Advances in Applied Mathematics 应用数学进展, 2023, 12(8), 3576-3586
Published Online August 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.128356

Hans汉斯

基于随机森林分类模型的葡萄干特征分析

余丽萍*，吴喜之，王　涛

云南师范大学数学学院，云南　昆明

收稿日期：2023年7月18日；录用日期：2023年8月8日；发布日期：2023年8月16日

摘　要

为了实现两种葡萄干的高效率分类，以R语言作为工具，将两种土耳其葡萄干(Besni和Kecimen)的900颗(每种450颗)葡萄干图像数据作为数据集，通过图像提取技术，提取7种形态学特征：Area、Perimeter、MajorAxisLength、MinorAxisLength、Eccentricity、ConvexArea、Extent，数据集经过归一化和清除噪音的处理，选择随机森林算法建立分类模型，与SVM模型相比较，结果表明：随机森林模型使用混淆矩阵进行综合评价结果显示与SVM模型不分上下，但对于葡萄干数据而言，使用随机森林模型对变量重要性的解读更适合，研究表示Perimeter和MajorAxisLength这两个形态学特征对随机森林的分类模型十分重要。

关键词

机器学习，随机森林，特征分类，R语言

Characterization of Raisins Based on Random Forest Classification Model

Liping Yu*, Xizhi Wu, Tao Wang

College of Mathematics, Yunnan Normal University, Kunming Yunnan

Received: Jul. 18th, 2023; accepted: Aug. 8th, 2023; published: Aug. 16th, 2023

Abstract

In order to realize the efficient classification of two kinds of raisins, R language was used as a tool, and the image data of 900 raisins (450 raisins each) of two kinds of Turkish raisins (Besni and Kecimen) were used as a dataset, and seven morphological features were extracted by image extraction technique: Area, Perimeter, MajorAxisLength, MinorAxisLength, Eccentricity, ConvexArea, and Extent, the dataset was normalized and noise removal, and the Random Forest algorithm was

*通讯作者。

文章引用：余丽萍，吴喜之，王涛．基于随机森林分类模型的葡萄干特征分析[J]．应用数学进展，2023，12(8)：3576-3586. DOI: 10.12677/aam.2023.128356

selected to build the classification model, which was compared with the SVM model, and the results showed that: the Random Forest model using the confusion matrix for the comprehensive evaluation of the results showed that it was indistinguishable from the SVM model, but for the raisin data, the interpretation of the importance of the variables using the random forest model is more appropriate, and the study indicated that the two morphological features of Perimeter and MajorAxisLength are important for the classification model of the random forest.

Keywords

Machine Learning, Random Forest, Feature Classification, R Language

1. 引言

葡萄干是富含碳水化合物和营养物质的浓缩食物，含有抗氧化剂、钾、纤维和铁[1] [2]。随着机器学习算法的日益成熟，使用机器结合机器学习的图像处理方法，葡萄干的分类方法也向人工智能方向发展。Ilkay CINAR [1]等人将收集的两种土耳其葡萄干图像数据经过图像处理，归一化和去除噪音从而提取到 7 种特征，使用 Logistic Regression (LR)、Multilayer Perceptron (MLP)和支持向量机(SVM)机器学习技术创建了模型，并进行了性能测量，LR 的分类准确率为 85.22%，MLP 的分类准确率为 86.33%，SVM 的分类准确率为 86.44%，准确率都在 80%以上。Navab Karimi [3]等人拍摄了 1400 幅葡萄干图像，利用图像提取技术共获得 146 个纹理特征，接着利用主成分分析(PCA)从提取的特征中找到最佳特征，使用人工神经网络(ANN)和支持向量机(SVM)对混合物进行分类。与人工神经网络相比，使用前 50 个特征，SVM 分类器的分类结果更有效、更准确。

因此本研究在 Ilkay CINAR [1]等人收集的葡萄干数据的基础上，在 R 上使用 SVM、随机森林等分类方法，建立各种分类模型，选择其中分类效果较好的方法，对其进行更加深入地分析其分类的结果，为探寻分类方法分析特征变量的重要性，从而改进葡萄干识别技术，提供了新的思路。

2. 数据

本研究获取的葡萄干样本数据，来自 Ilkay CINAR [1]等人提供的葡萄干样本图像数据，一共有两个品种，分别是 Besni 和 Kecimen，如图 1 所示，每种葡萄干有 450 粒样本，共计 900 粒葡萄干。Ilkay CINAR [1]等人在前人的基础上，在葡萄干众多的特征类型中，根据形态学特征进行了特征提取过程，每一种葡萄干共挑选出 7 个形态学特征，这些形态学特征的具体描述如下：

