计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900184

基于 GA-SVM 的农产品质量分类

马 创[1] 吕孝飞[2] 梁炎明[2]
1 重庆邮电大学软件学院 重庆 400065
2 重庆邮电大学计算机科学与技术学院 重庆 400065
(machuang@cqupt.edu.cn)

摘 要 传统方法对农产品进行细粒度划分,确定影响分类效果的关键因素,但忽略了农产品的质量特征。对农产品的质量进行科学的分类,能够更好地反映农产品在质量方面的变化,还可以显著提升农产品后续的处理效率。从农产品的质量特征出发,将农产品进行分类,对不同类别的农产品按照不同的方法进行处理,以在保证农产品质量的同时提高农产品的附加值。分类方法与模型参数的选取对于农产品质量分类的准确度尤为重要。传统支持向量机 SVM 对模型参数的选择具有盲目性,为提高分类的准确度,文中提出一和中将因子分析(Factor Analsysi,FA)与基于遗传算法改进的支持向量机(Genetic Algorithm-Support Vector Machine,GA-SVM)结合的分类模型。实验结果表明,改进后的 SVM 能够快速、有效地判别农产品质量类别,显著改善农产品质量的分类精度,评估过程较为简单,可广泛应用于农产品质量的评估。

关键词:农产品;质量分类;因子分析;遗传算法;支持向量机
中图法分类号 TP391

Agricultural Product Quality Classification Based on GA-SVM

MA Chuang[1],LV Xiao-fei[2] and LIANG yan-ming[2]
1 College of Software Engineering,Chongqing University of Posts and Telecommunications,Chongqing 400065,China
2 College of Computer Science and Technology,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract Traditional methods classify agricultural products by fined-grained level and determine the key factors affecting the classification effect,but ignore the quality characteristics of agricultural products. Scientific classification of agricultural products quality can not only effectively improve the speed of subsequent processing of agricultural products,but also better reflect the changes in the quality of agricultural products. Starting from the quality characteristics of agricultural products,agricultural products are classified,and different types of agricultural products are processed in different methods,so as to ensure the quality of agricultural products and increase their added values. The classification method and the selection of model parameters are especially important for the accuracy of agricultural product quality classification. Traditional support vector machine (SVM) has blindness in the selection of model parameters. In order to improve the classification accuracy of agricultural product quality,a product quality classification model combining factor analysis (FA) and improved support vector machine (GA-SVM) is proposed. Experimental results show that the improved SVM can quickly and effectively identify the quality categories of agricultural products,significantly improve the classification accuracy of agricultural product quality. The evaluation process is relatively simple and can be widely applied to the evaluation of agricultural product quality.

Keywords Agricultural products,Quality classification,Factor analysis,Genetic algorithm,SVM

随着科技的发展和社会的进步,我国农业生产能力得到了极大的提高,已经基本解决了农产品的数量安全问题,人们对农产品的需求也从过去的重数量转向现在的重质量。文献[1]中提到从目前已经实施的无公害农产品、绿色食品、有机食品、良好农业规范(GAP)认证和地理标志农产品登记来看,大多侧重于农产品卫生安全方面的评价和地域特性,我国尚

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金面上项目(6172099);重庆市"三百"科技创新领军人才支持计划(CSTCCXLJRC201917);重庆市人工智能技术创新重大主题专项(CSTC2017rgzn-zdyf0140);重庆市技术创新与应用示范重大主题专项项目(CSTC2018JSZX-CYZTZX0178,CSTC2018JSZX-CYZTZX0185)
This work was supported by the National Natural Science Foundation of China (6172099),Chongqing "300" Science and Technology Innovation Leader Talents Support Plan (CSTCCXLJRC201917),Chongqing Artificial Intelligence Technology Innovation Major Theme Project (CSTC2017rgzn-zdyf0140) and Chongqing Technical Innovation and Application Demonstration Major Theme Project (CSTC 2018JSZX-CYZTZX0178,CSTC 2018JSZX-CYZTZX0185).
通信作者:吕孝飞(424085240@qq.com)

