文章编号 1004-924X(2013)08-2121-08

面向田间籽棉成熟度判别的二种特征选择算法比较

王 玲,刘德营,姬长英*

(南京农业大学 工学院/江苏省现代设施农业技术与装备工程实验室,江苏 南京 210031)

摘要:为了快速、准确地判别田间籽棉的成熟度,提取了描述棉瓣形状的 15 个结构特征,基于 10 折交叉验证比较了封装器下穷举搜索并基于封装器停止搜索(WE-W)和过滤器下启发式搜索并基于封装器停止搜索(FH-W)这二种特征选择算法的执行效率和分类性能。分别以验证集上 Bayes 分类器的误分率(WE-W)和训练集上的类可分性测量值(FH-W)为评价函数,在训练集上穷举搜索(WE-W)和启发式搜索(FH-W)最优 l 维特征子集,l=1, 2, …, 15,并于 Bayes 分类器在验证集上的平均误分率极小时停止搜索(WE-W 和 FH-W)。结果显示,WE-W 和 FH-W 算法在预测集上于 l=3 处分别获得了 85.39%(WE-W)和 85.28%(FH-W)的平均识别率,表明 FH-W 算法执行效率高、分类性能好,对实际应用有参考意义。

关 键 词:籽棉成熟度;封装器;穷举搜索;过滤器;启发式搜索;特征选择

中图分类号:TP391.4 **文献标识码**:A **doi**:10.3788/OPE.20132108.2121

Comparison of two feature selection algorithms oriented to raw cotton ripeness discrimination

WANG Ling, LIU De-ying, JI Chang-ying*

(*College of Engineering, Nanjing Agricultural University/Jiangsu Province Engineering Lab for Modern Facitity Agriculture Technology & Equipment, Nanjing* 210031, *China*)

* *Corresponding author, E-mail: chyji@njau.edu.cn*

Abstract: To discriminate the ripeness of cotton quickly and accurately, 15 shape structure features were extracted from cotton images and the execute efficiencies and classification accuracy of their feature subset selection algorithms such as Wrapper-based Exhaustive searching and Wrapper-based stopping(WE-W) and Filter-based Heuristic searching and Wrapper-based stopping(FH-W) were compared by using 10-fold cross-validation. By taking the error rate of a Bayes classifier on validation set (WE-W) and the class-separability measuring value on a training set (FH-W)as assessing functions, the optimal l (l=1,2,3, …, 15) feature subset was searched by using exhaustive (WE-W) and heuristic (FH-W) strategies on the training set, which stops at the minimum error rate of Bayes-classifier on the validation set(WE-W and FH-W). Experimental results show that the average classification rates of WE-W and FH-W algorithms on the prediction set are 85.39% (WE-W) and 85.28% (FH-W) at l=3, respectively. It concludes that the FH-W algorithm can be a reference in practice for its

收稿日期:2013-02-07;**修订日期**:2013-03-14.

基金项目:国家 863 高技术研究发展计划资助项目(No, 2006AA10Z259);江苏省农机基金资助(No, GXZ10007)

higher execute efficiency and good classification accuracy.

Key words: cotton ripeness; wrapper; exhaustive search; filter; heuristic search; feature selection

1 引　言

棉花采摘机器人可从源头上解决我国棉花采摘、收购过程中存在的棉包一致性差等质量问题，从而提高棉花等级相符率。田间籽棉成熟度判别是棉花采摘机器人视觉系统的关键技术之一。描述田间籽棉成熟度一般基于形态信息，但这可能存在维数灾难，故必须进行特征选择。特征选择一般包括搜索策略、评价函数、停止条件和验证结果4个步骤。其中，搜索策略有穷举搜索和启发式搜索两种方法，穷举搜索可以挖掘最优解，但计算量大，处理高维问题有时几乎无法实现；启发式搜索计算量小，是一种次优搜索，有时能够达到穷举搜索类似的效果。评价函数分封装器和过滤器，封装器通过分类器的误分率来评价特征子集。特征拟合分类器虽然分类性能好，但速度慢，时间主要消耗在成千上万次分类器的训练及其性能验证上；而过滤器是基于数据本身的特性来评价特征子集的，速度快，但不一定能真正反映分类器对特征的侧重，分类性能不确定[1]。可见，在兼顾特征选择算法的执行效率和分类性能方面，搜索策略与评价函数的匹配问题已成为一个NP难题。