Area：给出葡萄干颗粒边界内的像素数。

Perimeter：它通过计算葡萄干颗粒的边界和周围像素之间的距离来测量环境。

MajorAxisLength：给出了主轴的长度，这是可以在葡萄干上画出的最长的线。

MinorAxisLength：给出了小轴的长度，这是可以在葡萄干上画出的最短的线。

Eccentricity：它给出了椭圆的偏心度，它与葡萄干有相同的时刻。

ConvexArea：给出了由葡萄干颗粒形成的区域的最小凸壳的像素数。

Extent：给出葡萄干颗粒形成的区域与边界框中总像素的比率。

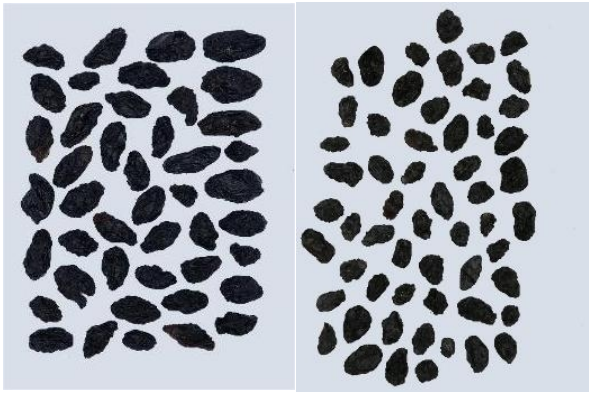


Figure 1. Sample images of the raisin varieties used in the study (Besni (left), Kecimen (right))
图 1. 研究中使用的葡萄干品种的样本图像(Besni (左), Kecimen (右))

3. 研究方法

Ilkay CINAR [1]等人采用的方法分别为 Logistic Regression (LR)，Multilayer Perceptron (MLP)和 Support Vector Machine (SVM)建立模型，根据葡萄干颗粒的特征对其进行分类。其中 LR 和 MLR 属于参数模型，SVM 的使用属于非参数模型，因此本研究也将继续使用其他同类型的方法建立模型。

3.1. 各类算法介绍

3.1.1. 参数模型

参数模型，顾名思义建立模型需要先假设样本的分布服从某一个我们所知道的分布，因此我们可以确定该分布的某一些参数值，比如正态分布的均值和方差。本次研究的参数模型如下：

1) lda：线性判别分析(Linear Discriminant Analysis)通常假定服从多元正态分布，使用贝叶斯的最大后验分布估计(maximum a posteriori estimation, MAP)来判别，寻求特征的线性组合[4]。

2) mda：混合线性判别分析(Mixed linear Discriminant Analysis)是基于高斯混合模型(Gaussian mixture model, GMM)的线性判别分析，假定对于第 k 类的分布为混合的正态(高斯)分布，还需 EM 算法计算必要值，然后用最大后验分布(MAP)法来做判别分析[5]。

3) Logit：logistic 回归(Logistic Regression)假设服从伯努利(Bernoulli)分布，建立的模型阐明了因变量和自变量之间的关系，因此来提取特征。

3.1.2. 非参数模型

非参数模型，建立模型不需要对样本做假设，因此算法需从数据中不断学习，最终建立出一个合理的模型。本次研究的非参数模型如下：

1) SVM：支持向量机(Support Vector Machine)能够通过分离机制将数据分为二维空间的线性空间数据、三维空间的平面数据和多维空间的超平面数据。SVM 找到分离数据的最佳超平面，并执行分类过程，因此 SVM 的最佳超平面是在两个类之间边缘最大的超平面[6]。

2) bagging：bootstrap aggregating 的缩写，基于自助法(bootstrap)抽样的组合方法，自助法抽样是从样本中重复进行放回抽样，是自助法(bootstrap)与决策树的组合方法[7]。

3) RF：随机森林(Random Forest)基于 bagging 算法，也就是说以决策树为基础，并且能够进行随机选择的一种组合方法，因此决策树存在的问题大都得到改善，比如过拟合[8]。

4) knn：k 最近邻方法(K-Nearest Neighbor)根据测试集自变量观测值与训练集自变量观测值的距离最近的 k 个点对测试集的因变量做加权平均[9]。

5) adaboost：全称为 adaptive boosting 方法，是监督学习中的一个二分类模型，与 bagging 不同之处在于 adaboost 每次用自助法抽样来构建分类树时，都会根据前一棵树的结果对误判的观测值，选择判错率高的树的增加抽样权重，使得判错率高的树更具有代表性，使得下一棵树能够令误判的观测值有更多代表性，最终的结果由所有的树加权投票得到[10]。

3.1.3. 交叉验证

交叉验证(Cross Validation)是一种为提高分类安全性而开发的误差预测方法，可以在任何模型之间做客观的评价[1]。本研究使用的为 k 折交叉验证：随机将数据集分成 k 份，随机选择一份作为测试集，另外 $k-1$ 份合并为训练集，用该训练集建模，然后用测试集，测试，算出平均误判率 $error$，公式为

$$error = \frac{1}{k}\sum_{i=1}^{k} err_i$$

其中 err_i 表示模型在第 i 组测试集上观测值被分类错误的个数，$error$ 越接近 0，则说明模型分类效果越好。

因此我们能够根据测试集的判错误差大小对不同模型的比较来判定模型的好坏，比较不同模型的误判率。Ilkay CINAR [1]等人采用的交叉验证为十折，为了减少误差，本次研究也是使用十折交叉验证。

3.2. 精度评估

3.2.1. 混淆矩阵

创建分类问题所需的新模型或使用现有模型后，在该模型上取得成功是通过正确估计的数量来计算的。为了估计分类模型的正确性，我们使用混淆矩阵(Confusion matrix)来评估性能指标，判断模型。在混淆矩阵中有四个参数。这些结果被命名为 tp (true positives)：真阳性；fp (false positives)：假阴性；fn (false negatives)：假阴性；tn (true negatives)：真阴性。正确地归为阳类的例子称为真阳性，正确地归为阴类的例子称为真阴性，被错误归类为阴性的阳性类的例子称为假阴性，而被错误归类为阳性的阳性类的例子称为假阳性，混淆矩阵的四个参数见表 1。