未推行统一的根据农产品外观、大小、颜色等要素的质量分类评定工作，一些国际组织和发达国家在农产品质量分类评定方面已拥有成功的经验和做法。大多数的农产品是鲜活产品，它的形状、色泽以及新鲜程度等外在品质特征十分重要。这些质量特征不仅是影响消费者购买的关键因素，也是其商品属性的具体体现。随着社会的发展，我国群众生活质量不断提升，农产品的消费呈现出多样化的趋势和明显的层次性。对农产品的质量进行分类，不仅可以满足消费者不同层次需求，也是对农产品消费多样化的适应。实施农产品质量分类，对满足消费多样化需求、加快我国农业生产方式转变、推进农业标准化进程有着深刻的意义。农业生产方式的转变是农业现代化发展的核心推动力，是“三农”建设的根本。

近年来，计算机视觉为实现农产品质量分类提供了思路，一些基于图像分析和处理的检测方法已成功用于肉类和鱼类、比萨饼以及奶酪和面包的质量分析[2]。文献[3]提到了一种水果自动分级机器视觉系统，用于在线估算橙子、桃子和苹果的质量，但是该方法计算过程较为复杂，分类效果较差。Yang 等[4]提出的基于阶段特性的 APT 攻击行为分类与评估方法可以应用到农产品质量分类中。通过总结农产品的特征，在细粒度方面对农产品进行划分，能够确定影响分类效果的关键因素，但是分类耗时较长且计算过程和操作复杂。Zhang 等[5]将代价敏感的反馈神经网络应用于农产品的分类，利用相关的农产品分类数据对评估模型进行调整，实现了大数据中农产品质量分类。但该方法在应用过程中，没有对农产品的质量进行分析，分类效果较差。Kong 等[6]通过提取几何与形状特征来实现马铃薯的质量分级，具体采用线性回归分析建立马铃薯质量大小与几何特征参数之间的回归模型，然后提取出马铃薯的六个不变矩参数，将其输入已经训练完成的神经网络中，进而实现分级。实验结果表明，此方法的质量分级准确率为 95.3%，满足实际检测分级的要求。An 等[7]提出了一种大数据中农产品质量特征分类方法，通过对不同的特征进行权重计算，完成对农产品质量特征的分类，分类效果较好。

农产品的质量受到多种因素的影响，对其进行分类是一个复杂的非线性问题，所以需要建立一个对影响因素众多且非线性适用的综合分类模型。基于此，本文探索了一种基于 GA-SVM 的农产品质量分类模型。该模型首先对农产品的质量特征进行降维处理，减少对分类模型的输入，从而提高农产品质量分类的效率，并结合遗传算法(Genetic Algorith, GA)的全局寻优能力强、可以并行、执行高效等特点，对支持向量机(SVM)存在的核函数构建及参数选择较为困难等问题进行改进。本文对所提模型进行实验，证明了该模型的准确性和安全性。

1 相关理论基础

1.1 因子分析

因子分析通过研究变量间的相关系数矩阵，把这些变量间错综复杂的关系归结成少数几个综合因子，这些综合因子包含原始变量的信息。简单来说，因子分析就是找出隐藏在多个变量背后具有共性的公共因子。这些公共因子是相互独立的，具有较低的相关性，其线性组合可以表示所有的原始变量。假设有 N 个样本，P 个指标，$\boldsymbol{X}=(x_1,x_2,\cdots,x_1)^{\mathrm{T}}$ 为随机向量，要寻找公共因子 $\boldsymbol{F}=(F_1,F_2,\cdots,F_M)^{\mathrm{T}}$，则因子模型为：

$$\begin{cases} X_1=a_{11}F_1+a_{12}F_2+\cdots+a_{1m}F_m+\varepsilon_1 \\ X_2=a_{21}F_1+a_{22}F_2+\cdots+a_{2m}F_m+\varepsilon_2 \\ \quad\vdots \\ X_1=a_{11}F_1+a_{12}F_2+\cdots+a_{1m}F_m+\varepsilon_1 \end{cases} \tag{1}$$

其中，a_{ij} 为因子载荷，是每个公共因子 F_j 和每个原始变量 X_i 之间的相关系数，反映了变量对因子的重要性，其值越大，表示两者的依赖程度越大。当有多个因子的时候，因子载荷将构成一个矩阵，称为因子载荷矩阵，记 $\boldsymbol{A}=(a_{ij})$ 为因子载荷矩阵。ε 为特殊因子，实际分析时可以忽略。上述模型的矩阵形式为：

$$\boldsymbol{X}=\boldsymbol{A}\boldsymbol{F}+\varepsilon \tag{2}$$

1.2 遗传算法

遗传算法是一种模拟自然界生物进化机制的随机全局搜索和优化方法，其本质是一种高效、并行、全局搜索的方法，具有很强的解决问题能力和广泛的适应性。遗传算法操作使用适者生存的原则，在潜在的解决方案种群中逐次产生一个近似最优解的方案，在每一代中，根据个体在问题域中的适应度值和从自然遗传学中借鉴来的再造方法进行个体选择，产生一个新的近似解。这个过程导致种群中个体的进化，得到的新个体比原来的个体更能适应环境，就像自然界中的改造一样。