近年来，混合过滤器、封装器和启发式搜索已成为研究热点，并获得了满意的效率和效果[2-12]，其中，过滤器涵盖Bhattacharyya距离、Jeffries-Ma-tusita距离、相关性增益、互信息、ReliefF、统计显著性、最小生成树等评价准则，封装器涉及人工神经网络、支持向量机、1-近邻图等分类器，启发式搜索包括最优特征排序、顺序搜索、浮动搜索、梯度搜索、遗传算法、蚁群优化。相关研究表明，基于距离的过滤器性能优于互信息和ReliefF，新兴的类可分性过滤器性能近似封装器[4,7]；基于训练集附近的局部泛化误差设计的封装器能快速地剔除90%的特征[2]；针对生物信息数据高维和小样本情形，基于单变量t检验的过滤器和基于大样本、多变量ReliefF的封装器更合适[8]；用新特征替换辨识力较弱的特征改进浮动搜索可获得近似最优解[9]；基于支持向量机和蚁群优化搜索特征可获得95%以上的识别率[10]；遗传算法在解空间进行高效启发式搜索，并且噪声较大时，可有效辨识伺服系统的模型参数[11]；基于JM距离搜索最优高光谱波段的梯度算法，算法简单，但接近最优解时收敛速度很慢，有时得不到最优解[12]。

针对田间籽棉成熟度判别，本文提出采用“类可分性过滤器”匹配“启发式搜索”选择关键特征子集，并比较“Bayes封装器”匹配“穷举搜索”获得的最优分类性能，以寻找执行效率高、分类性能好的特征选择算法。由于经过特征选择的数据集最终都要用来设计分类器，本文拟基于封装器的最优解停止特征选择，并在新的数据集上验证结果，力求以一个较小的特征子集获取较高的分类性能。

2 样本采集与划分

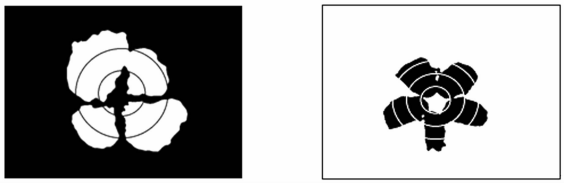
实验用籽棉正面图像样本共计540幅，其中，成熟籽棉390幅，未成熟籽棉150幅，于2009年秋在南京农业大学苏棉12号试验田用CCD数码相机获取，采集分辨率为640×480，全部样本已进行二值化处理，电脑配置为Intel core(TM) 2 Duo CPU，2.93 GHz主频，2 GB内存。

由于模式识别领域通常要求特征数量尽可能少，并且每类样本的数量比特征数量至少大5～10倍，分类器才比较稳定[13]。因而，特征数量过多的复杂分类器或小数据集都有过度拟合数据的危险；而10-交叉验证在样本量不足的情形下特别有用，目前已成功地应用于实践中。本文将全部样本的2/3划分为10等份(每次用其中的9份作为训练集，剩余1份为验证集)，将剩余的1/3样本作为预测集。

3 特征提取

田间籽棉通常由3～5个棉瓣构成，未成熟籽棉外围突兀、棉芯紧实，成熟籽棉外围饱满、棉芯绽开，过熟籽棉棉瓣僵硬、稀疏，籽棉的形态信息能够客观反映籽棉的成熟度。考虑到精度与速度的匹配问题，用1～3个同心圆切割获取以下结构

特征(图 1)：①面积特征(棉瓣与外接圆的面积比)；②一个同心圆切割：径向切割区域特征 1(棉瓣内圈与外圈的面积比)，圆周向切割区域特征 1、2(棉瓣内圈、外圈占其所在圆或圆环的面积比)；③二个同心圆切割：径向切割区域特征 2、3(棉瓣内圈与中圈、外圈与中圈的面积比)，圆周向切割区域特征 3、4、5(棉瓣内圈、中圈、外圈占其所在圆或圆环的面积比)；④三个同心圆切割：径向切割线特征 1、2(棉瓣内切割线与中切割线、外切割线与中切割线的长度比)，圆周向切割线特征 1、2、3(棉瓣内、中、外切割线占其所在圆周长的长度比)；⑤计盒维数：以不同尺度 r 的栅格度量棉瓣区域，统计栅格子数 $N(r)$，基于线性回归求取 $D=\log N(r)/\log r$。



(a)成熟籽棉切割区域　(b)不成熟籽棉切割线

(a) Cutting region of ripe raw cotton　(b) Cutting line of unripe raw cotton

图 1　棉瓣切割示意图

Fig. 1　Cutting diagrams of cotton flap

将面积特征，径向切割区域特征 1～3，圆周向切割区域特征 1～5，径向切割线特征 1～2，圆周向切割线特征 1～3 和计盒维数作为特征集，记作{ x_1, x_2, …, x_{15}}，其中，{ x_1, x_2, x_{15}}描述了籽棉的整体形态，{ x_4, x_6, x_8, x_9, x_{11}, x_{13}, x_{14}}、{x_3, x_5, x_7, x_{10}, x_{12}}描述了籽棉的外围、棉芯结构。