Table 1. Confusion matrix
表 1. 混淆矩阵

		Predicted	
		Kecimen	Besni
Actual	Kecimen	tp	fp
	Besni	fn	tn

混淆矩阵提供了通过在测试数据上的分类模型执行的估计类和真实类的信息，因此有以下指标综合判断分类是否成功。如下为指标及其计算式：

准确率(Accuracy)：

$$\text{Accuracy} = \frac{\text{tp}+\text{tn}}{\text{tp}+\text{fp}+\text{tn}+\text{fn}} \times 100$$

精确率(Precision 或真阳性率 True Positive)：

$$\text{Precision} = \frac{\text{tp}}{\text{fp}+\text{tp}} \times 100$$

敏感率(Sensitivity)：

$$\text{Recall} = \frac{\text{tp}}{\text{tp}+\text{fn}} \times 100$$

假阳性率(Flase Positive):

$$\text{Flase Positive Value} = \frac{\text{fp}}{\text{tn}+\text{fp}} \times 100$$

3.2.2. OOB 误差

OOB (Out Of Bag)：在随机森林中，会有 1/3 左右的样本不会出现 bootstrap 所采集的样本集合中，因此没有参加随机森林中决策树的建立，而这部分未被采集的样本就是袋外数据 OOB，袋外数据就可以用来检测模型的泛化能力，计算公式和混淆矩阵的分类准确率相似，即分类错误的样本数占总样本数的比值.

3.2.3. 决策树分类模型拆分变量的标准

纯度是变量的一个度量，简单说就是该数据集中这个变量的每个观测值之间越相似，该变量所有观测值越接近某个值则纯度就越高。因此对数据中的不同变量，有衡量标准：Gini 不纯度(Gini index 或 Gini impurity)和信息熵(information entropy)。

设因变量有 k 类，记 p_i 为第 i 类在一个节点中的比例，则在一个节点中两种不纯度定义为：

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{k} p_i^2 = 1 - \left(p_1^2 + p_2^2 + \cdots + p_k^2 \right)$$

$$\text{entropy} = -\sum_{i=1}^{k} p_i \log_2 p_i$$

变量纯度是根据该变量根据某标准来拆分时生成的节点的不纯度，再由样本的加权平均值来确定。

4. 结果与分析

4.1. 各种方法的比较

为了在 R 上找到误差最小的判断模型，本次研究分别建立了 lda (线性判别分析的模型)、mda (混合线性判别分析)、Logit (logistic 回归)、SVM (支持向量机)、bagging、RF (随机森林)、knn (k 最近邻方法)和 adaboost 的模型，计算出了这些模型的平均误判率，见图 2。

由图 2 可以看出 SVM 建模的平均误判率是最小的(本次使用的核函数是 R 程序中 ksvm( )函数自动选择的径向基核核函数而不是多项式核核函数[1])，误判率值较小的方法其次为 RF(随机森林)算法，bagging 与 RF 的值相差较小，RF 也是在决策树基础上的组合方法[10] [11]。

与 RF 的平均误判值相差较小的还有 lda (线性判别分析的模型)法，这个方法假定样本的总体服从正态分布，由 Ilkay CINAR 等人采用的方法可知，这些形态特征的分布[1]近似正态分布，所以使用 lda 法和 mda 法也可建模，但 logit (logistic 回归)与这两类方法的差值也只有 0.01 左右，与机器学习相比精度较差。因此选择 RF 法作为下一步详细的研究。

4.2. 随机森林法的特征分析

本次研究使用随机森林法使用 R 语言，建模利用 random Forest( )函数，默认值为 500 棵树，计算得出的 OOB 误差为 13.78%，在 R 程序上计算得到的混淆矩阵见表，由表可知，RF 法的准确率为 86.22%，精确率为 80.6%，召回率为 84.79%，这三个性能指标表现都在 80%以上，说明 RF 法的分类模型在葡萄干样本数据上是非常可行的，且假阳性率为 10.44%，比起 Ilkay CINAR [1]等人的建立的三种模型的假正