1.3 支持向量机

支持向量机(SVM)基于寻找最能将特征分离到不同域的超平面的思想，它将向量映射到一个高维的空间里，在这个空间里有一个最大间隔超平面。在分开数据的超平面两侧有两个互相平行的超平面，分隔超平面使两个平行超平面的距离最大化。假定平行超平面间的距离或差距越大，分类器的总误差越小。

假定训练样本集为 (x_i,y_i)，$i=1,2,\cdots,n$，x_i 为第 i 个输入样本，y_i 为第 i 个输 n 出样本。SVM 通过求解最优化问题来解决分类问题：

$$\begin{aligned} &\min\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\varepsilon_i \\ &\text{s.t.}\ \begin{cases} y_i(wx_i+b)\geqslant 1-\varepsilon_i \\ \varepsilon_i>0 \end{cases},i=1,2,\cdots,n \end{aligned} \tag{3}$$

其中，ε_i 是引入的松弛变量；$C>0$ 是惩罚因子。

为了求解式(2)，需要引入拉格朗日函数将其转化为对偶问题：

$$\begin{aligned} &\max\ w(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \\ &\text{s.t.}\ \begin{cases} \sum_{i=1}^{n}\alpha_i\alpha_j=0 \\ 0\leqslant\alpha_i\leqslant C \end{cases},i=1,2,\cdots,n \end{aligned} \tag{4}$$

其中，α_i 代表拉格朗日乘子，$K(x_i,x_j)$ 为核函数。该核函数为径向基核函(Radial Basis Function, RBF)，$K(x_i,x_j)=\exp\{-g|x-x_i|^2\}$，$g$ 为核函数参数。

SVM 决策函数可以表示为 $f(x)=sign[\sum_{i=1}^{n}\alpha_i j_i K(x_i,x)+b]$。

1.4 GA 优化 SVM

惩罚因子 C 和 RBF 参数 g 的取值会直接影响 SVM 分

类的准确率。惩罚因子 C 与数据的拟合程度成正比关系，C 取值越大则数据拟合程度越高；RBF 参数 g 则决定了线性不可分样本数据映射到高维特征空间后的径向作用范围，过大的参数 g 会降低对农产品的分类效果。因此，参数值的合理选择，对于保证和提高 SVM 的分类性能至关重要。