相关分析表明，切割区域与切割线特征之间、径向与圆周向切割特征之间的部分相关系数超过了 0.9(表 1)。

表 1　相关系数大于 0.9 的特征

Tab. 1　Features with correlation coefficient greater than 0.9

(x_i, x_j)	相关系数	(x_i, x_j)	相关系数
(x_3, x_{10})	0.932	(x_6, x_{14})	0.961
(x_4, x_9)	0.907	(x_7, x_{12})	0.915
(x_4, x_{11})	0.905	(x_8, x_{13})	0.943
(x_6, x_9)	0.982	(x_9, x_{14})	0.939

单因素方差分析表明，在 $\alpha=0.05$ 的显著性水平下，仅 x_7、x_{12} 在二类样本上的均值差异不显著，但方差差异显著(图 2)。

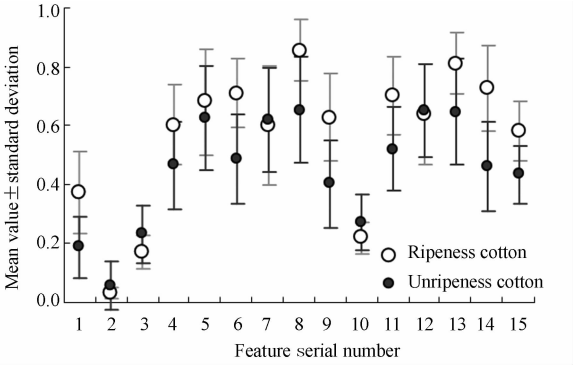


图 2　特征集在二类样本上的均值和标准差

Fig. 2　Mean value and standard deviation of feature set in two category samples

4　WE-W 特征选择

4.1　WE-W 算法

WE-W 算法，即在封装器(Wrapper)下穷举搜索(Exhaustive)特征子集并基于封装器(Wrapper)停止搜索，分类性能好但执行效率低，其实用性是一个值得关注的问题，算法可描述为：

步骤 1：假设

(1)Subset_l 为 m 维特征集 Data_m 的 l 维特征子集

$Data_m=Data(Feature_1, Feature_2, \cdots, Feature_m)$；

步骤 2：封装器下穷举搜索

For $l=1$ to m

(2) 穷举 Data_m 中所有可能的 l 维特征子集(Subset_l)，共计 $m!/(l!(m-l)!)$个，

$Subset_l=\text{Exhaustive\_Search}(Data_m)$；

(3) 用 Subset_l 在训练集(Train Data)上训练 Bayes 分类器(Classifier_l)

$Classifier_l=Build(\text{Train Data}, \text{Subset}_l)$，

(4) 计算 Classifier_l 在验证集(Validate Data)上的误分率(Error_l)

$Error_l=Test(\text{Validate Data}, \text{Classifier}_l)$，

(5) 获取 $m!/(l!(m-l)!)$个 Error_l 的最小值(Error_{best_l})及相应的最优 l 维特征子集(Subset_{best_l})

$$Error_{best\_l}=\mathrm{Min}\{Error_l\}_{(m!/(l!(m-l)!))}$$

$$Subset_{best\_l}=Error_{\min}\{Subset_l\}_{(m!/(l!(m-l)!))}$$

End;

步骤 3:停止条件

(6) 以尽可能小的子集容量(l)获取 m 个 $Error_{best\_l}$ 中的极小值($Error_{best}$)及相应的最优特征子集($Subset_{best}$)

$$Error_{best}=\mathrm{Min}\{Error_{best\_l}\}_{(l=1,2,\cdots,m)};$$

$$Subset_{best}=Error_{\min}\{Subset_{best\_l}\}_{(l=1,2,\cdots,m)}$$

步骤 4:验证结果

(7) 计算 $Subset_{best}$ 在预测集(Prediction Data)上的泛化误差($Error_{generalization}$).