例率小，因此从假阳性率上看随机森林模型更适合研究两种这 7 种形态学特征的分类问题[12] [13]。

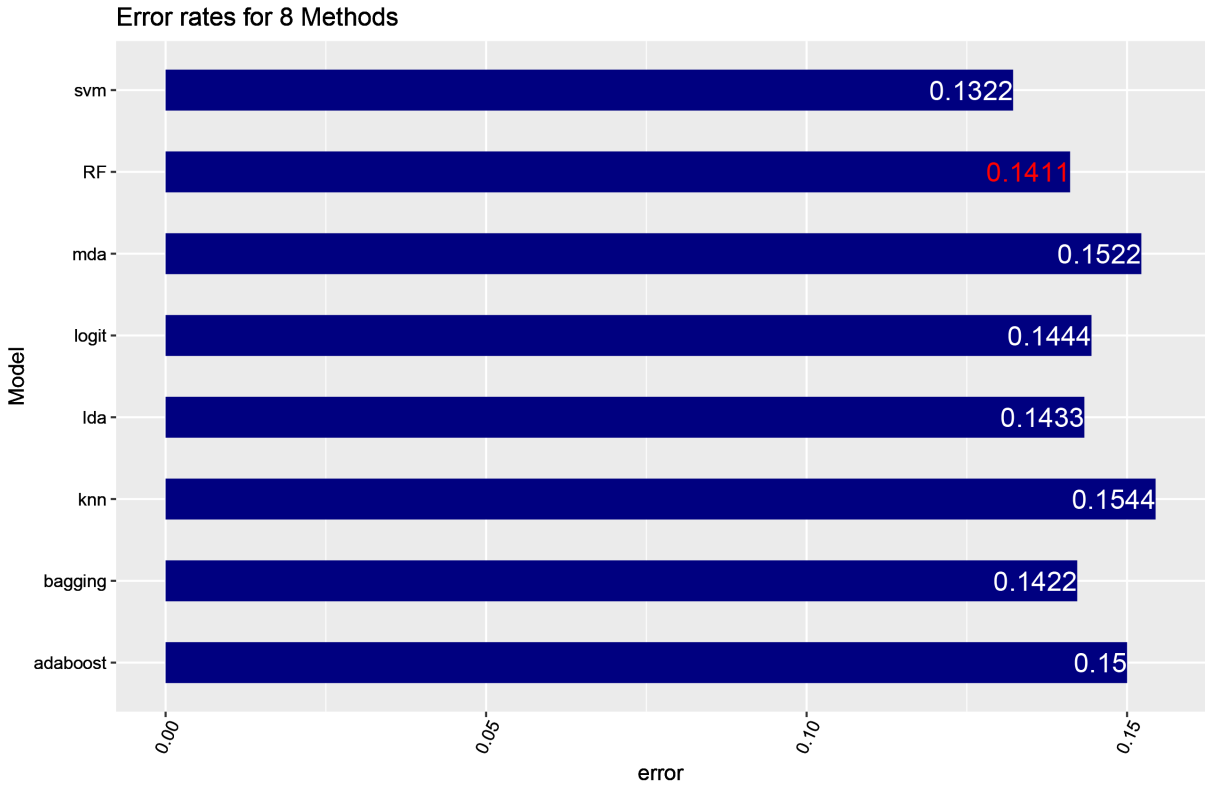


Figure 2. Bar graph comparing the false positive rate of the eight methods
图 2. 八种方法的误判率对比条形图

Table 2. Confusion matrix for two raisin data
表 2. 两种葡萄干数据的混淆矩阵

Algorithms		Predicted	
		Kecimen	Besni
Actual	Kecimen	403	47
	Besni	77	373

随机森林模型的优点之一就是该模型尽可能的提供关于观测值和特征变量的各类信息，这也是本文的研究目的[14]。随机森林有两个重要参数，一个是单棵树的树节点所选的变量个数，另一个是总体规模，也就是随机森林中树的数目[15]。因此评价一个随机森林的模型就从以下两个方面看：1) 每个树越生长的茂盛(分类能力越强)，组成森林的分类效果就越好；2) 每棵树之间的相关性越差，换句话说树与树之间越接近独立分布，则随机森林的分类性能越好。利用 R 中的 treesize( )函数，选择默认值我们可以得到所有树的终节点个数，terminal = T 得到每棵树的所有节点的个数，然后使用画图函数 hist( )得到随机森林每个树的所有节点数和终节点数的直方图，见图 3。

由图 3 可知，左图中大部分的树的节点都超过 20 个，50 棵左右的树平均每棵树约有 60 个节点，右图中每棵树的终节点个数大部分都超过 20 个，终节点超过 50 个的树大约有 40 棵，图 6 表明随机森林模型的 500 棵树生长情况较好而且枝繁叶茂，说明分类能力较强。

最后一个重要参数就是树的深度，深度是指树的根节点到叶子(分叉)节点的路径上节点的数量，因此

延伸出最小深度(min-depth)这个定义，最小深度是根节点到最近节点的最短路径，在选择特征变量时，需要关注最小深度，因为随机森林是由一颗颗独立的决策树组成，因此进行分类时，特征变量越是优先作为分类依据，说明该特征变量对于观测值的分类越重要，最小深度越大说明这个特征变量在树上是处于表层位置的，在 R 中使用 plot_min_depth_distribution ( )函数，因此每个特征变量的最小深度分布图如图 4 所示。

