现了检测的目的，还可以保证样本的无损检测顺利进行。为使回归预测模型有更高的准确性，采用了不同的方法对参数进行寻优。为将来实际的应用可以提供理论依据。但后期还需要对建模的预测集样本、校正集样本数增加数量，期得到更好的效果。

6.2 对于可视化方面，最好能呈现展现出菌落总数的多维显示，这样更有利于人们对于菌落总数的更加直观和准确的认识。

参考文献

[1] 王虎虎,刘登勇,徐幸莲,等.我国传统腌香肠制品产业现状及发展趋势[J].肉类研究,2013,27(9):36-40
WANG Hu-hu, LIU Deng-yong, XU Xing-lian, et al. Traditional pickled sausage products industry present situation and development trend in China [J]. Journal of Meat Research, 2013, 27(9): 36-40

[2] 赵俊华,郭培源,邢素霞,等.基于高光谱成像的腊肉细菌总数预测建模方法研究[J].中国调味品,2016,2:74-78
ZHAO Jun-hua, GUO Pei-yuan, XING Su-xia, et al. Based on set highlights like bacon bacterium total forecast modeling method research [J]. Chinese Seasoning, 2016, 2: 74-78

[3] 马飞.基于多光谱成像技术的香肠多元品质无损检测研究[D].合肥:合肥工业大学,2015
MA Fei. Sausage multivariate quality based on the technology of multi-spectral imaging nondestructive testing research [D]. Hefei: Hefei University of Technology, 2015

[4] Barbin D F, Elmasry G, Sun D W, et al. Predicting quality and sensory attributes of pork using near-infrared hyperspectral imaging [J]. Analytica Chimica Acta, 2012, 719(10): 30-42

[5] Kamruzzaman M, Elmasry G, Sun D W, et al. Prediction of some quality attributes of lamb meat using near-infrared hyperspectral imaging and multivariate analysis [J]. Analytica Chimica Acta, 2012, 714(3): 57-67

[6] 刘娇.基于高光谱技术的不同品种猪肉品质检测模型传递方法研究[D].武汉:华中农业大学,2015
LIU Jiao. Different varieties of pork quality detection based on hyperspectral technology model transfer method research [D]. Wuhan: Huazhong Agricultural University, 2015

[7] 田潇瑜.基于光谱与图像分析的生鲜牛肉嫩度快速检测技术研究[D].北京:中国农业大学,2014
TIAN Xiao-yu. Based on the spectrum analysis and image analysis of fresh beef tenderness, rapid detection technology research [D]. Beijing: China Agricultural University, 2014

[8] 王伟,彭彦昆,张晓莉.基于高光谱成像的生鲜猪肉细菌总数预测建模方法研究[J].光谱学与光谱分析,2010,30(2):411-415
WANG Wei, PENG Yan-kun, ZHANG Xiao-li. Based on the set highlights like fresh pork total bacterial count of predictive modeling method research [J]. Spectroscopy and Spectral Analysis, 2010, 30(2): 411-415

[9] 周志光.体数据特征的高效可视化方法研究[D].浙江:浙江大学,2012
ZHOU Zhi-guang. The data characteristics of the efficient visualization method research [D]. Zhejiang: Zhejiang University, 2012

[10] Ikram S T. Improving accuracy of intrusion detection model using PCA and optimized SVM [J]. Journal of Computing & Information Technology, 2016, 24(2): 133-148

[11] 孙俊,张梅霞,毛罕平,等.基于高光谱图像的桑叶农药残留种类鉴别研究[J].农业机械学报,2015,46(6):251-256
SUN Jun, ZHANG Mei-xia, MAO Han-ping, et al. Based on hyperspectral image of mulberry leaf pesticide residue species identification study [J]. Journal of Agricultural Machinery, 2015, 46(6): 251-256

[12] Zheng C, Jiao L. Automatic parameters selection for SVM based on GA [C]// Intelligent Control and Automation, 2004. WCICA 2004. Fifth World Congress on. 2004:1869-1872

[13] 徐海龙,王晓丹,廖勇,等.一种基于 PSO 的 RBF-SVM 模型优化新方法[J].控制与决策,2010,25(3):367-370
XU Hai-long, WANG Xiao-dan, LIAO Yong, et al. A new way to optimize RBF based on PSO-SVM model [J]. Control and Decision, 2010, 25(3): 367-370

[14] 王德敏.空气污染数据可视化方法研究及可视化系统实现[D].山东:山东大学,2012
WANG De-min. Air pollution data visualization method research and the realization of visualization system [D]. Shandong: Shandong University, 2012

[15] 朱琪,于济坤,王明德,等.社会网络数据的可视化[J].吉林大学学报(信息科学版),2015,33(5):584-587
ZHU Qi, YU Ji-kun, WANG Ming-de, et al. Social network data visualization [J]. Journal of Jilin University (Information Science Edition), 2015, 33(5): 584-587

[16] 梁琨,杜莹莹,卢伟,等.基于高光谱成像技术的小麦籽粒赤霉病识别[J].农业机械学报,2016,47(2):309-315
LIANG kun, DU ying-ying, LU wei, et al. Identification of fusarium head blight wheat based on hyperspectral imaging technology [J]. Transactions of the Chinese Society of Agricultural Machinery, 2016, 47(2): 309-315