$Error_{generalization}=Test$ (Prediction Data, $Subset_{best}$)。

在第(3)、(4)步中,分类器包括 Bayes 分类器以及最近邻、神经网络、决策树、支持向量机等几何分类器,其中,Bayes 分类器简单、有效、语义明确、容易理解,已成功地应用于实践中,其性能优于或相当于其他分类器[14-15]。Bayes 分类器基于最小错误率决策准则,利用 ω_i 类的均值 u_i、先验概率 $p(\omega_i)$、类条件概率密度函数 $p(x|\omega_i)$(式 1)和 Bayes 公式计算后验概率 $p(\omega_i|x)$(式 2),将样本 x 归入后验概率最大的类别。实际应用中,假定 $p(x|\omega_i)$ 服从多元正态分布 $N(u_i, s_i)$ 且各类协方差 $s_i=E[(x-u_i)(x-u_i)^{\mathrm{T}}]$ 同质,利用 0-1 损失函数可获取线性 Bayes 分类器用以测量 l 维特征子集的分类性能。

$$p(x|\omega_i)=\frac{\exp(-(x-u_i)^{\mathrm{T}}s_i^{-1}(x-u_i)/2)}{\sqrt{(2\pi)^l|s_i|}}, \quad (1)$$

$$p(\omega_i|x)=\frac{p(x|\omega_i)p(\omega_i)}{\sum_{i=1}^{l}p(x|\omega_i)p(\omega_i)}, \quad (2)$$

在第(6)步中,随着特征子集容量 $l=1, 2, 3, \cdots, m$ 的不断增加,拟合方法适应训练集,训练集误分率将逐渐减小;当 l 增加到足够大时,训练集误分率会减小到零,此时,验证集误分率($Error_{best\_l}$)通常很大,期间经历了逐渐减小至增大的过程,其拐点处产生 $Error_{best}$ 和 $Subset_{best}$。

4.2 WE-W 实验结果

在各训练集上穷举 C_{15}^{l} 个 l 维特征子集(X_l),$l=1, 2, \cdots, 15$,用 X_l 建立线性 Bayes 分类器将训练集、验证集和预测集样本划分为 2 类,在验证集上具有最小误分率的 X_l 为最优解。基于 10 个训练集穷举搜索的最优 X_l 各异(表 2),运行总耗时为 1 207.63 s。

表 2　穷举搜索的最优 X_l

Tab. 2　Optimal X_l based on exhaustive search

l	训练集 1	训练集 2	训练集 3	训练集 4	训练集 5
1	2	14	2	13	11
2	5 8	14 15	5 8	8 12	9 14
3	5 7 8	6 9 15	6 11 13	5 7 8	3 11 13
4	5 7 8 14	5 7 8 9	5 10 12 15	2 5 7 8	3 6 11 13
5	1 5 7 8 14	5 7 8 9 12	3 5 8 10 13	5 7 8 14 15	4 5 6 7 11
…	…	…	…	…	…
l	训练集 6	训练集 7	训练集 8	训练集 9	训练集 10
1	9	14	14	14	15
2	3 7	3 7	8 11	11 13	5 8
3	5 7 8	5 7 8	6 9 11	5 7 8	2 6 9
4	2 5 7 8	5 7 8 15	1 6 9 11	5 7 8 12	5 7 8 13
5	1 5 7 8 12	5 7 8 9 15	5 6 7 8 14	1 5 7 8 15	5 7 8 11 15
…	…	…	…	…	…

随着 l 的不断增加($l=1, 2, \cdots, 15$),最优 X_l 在验证集上的平均误分率急剧下降,直至 $l=3$ (12.78%±1.43%)以后趋于平缓(图 3),最优 X_l 在 $l=3$ 处的训练集、预测集平均误分率分别为 11.23%±1.78%、14.61%±2.18%,本着特征数量尽可能少的原则,确定最优特征子集容量 $l=3$。

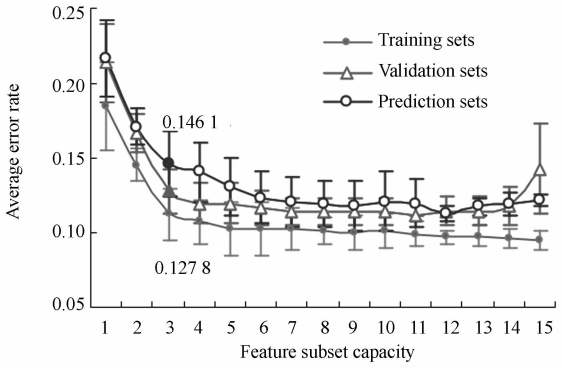


图 3　穷举搜索策略下 Bayes 分类器在数据集上的平均误分率随着特征子集容量而变化

Fig. 3　Average error rate of Bayes classifier on data set varies with feature subset capacity by using exhaustive search

最优特征子集,如图 4 所示。在训练集、验证集和预测集上的平均识别率比较稳定(88.77% > 87.22% > 85.39%)。其中,在 5 个训练集上

获取的{x_5，x_7，x_8}分类性能较好，相关系数小于 0.9，可以较好地描述棉芯的结构；在 3 个训练集上获取的{x_6，x_9，$x_2/x_{11}/x_{15}$}分类性能一般，相关系数大于 0.9，可以描述籽棉的外围结构和整体形态；在 2 个训练集上获取的{x_{11}，x_{13}，x_3/x_6}分类性能较差，相关系数小于 0.9，主要描述了籽棉的外围结构。