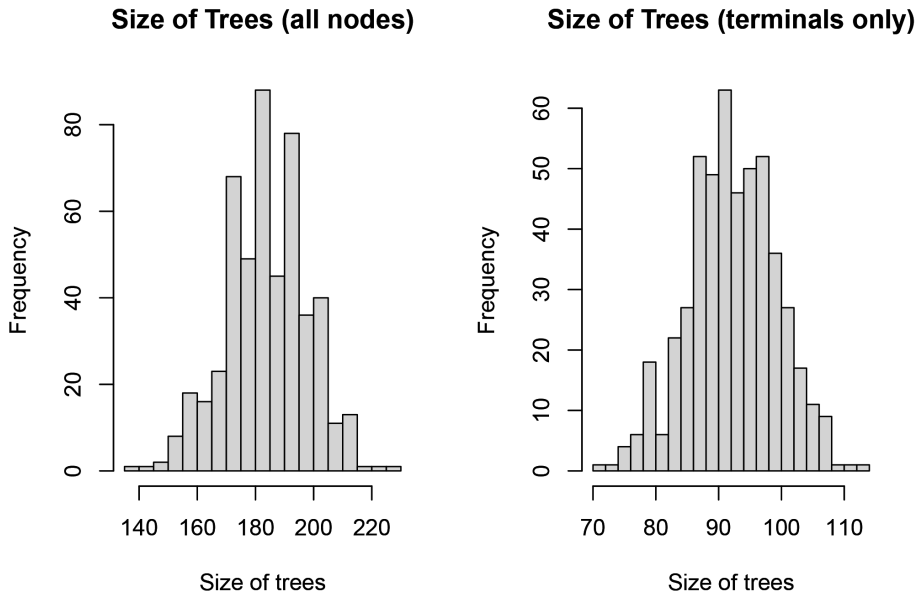


Figure 3. Histograms of the number of all nodes (left) and the number of end nodes (right) for the random forest fit to the raisin data
图 3. 随机森林对葡萄干数据拟合的所有节点数(左)和终节点数(右)的直方图

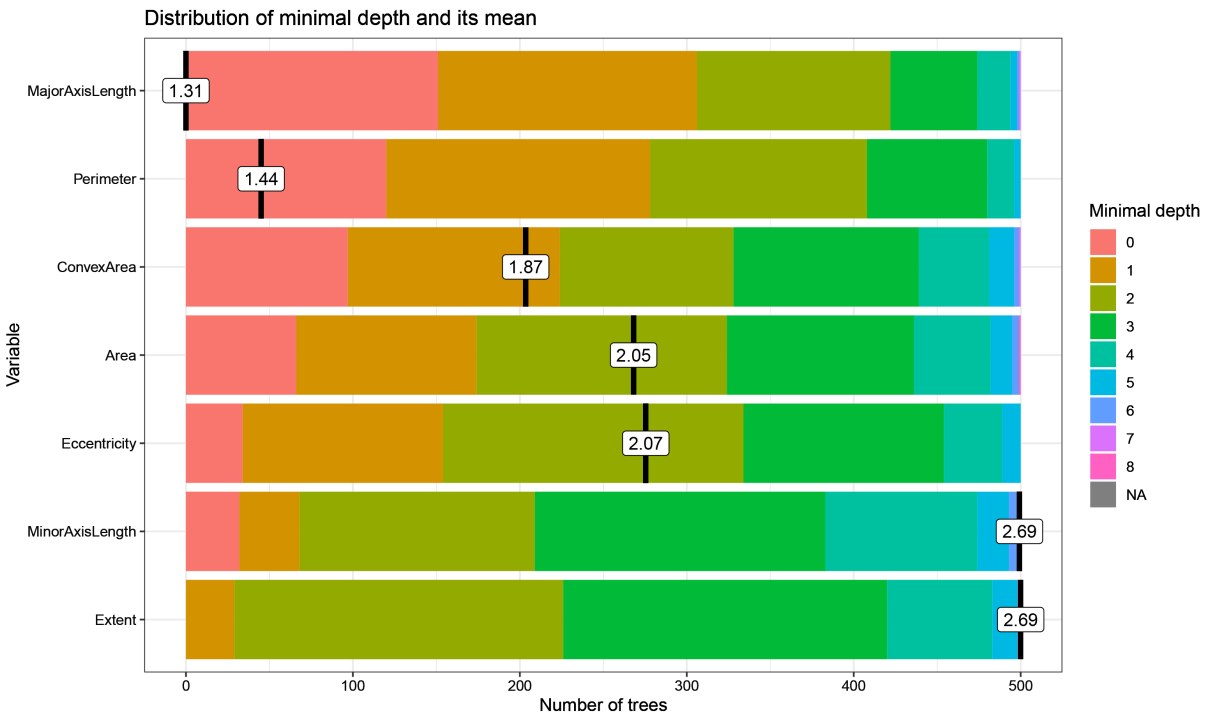
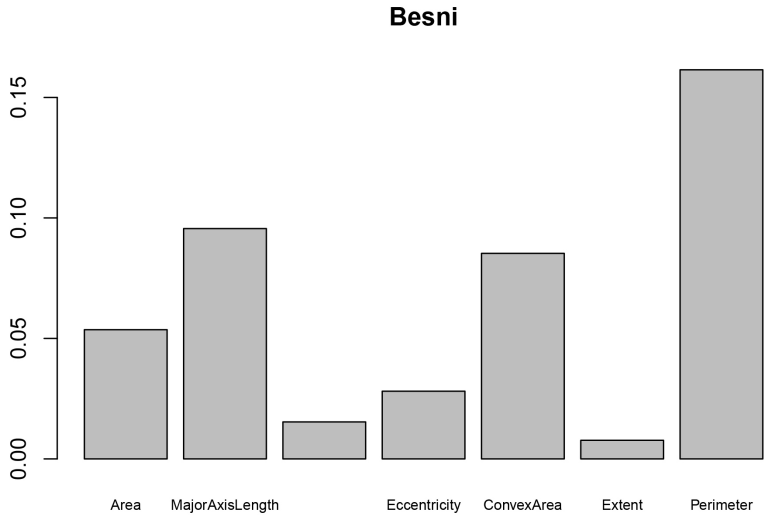