传统 SVM 随机生成参数值，分类精度存在不稳定因素。遗传算法(GA)具有较强的寻优能力。本文采用 GA 来优化 SVM 参数选择，具体步骤如下。

步骤 1　数据预处理。对训练样本和测试样本进行归一化处理，消除原始变量之间的量纲差异。

步骤 2　种群初始化与编码。构建一定数量的初始种群，对惩罚函数 C 和核函数参数 σ 进行二进制编码。

步骤 3　解码与计算适应度函数。解码种群中的染色体，获取 C 及 σ 并进行训练，通过测试样本集的预测精度 E 计算适应度函数。

步骤 4　选择。通过判断优化过程是否满足遗传算法的终止条件，选择输出最优参数组合，进行步骤 5，或者进行交叉、变异等操作产生新一代种群，开始新的遗传。

步骤 5　将最优参数组合代入 SVM 模型中进行训练。

步骤 6　对测试集进行分类预测，得出分类准确率。

具体流程如图 1 所示。

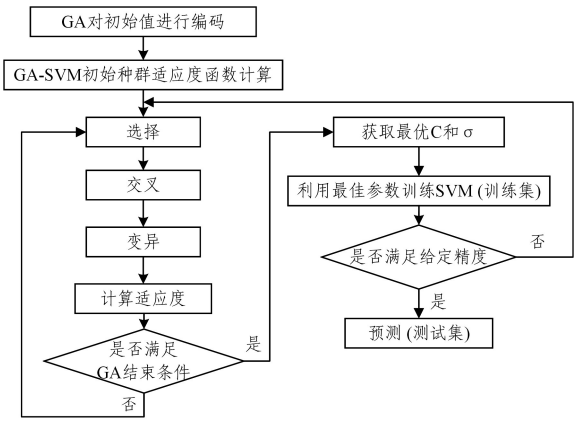

图 1　GA 优化 SVM 流程

Fig. 1　GA optimizes SVM process

2　GA-SVM 农产品质量分类模型及应用

2.1　判别因子的选取

农产品质量分类是受诸多不确定因素影响与控制的复杂非线性问题。影响农产品质量的因素众多，且不同种类的农产品所受影响的因素也有所不同，大致可以概括为形状、大小、颜色、纹理和缺陷程度 5 类。本文实验所选取的农产品为红富士苹果。我国是一个苹果产量大国，2014 年我国水果的总产量为 1.35 亿吨，占全国水果总产量的 26.12%，而且苹果也是我们日常生活中最常见的水果，是十分具有代表性的农产品。本文根据《鲜苹果》(GB/T 10651-2008)，遵循指标选取的重要性、独立性和易测性原则，选取最大横切面平均直径、圆形度、红色区域面积比例和缺陷面积等 4 个特征值作为分类的判别因子。

2.2　因子分析

模型所用 36 组数据均引自文献[8]。运用 SPSS19.0 统计软件对选取的 4 个特征值进行 Pearson 相关性分析和因子分析检验。结果表明，这 4 个特征值之间存在明显的相关关系，特征值上隐含重叠的信息。由 KMO 检验和 Bartlett 球度检验结果(见表 1)可知，样本的 KMO 值为 0.865，Bartlett 球度检验的 Sig. 值为 0，小于 0.01，所以适合做因子分析。以上两个检验结果都表明这 4 个特征值可通过因子分析处理以实现降维的目的。

表 1　KMO 和 Bartlett 球形检验

Table 1　KMO and Bartlett sphericity test

KMO Sample Test		q0.865
Bartlett Spherical Test	Chi-square Value	218.921
	Degree of Freedom	20
	Sig. Value	0

根据实际需要，将最大横切面平均直径、圆形度、红色区域面积比例和缺陷面积这 4 个特征值做因子分析处理，得到公共因子总方差解释表(见表 2)。

表 2　总方差解释表

Table 2　Total variance explain table

Serial Number	Eigenvalues	Contribution Rate/%	Cumulative Contribution Rate/%
1	3.494	87.362	87.362
2	0.205	5.113	92.745
3	0.159	3.974	96.719
4	0.142	3.551	100

为了较少地损失信息，提高分类准确率，由表 2 可得，提取前两个公因子，累计方差贡献率可以达到 92.745%，即前 2 个公因子能够对原始指标变量 92.745%的信息进行解释，将两个公因子分别记为 F_1 和 F_2。

初始因子的综合性过于明显，实际意义难以得到显示，因此方差最大化旋转因子载荷矩阵，最小化每个因子的最高载荷变量数，可以降低因子的综合性。对旋转后的因子载荷矩阵，运用 Regression 法计算各因子得分。

2.3　GA－SVM 农产品质量分类模型的构建

根据《鲜苹果》(GB/T 10651－2008)，将红富士苹果质量等级分为 4 类(优等果、一等果、二等果和等外果)。本实验一共选取了 36 组数据，将其中 29 组红富士苹果的样本数据作为训练样本，7 组样本数据作为测试样本。首先，对红富士苹果的 4 个判别因子进行分析，得到 2 个主要的判别因子 F_1 和 F_2 作为模型的输入，将 4 个质量类别作为输出，建立基于因子分析的 GA-SVM 红富士苹果质量分类模型。

在训练过程中，在经过了 GA 优化 SVM 分类模型参数的 200 次进化后，得到了一个最佳适应度的稳定迭代值，如图 2 所示。

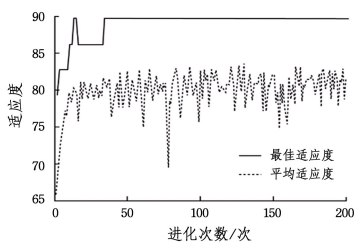

图 2　遗传算法适应度曲线

Fig. 2　Fitness curve of genetic algorithm

3 实验结果及对比

3.1 实验结果

由图 2 可知，GA-SVM 模型的最优惩罚函数为 $C=10.6395$，核函数参数的值为 $\sigma=1.941$。将其代入 SVM 模型进行训练，得到训练集的分类结果如图 3 所示。