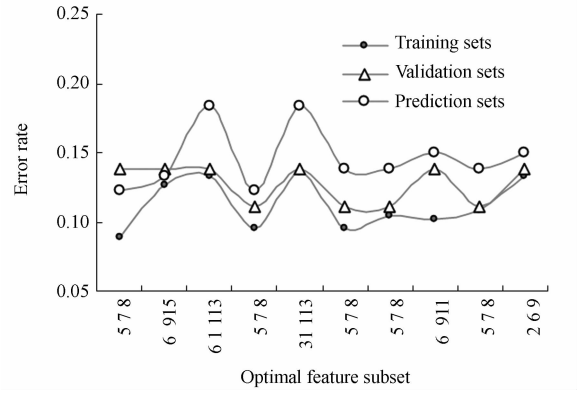


图 4　最优特征子集在 $l=3$ 处的误分率

Fig. 4　Error rate of optimal feature subset at $l=3$

5　FH-W 特征选择

5.1　FH-W 算法

FH-W 算法，在过滤器(Filter)下启发式搜索(Heuristic)特征子集并基于封装器(Wrapper)停止搜索，它执行效率高、分类性能不确定，是一种比较实用的方法，算法可描述为：

步骤 1：假设

(1) Subset_0 为 m 维特征集 Data_m 的 0 维特征子集

$Data_m = Data(Feature_1, Feature_2, \cdots, Feature_m)$；

步骤 2：过滤器下启发式搜索

For $l=1$ to m

(2) 从 Subset_0 开始，启发式搜索类可分性测量值(J)最大的最优 l 维特征子集($\text{Subset}_{\text{best\_}l}$)

$Subset_{\text{best\_}l} = Heuristic\_Search(Data_m, Subset_0, J)_{(J=\max)}$

End；

步骤 3：停止条件

For $l=1$ to m

(3) 用 $\text{Subset}_{\text{best\_}l}$ 在训练集(Train Data)上训练 Bayes 分类器($\text{Classifier}_{\text{best\_}l}$)

$Classifier_{\text{best\_}l} = Build(\text{Train Data}, \text{Subset}_{\text{best\_}l})$，

(4) 计算 $\text{Classifier}_{\text{best\_}l}$ 在验证集(Validate Data)上的误分率($\text{Error}_{\text{best\_}l}$)

$Error_{\text{best\_}l} = Test(ValidateData, Classifier_{\text{best\_}l})$

End，

(5) 以尽可能小的子集容量(l)获取 m 个 $\text{Error}_{\text{best\_}l}$ 中的极小值($\text{Error}_{\text{best}}$)及相应的最优特征子集($\text{Subset}_{\text{best}}$)

$$Error_{\text{best}} = \text{Min}\{Error_{\text{best\_}l}\}_{(l=1,2,\cdots,m)}$$

$$Subset_{\text{best}} = Error_{\min}\{Subset_{\text{best\_}l}\}_{(l=1,2,\cdots,m)};$$

步骤 4：验证结果

(6) 计算 $\text{Subset}_{\text{best}}$ 在预测集(PredictionData)上的泛化误差($\text{Error}_{\text{generalization}}$)

$Error_{\text{generalization}} = Test(\text{PredictionData}, \text{Subset}_{\text{best}})$.

在第(2)步中，距离测度是所有过滤器评价函数中研究时间最长、理论最完善的准则之一[1]，包括概率距离、类内和类间距离等。类可分性测量值(J)基于类间/类内最大化准则表达多维空间样本间的分布，根据类内散布矩阵($\boldsymbol{S}_w$)、类间散布矩阵($\boldsymbol{S}_b$)和混合散布矩阵($\boldsymbol{S}_m$)测量 l 维特征子集的分类性能，如式 3～式 6 所示。

$$\boldsymbol{S}_w = \sum_{i=1}^{l} \boldsymbol{p}(\omega_i)\boldsymbol{s}_i, \tag{3}$$

$$\boldsymbol{S}_b = \sum_{i=1}^{l} \boldsymbol{p}(\omega_i)(\boldsymbol{u}_i - \boldsymbol{u}_0)(\boldsymbol{u}_i - \boldsymbol{u}_0)^{\mathrm{T}}, \tag{4}$$