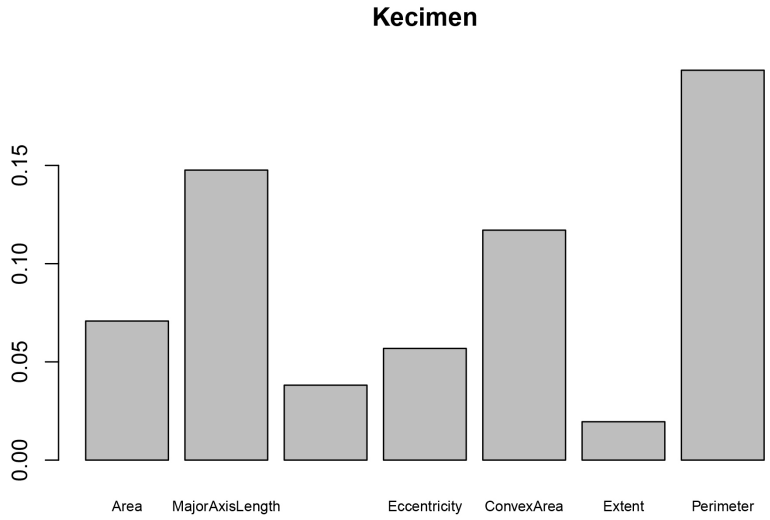
Figure 4. Minimum depth plot for each feature variable
图 4. 各个特征变量的最小深度图

由图 4 可知，图中横坐标为树的数量，纵坐标为每个特征变量的最小深度值分布，右边的图例是按照最小深度由小到大的排序，图中标出的数值为最小深度的平均值，在图上显示越往上的特征变量其最小深度越小，可看出 MajorAxisLength 是在树的最表层的，说明大部分的树都选择 MajorAxisLength 作为表层节点来划分，也就说明这个特征变量显著重要，Perimeter 的最小深度分布中节点数目在 0 的比 MajorAxisLength 的少，节点数为 1、2 和 3 的比较多，说明在后续的节点中会选择 Perimeter，而 Extent 大约有 400 棵树的最小深度值是在 2 到 3 个节点，也就是说几乎 80%的树会选择 Extent 较为深的叶子节点，对比其他特征变量其重要性最小。

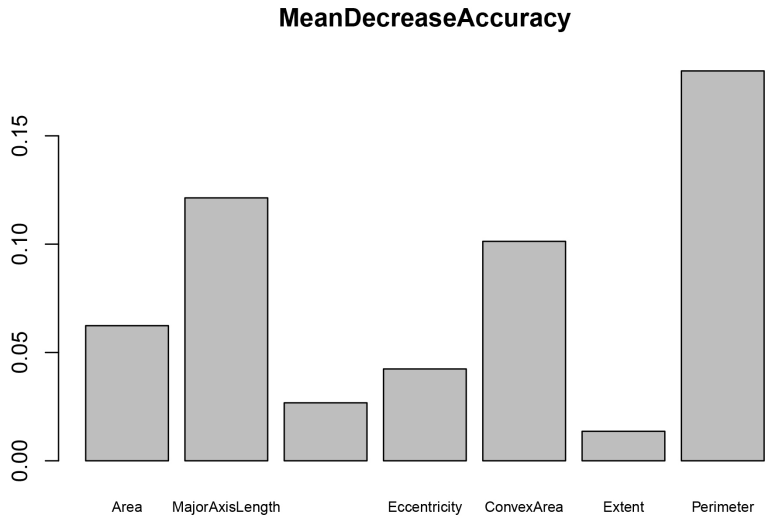
由图 5 可知，上面两幅是关于两类的各个变量的重要性，下面两幅图是综合分类的变量重要性图。从两个葡萄干品种的特征重要性图可以看出虽然这两种葡萄干存在明显的差异，但是从变量重要性图可以看出最具代表性的形态特征是一致的。但是从综合分析的两个图上可知，MeanDecreaseAccuracy 图上最重要的特征变量是 Perimeter，说明该特征被打乱特征值的顺序会影响模型的精确度，因此在该图上 Perimeter 是重要性最大的，但 MeanDereaseGini 图上最重要的是 MajorAxisLength，表示该特征值的基尼系数平均值下降最快，Gini Impurity 记录被错误分类的频率，与随机森林每棵树的节点有关。



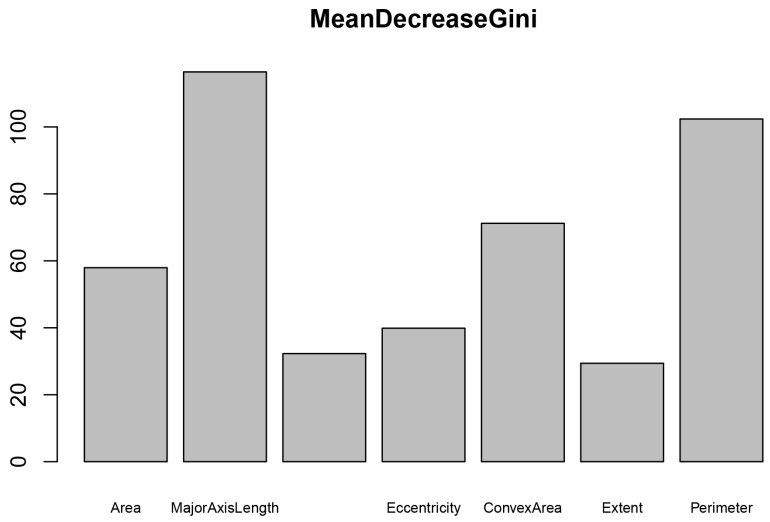
(a)