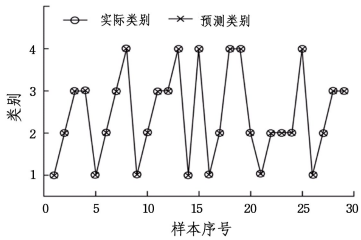

图 3 训练集预测结果

Fig. 3 Prediction results of training set

利用训练好的 GA-SVM 模型对 7 组测试样本数据进行预测分类，结果如图 4 所示。

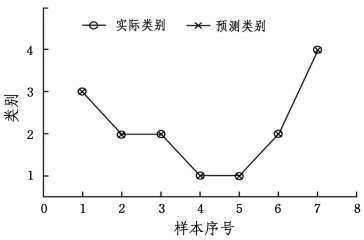

图 4 测试集预测结果

Fig. 4 Prediction results of testing set

由图 3 和图 4 可知，GA-SVM 模型对农产品质量的分类准确且稳定。

3.2 不同模型对比

为了验证 GA-SVM 模型的优越性，将其与 SVM 模型、BP 神经网络模型进行比较，结果如表 3 所列。与 GA-SVM 不同，SVM 模型和 BPNN 模型的输入数据是没有经过因子分析处理的原始数据。根据表 3 可知，GA-SVM 分类的正确率为 100%，明显高于其他 2 种模型的正确率。由此可证，基于因子分析建立的 GA-SVM 农产品质量分类模型安全可靠，能够实现对农产品质量快速、有效的分类。

表 3 GA-SVM 模型的测试集

Table 3 Test set of GA-SVM model

序号	类别	预测结果		
		GA-SVM	SVM	BP
I	3	3	3	3
II	2	2	2	2
III	2	2	2	2
IV	1	1	2	1
V	1	1	1	2
VI	2	2	2	2
VII	4	4	4	4

结束语 本文针对当前对农产品的质量进行分类存在的准确率较低、分类效果较差以及分类过程复杂的问题，提出了一种基于因子分析的 GA-SVM 农产品质量分类模型，将统计学理论及机器学习算法应用于农产品质量分类，减少了特征值间的冗余信息，简化了 GA-SVM 模型分类过程的工作量，提高了分类的效率。实验结果表明，该模型能够对农产品的质量进行准确、有效的判别，保证了分类的可靠性。因此，该模型适用于农产品质量分类问题。

参 考 文 献

[1] YUAN G Y. Discussion on the Specification of Agricultural Product Quality Grade[J]. Quality and Safety of Agro-products, 2016(4): 23-27.

[2] TADHG B, DA W S. Improving quality inspection of food products by computer vision—a review[J]. Journal of Food Engineering, 2004, 61(1): 3-16.

[3] BLASCO J, ALEIXOS N, MOLTÓ E. Machine Vision System for Automatic Quality Gradingof Fruit[J]. Biosystems Engineering, 2003, 85(4): 415-423.

[4] YANG H P, WANG K. Phase-based Classification and Evaluation of APT Attack Behaviors[J]. Computer Engineering and Applications, 2017, 53(22): 97-104.

[5] ZHANG Z L, TANG J G, QIAO D N, et al. A Method for Cigarette Sensory Quality Evaluation Based on Cost-sensitive Learning[J]. Chinese Tobacco Science, 2016, 37(5): 75-81.

[6] KONG Y L, GAO X Y, LI H L, et al. Potato Grading Method of Mass and Shapes Based on Machine Vision[J]. Transactions of the Chinese Society of Agricultural Engineering, 2012, 28(17): 143-148.

[7] AN F J, YU X, MENG Q Q. Simulation of the Classification Effect of Quality Characteristics of Agricultural Products in Large Data[J]. Computer Simulation, 2018(8): 439-442.

[8] YU Y. Research on Classification and Grading Technique of Apple Based on Multi-features[D]. Zhengzhou: Henan Agricultural University, 2017.

MA Chuang, born in 1984, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include complex network, and machine learning.

LV Xiao-fei, born in 1995, postgraduate, is a member of China Computer Federation. His main research interests include machine learning.