$$\boldsymbol{S}_m = \boldsymbol{S}_w + \boldsymbol{S}_b, \tag{5}$$

$$J = \text{Trace}\{\boldsymbol{S}_w^{-1}\boldsymbol{S}_m\}. \tag{6}$$

式中，$\boldsymbol{u}_0$ 为全局均值，Trace 为散布矩阵的迹。

启发式搜索策略有：①将前 l 个 J 值最大的标量特征组合形成 $\text{Subset}_{\text{best\_}l}$，不考虑特征之间的关系；②从一个空集开始，在 $\text{Subset}_{\text{best\_}l}$ 上建立所有 Subset_{l+1}，前向顺序搜索 J 值最大的 $\text{Subset}_{\text{best\_}l+1}$，考虑了特征之间的关系；③从整个特征集开始，在 $\text{Subset}_{\text{best\_}l+1}$ 下建立所有 Subset_l，后向顺序搜索 J 值最大的 $\text{Subset}_{\text{best\_}l}$，充分考虑了特征之间的关系；④前向浮动搜索允许回溯，消除了顺序搜索中的嵌套效应，可获得近似最优解，本文着重讨论①和④。

在第(3)～(5)步中，随着特征子集容量的不断增加($l=1, 2, 3, \cdots, m$)，$\text{Classifier}_{\text{best\_}l}$ 在 ValidateData 上的 $\text{Error}_{\text{best\_}l}$ 经历了逐渐减小至增大的过程，其拐点处产生 $\text{Error}_{\text{best}}$ 和 $\text{Subset}_{\text{best}}$。

5.2　FH-W 实验结果

5.2.1　最优特征组合策略

在各训练集上计算每个特征的类可分性测量值

(J),将前 l 个 J 值最大的特征组合形成最优 l 维特征子集(X_l),l=1, 2, …, 15。基于 10 个训练集的最优 X_l 不尽相同,最优 X_l 是最优 X_{l+1} 的子集(表 3),随着 l=1, 2, …, 15 不断增加,最优 X_l 的平均 J 值在 1.6281～14.9611 之间单调递增,运行总耗时为 0.8 s。

表 3　最优特征组合的最优 X_l

Tab. 3　Optimal X_l based on optimal feature combination

训练集	依 J 值下降的标量特征序号
1	14 15 1 9 6 11 8 4 13 7 12 5 3 10 2
2	14 1 15 9 6 11 8 4 13 7 12 5 3 10 2
3	14 1 15 9 6 11 8 4 13 7 12 5 3 10 2
4	14 1 15 9 6 11 8 4 13 7 5 12 3 10 2
5	14 1 15 9 6 11 8 4 13 7 12 5 3 10 2
6	14 1 15 9 11 6 4 8 7 12 5 13 3 10 2
7	14 1 15 9 6 11 8 4 13 7 12 5 3 10 2
8	14 1 15 6 9 11 8 13 4 5 7 12 3 10 2
9	14 15 1 9 6 11 8 4 13 7 12 5 3 2 10
10	14 1 15 9 6 11 8 13 4 7 12 5 3 10 2

在各训练集上,用最优 X_l 建立 Bayes 分类器,将训练集、验证集和预测集样本划分为 2 类。随着 l 的不断增加(l=1, 2, …, 15),最优 X_l 在验证集上的平均误分率逐渐下降至 l=3(14.44%±7.61%)以后趋于平缓(图 5),最优 X_l 在 l=3 处的训练集、预测集平均误分率分别为 13.77%±0.8%、14.72%±0.39%,最优特征子集容量 l=3。

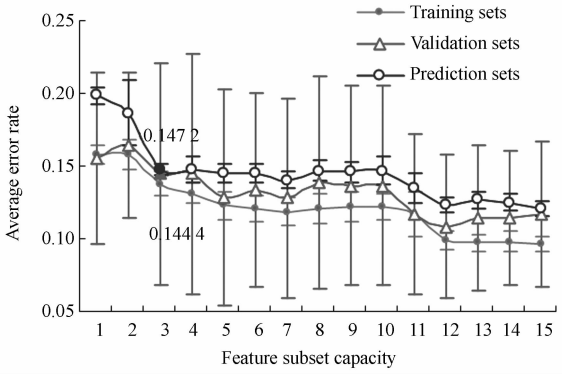


图 5　最优特征组合策略下 Bayes 分类器在数据集上的平均误分率随着特征子集容量而变化

Fig. 5　Average error rate of Bayes classifier on data set varies with feature subset capacity by using optimal feature combination

最优特征子集(表 3)在训练集、验证集和预测集上的平均识别率不太稳定(86.23%＞85.56%＞85.28%),在 10 个训练集上获取的{x_{14}, x_1, x_{15}}描述了籽棉的外围结构和整体形态,相关系数＜0.9,分类性能不如 WE-W 算法。