(b)



(c)



(d)

Figure 5. Variable significance plots (Besi's significance plots for each variable (a), Kecimen's significance plots for each variable (b), MeanDecreaseAccuracy significance plots for each variable (c), MeanDereaseGini significance plots for each variable (d))

图 5. 变量重要性图(Besi 的各个变量的重要性图(a)；Kecimen 的各个变量的重要性图(b)；MeanDecreaseAccuracy 变量重要性图(c)；MeanDereaseGini 变量重要性图(d))

如图 6 所示，图 6 为部分依赖图，它能够显示特征变量对模型(预测结果)有影响的相关关系，假定计算的每个特征变量相互独立，没有存在相关性，同时还忽略了特征变量的边际依赖性，在 R 中使用 partialPlot( ) 函数我们就可以画出每个特征变量的部分依赖图。见图 6 可知，特征变量 Area 和 MinorAxisLength 曲线说明这两个特征变量变化只会影响模型在一定范围内的预测结果，在趋于一定数值之后就不对模型有明显影响，Perimeter、MajorAxisLength 和 ConvexArea 这三个特征变量的曲线存在非常明显的变化会对分类预测结果有显著影响，而 Eccentricity、和 Extent 这两个变量的曲线只有无意义的抖动对模型分类预测结果的影响没有解读意义。

虽然随机森林模型的误判率低，但是我们还需要不断的调整，以达到更加低的误判率同时提高模型效率，在 R 中只需使用 plot( )函数，见图 7。

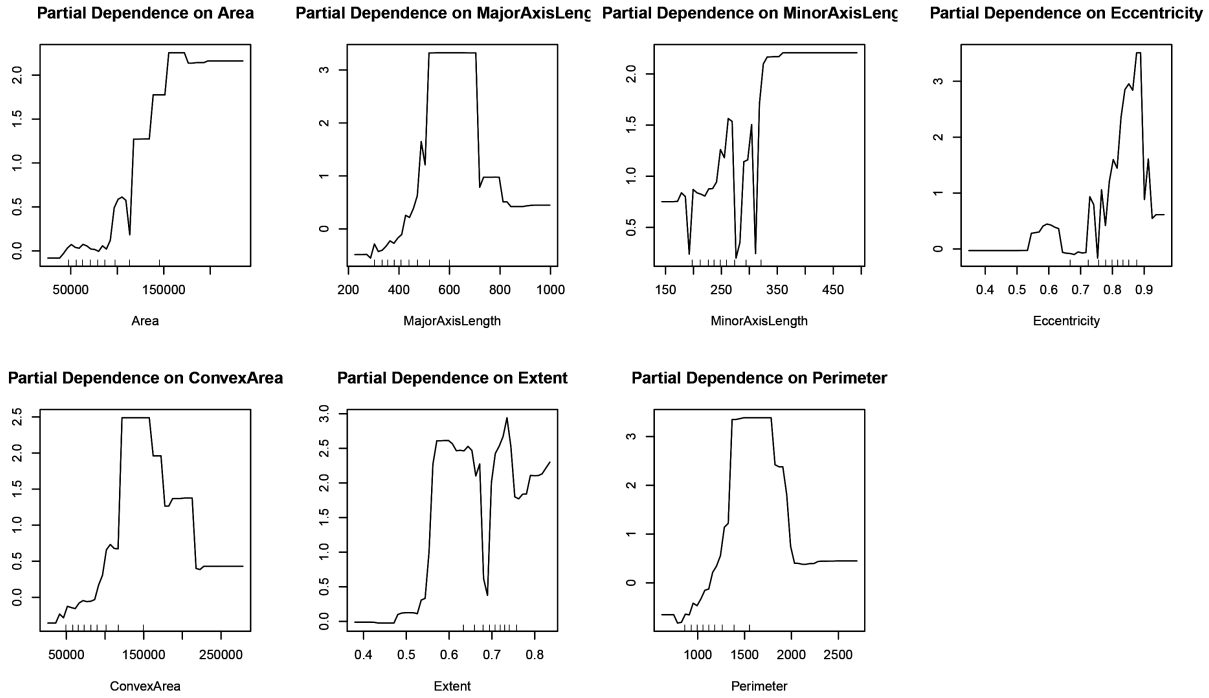


Figure 6. Partial dependency diagram
图 6. 部分依赖图

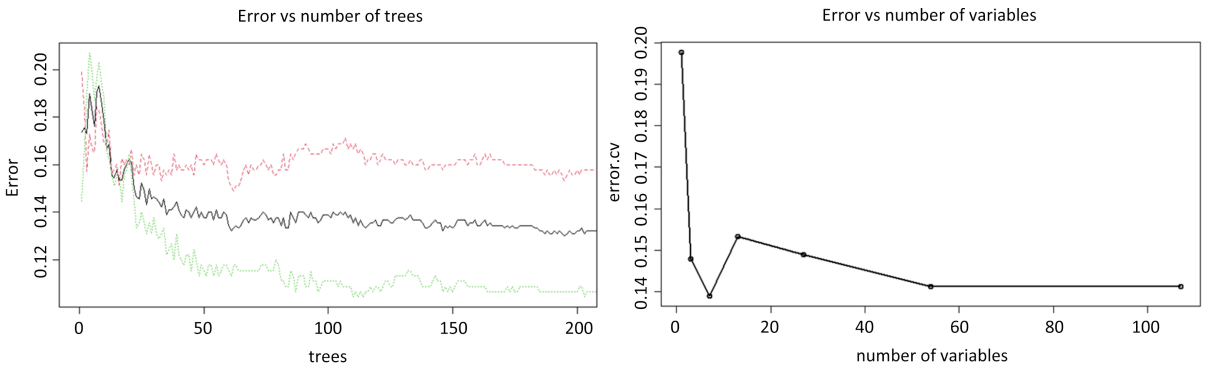


Figure 7. Plot of number of decision trees (left) and number of variables (right) against error
图 7. 决策树数目(左)及变量个数(右)与误差的关系图

图 7 的两图纵坐标都为误差率，左图横坐标为建立的树的数量，由左图可以看出在树的数量在(0, 20)的范围内，当决策树数量达到 25 棵左右，误差率才随数目增加而减小，由左图可以推断模型只需要 50 棵树就可以达到要求；右图为通过 10 折交叉验证不断随机挑选训练集和测试集得到的特征变量个数(横坐标)与误差率的关系图，图中根据变量的重要性来确定图中变量数量的变化顺序，可以看出当特征变量约为 8 个左右时，模型的误差达到最小，这也符合 Ilkay CINAR [1]等人研究前人文章从而提取特征的结果。

5. 结论

综上所述，Ilkay CINAR [1]等人将收集的两种土耳其葡萄干图像数据提取出 7 种形态学特征的数据，使用逻辑回归(LR)、多层感知器(MLP)和支持向量机(SVM)创建了模型，SVM 模型分类效果最佳。而本文使用了随机森林(RF)建立模型，综合比较 SVM 模型，随机森林更适合葡萄干品种特征分析问题。通过

使用随机森林算法对葡萄干数据进行特征分析，对这 7 个形态学特征的变量重要性的分析结果表明：对分类结果最重要的变量是 Perimeter 和 MajorAxisLength，其余五个特征变量对分类结果的影响稍差，总体来说这 7 个特征变量都是影响葡萄干分类结果的重要变量，并且每个特征变量存在相关性，不能假设特征变量是互不干涉。

参考文献

[1] Cinar, I., Koklu, M. and Tasdemir, S. (2020) Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods. *Gazi Journal of Engineering Sciences*, **6**, 200-209. https://doi.org/10.30855/gmbd.2020.03.03

[2] 李忠新, 朱占江, 杨莉玲, 杨忠强, 崔宽波, 刘奎, 刘佳, 沈晓贺, 买合木江. 推进新疆葡萄干走向国际市场的技术对策研究[J]. 新疆农业科学, 2012, 49(6): 1103-1109.

[3] Karimi, N., Kondrood, R.R. and Alizadeh, T. (2017) An Intelligent System for Quality Measurement of Golden Bleached Raisins Using Two Comparative Machine Learning Algorithms. *Measurement*, **107**, 68-76. https://doi.org/10.1016/j.measurement.2017.05.009

[4] Wen, J, Fang, X.Z, Cui, J.R., et al. (2019) Robust Sparse Linear Discriminant Analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, **29**, 390-403. https://doi.org/10.1109/TCSVT.2018.2799214

[5] 黄国宏, 刘刚. 一种新的基于高斯混合模型的线性判别分析[J]. 计算机工程与应用, 2007, 43(27): 75-77.

[6] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2011, 40(1): 2-10.

[7] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1007/BF00058655

[8] 李旭青, 刘世盟, 李龙, 等. 基于 RF 算法优选多时相特征的冬小麦空间分布自动解译[J]. 农业机械学报, 2019, 50(6): 218-225.

[9] Kumar, N.S. and Arun, M. (2017) Genetic Algorithm-Based Feature Selection for Classification of Land Cover Changes Using Combined LANDSAT and ENVISAT Images. *International Journal of Bio-Inspired Computation*, **10**, 172-187. https://doi.org/10.1504/IJBIC.2017.086700

[10] 黄衍, 查伟雄. 随机森林与支持向量机分类性能比较[J]. 软件, 2012, 33(6): 107-110.

[11] Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning*, **1**, 81-106. https://doi.org/10.1007/BF00116251

[12] Solomatine, D.P. and Shrestha, D.L. (2004) AdaBoost.RT: A Boosting Algorithm for Regression Problems. 2004 *IEEE International Joint Conference on Neural Networks*, Budapest, 25-29 July 2004, 1163-1168. https://doi.org/10.1109/IJCNN.2004.1380102

[13] 杨迎港, 刘培, 张合兵, 张文志. 基于特征优选随机森林算法的 GF-2 影像分类[J]. 航天返回与遥感, 2022, 43(2): 115-126.

[14] 王日升, 谢红薇, 安建成. 基于分类精度和相关性的随机森林算法改进[J]. 科学技术与工程, 2017, 17(20): 67-72.

[15] 李坤, 赵俊三, 林伊琳, 陈轲, 毕瑞. 基于RF和SVM模型的东川泥石流易发性评价研究[J]. 云南大学学报(自然科学版), 2022, 44(1): 107-115.