5.2.2　浮动搜索策略

在各训练集上计算 $X_1=\{x_i\}$ 的 J 值,i=1, 2, …, 15,从中选择 J 值最大的最优 X_1,并将最优 X_1 与特征 x_j 组成 $X_2=X_1\cap\{x_j\}$,j=1, 2, …, 15,$j\neq i$,再从中选择 J 值最大的最优 X_2;依此类推,在最优 X_l 的基础上前向浮动搜索最优 X_{l+1},l=2, 3, …, 14。基于 10 个训练集的最优 X_l 不尽相同,由于回溯的关系,最优 X_l 不一定是最优 X_{l+1} 的子集(表 4),随着 l 的不断增加(l=1, 2, …, 15),最优 X_l 的平均 J 值在 1.628 1～14.961 1之间单调递增,同比大于最优特征组合策略,程序运行耗时为 0.76 s。

表 4　浮动搜索的最优 X_l

Tab. 4　Optimal X_l based on floating search

l	训练集 1	训练集 2	训练集 3	训练集 4	训练集 5
1	14	14	14	14	14
2	14 15	14 15	15 1	14 15	14 15
3	14 15 1	14 15 1	15 1 14	14 15 1	14 15 1
4	14 15 1 9	14 15 1 9	15 1 14 9	14 15 1 9	14 15 1 9
5	14 15 1 9 5	14 15 1 9 5	15 1 14 9 5	14 15 1 9 5	14 15 1 9 5
…	…	…	…	…	…
l	训练集 6	训练集 7	训练集 8	训练集 9	训练集 10
1	14	14	14	14	14
2	14 15	14 15	15 1	14 15	14 15
3	15 5 1	14 15 1	15 1 14	14 15 1	14 15 1
4	15 1 14 9	14 15 1 9	15 1 14 9	14 15 1 9	14 15 1 9
5	15 1 14 9 5	14 15 1 9 5	15 1 14 9 5	14 15 1 9 5	14 15 1 9 5
…	…	…	…	…	…

在各训练集上,用最优 X_l 建立 Bayes 分类器,将训练集、验证集和预测集样本划分为 2 类。随着 l 的不断增加(l=1, 2, …, 15),最优 X_l 在验证集上的平均误分率逐渐下降至 l=3(14.44%±7.61%)以后趋于平缓(图 6),最优 X_l 在 l=3 处的训练集、预测集平均误分率分别为 14.07%±1.45%、14.78%

±0.39%，最优特征子集容量 $l=3$。

最优特征子集（表 4）在训练集、验证集和预测集上的平均识别率不够稳定（85.93%＞85.56%＞85.22%），在 9 个训练集上获取的{x_{14}，x_1，x_{15}}同5.2.1节。

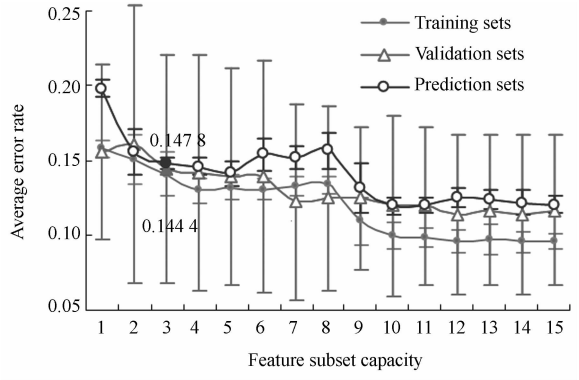


图 6　浮动搜索下 Bayes 分类器在数据集上的平均误分率随着特征子集容量而变化

Fig. 6　Average error rate of Bayes classifier on data set varies with feature subset capacity by using floating search

6　结　论

面向田间籽棉成熟度判别的特征选择问题属于 NP 难题，为了兼顾执行效率、分类性能以及特征子集的容量，本文以封装器的最优解为停止条件，比较了“穷举搜索＋封装器”、“启发式搜索＋过滤器”的匹配结果，结论如下：

（1）在户外光照条件下采集田间籽棉正面图像样本，所提取的描述籽棉形态信息的特征集能够区分籽棉的成熟度。

（2）“穷举搜索匹配封装器”的方法执行效率低、分类性能和稳定性好，所选择的 3 个最优特征描述了棉芯的结构，相关系数＜0.9，在新的数据集上获得了 85.39%的平均识别率。

（3）“启发式搜索匹配过滤器”的执行效率高、分类性能好，所选择的 3 个最优特征描述了籽棉的整体形态和外围结构，相关系数＜0.9，在新的数据集上获得了 85.28%（最优特征组合）和 85.22%（浮动搜索）的平均识别率。

本文的研究对判别田间籽棉成熟度时如何兼顾准确率和实时性具有参考意义。

参考文献：

[1] 孙伟艳. 模式分类中特征选择问题的研究[D]. 哈尔滨：哈尔滨理工大学，2009.
SUN W Y. *Research on feature selection for pattern classification* [D]. Harbin: Harbin University of Science and Technology, 2009. (in Chinese)

[2] NG W W Y, YEUNG D S, FIRTH M, *et al.*. Feature selection using localized generalization error for supervised classification problems using RBFNN [J]. *Pattern Recognition*, 2008, 41(12): 3706-3719.

[3] RUIZ R, RIQUELME J C, AGUILAR-RUIZ J S. Incremental wrapper-based gene selection from microarray data for cancer classification [J]. *Pattern Recognition*, 2006, 39(12): 2383-2392.

[4] LIANG J N, YANG S, WINSTANLEY A. Invariant optimal feature selection: A distance discriminant and feature ranking based solution [J]. *Pattern Recognition*, 2008(5), 41: 1429-1439.

[5] LIU Y, ZHENG Y F. FS_SFS: A novel feature selection method for support vector machines [J]. *Pattern Recognition*, 2006, 39(7): 1333-1345.

[6] SEBBAN M, NOCK R. A hybrid filter/wrapper approach of feature selection using information theory [J]. *Pattern Recognition*, 2002, 35(4): 835-846.

[7] DONG M, KOTHARI R. Feature subset selection using a new definition of classifiability [J]. *Pattern Recognition Letters*, 2003, 24(9-10): 1215-1225.

[8] HUA J P, TEMBE W D, DOUGHERTY E R. Performance of feature-selection methods in the classification of high-dimension data [J]. *Pattern Recognition*, 2009, 42(3): 409-424.

[9] NAKARIYAKUL S, CASSASENT D P. An improvement on floating search algorithms for feature subset selection [J]. *Pattern Recognition*, 2009, 42(9): 1932-1940.

[10] 李先锋，朱伟兴，纪滨，等. 基于图像处理和蚁群优化的形状特征选择与杂草识别[J]. 农业工程学报，2010，26(10)：178-182.
LI X F, ZHU W X, JI B, *et al.*. Shape feature selection and weed recognition based on image processing and ant colony optimization [J]. *Transactions of the* CSAE, 2010, 26(10): 178-182. (in Chinese)

[11] 董岩,张涛,李文明,等. 机载立体测绘相机滚转轴伺服系统的辨识与设计[J]. 光学 精密工程,2011, 19(7): 1580-1587.
DONG Y, ZHANG T, LI W M, *et al*.. Identification and design of roll axis servo system in airborne stereo mapping camera[J]. *Opt. Precision Eng.*, 2011, 19(7): 1580-1587. (in Chinese)

[12] 董超,田联房. 最速上升关联向量机高光谱影像分类[J]. 光学 精密工程,2012,20(6): 1398-1405.
DONG CH, TIAN L F. Hyperspectral image classification by steepest ascent relevance vector machine[J]. *Opt. Precision Eng.*, 2012,20(6):1398-1405. (in Chinese)

[13] WEBB A R. 统计模式识别[M]. 第二版. 王萍 等译. 北京:电子工业出版社, 2004.
WEBB A R. *Statistical Pattern Recognition*[M]. 2nd ed. Wang Ping *et al*. translation. Beijing: Publishing House of Electronics Industry, 2004. (in Chinese)

[14] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifiers[J]. *Machine Learning*, 1997, 29(2-3):131-163.

[15] 宋枫溪,高秀梅,刘树海, 等. 统计模式识别中的维数削减与低损降维[J]. 计算机学报,2005, 28(11): 1915-1922.
SONG F X, GAO X M, LIU SH H, *et al*.. Dimensionality reduction in statistical pattern recognition and low loss dimensional ity reduction[J]. *Chinese Journal of Computers*, 2005, 28(11): 1915-1922. (in Chinese)

作者简介:

王　玲(1966—),女,江西南昌人,博士,副教授,1988年于南京航空航天大学获得学士学位,2006年、2009年于南京农业大学分别获得硕士、博士学位,主要从事图像处理与模式识别技术方面的研究。E-mail: Lingw@njau.edu.cn

刘德营(1963—),浙江义乌人,博士,副教授,1986年于浙江大学获得学士学位,2010年于南京农业大学获得博士学位,现为南京农业大学工学院电气工程系教师、实验中心主任,主要从事农业电子与信息技术方面的研究。E-mail: dyliu@njau.edu.cn

姬长英(1955—),男,山东人,博士生导师,教授,1982年、1984年于江苏大学分别获得学士、硕士学位,1994年于南京农业大学分别获得博士学位,现为南京农业大学工学院农机化系主任,主要从事农业机器人、水田土壤流变机理等研究。E-mail: chyji@njau.edu.cn