马蹄蟹追随者状态数据分析

姓名：隆征帆　数学与统计学院　学号：2013012422

摘要：本文综合运用统计图表、假设检验、广义线性模型（logit、probit、AIClogit、AICprobit，BIClogit、BICprobit）、机器学习分类模型（决策树、bagging、随机森林、adaboost）等对马蹄蟹追随者状态作了简要分析。同时利用 TPR（true positive rate）、FPR（false positive rate）、Youden 指数（TPR-FTR）、错分率等四项统计指标，并结合 k 折交叉验证方法（$k = 5,6,\cdots,10$）对以上 10 种模型进行了选择。本文最终从 Youden 指数和错分率两方面综合考虑选取了随机森林模型，利用全部数据给出了该模型的建模结果及样本内预测精度和犯错概率，并对整个分析过程作了总的评价。

关键词:列联表　散点图　箱线图　Q-Q 图　t 检验　logit　probit　AIC　BIC　决策树　bagging　随机森林　adaboost　TPR　FPR　Youden　错分率　k 折交叉验证

一、研究目的

研究马蹄蟹体表特征与其有无追随者的关系具有十分重要的意义，它将帮助科研人员尤其是海洋动物行为领域的科研人员们更好地认识马蹄蟹的行为特征与其生理特征间的关系。而如果能够通过马蹄蟹体表特征较为准确地得知其有无追随者的行为状态，那么人类对马蹄蟹行为特征的大规模分析将有可能成为现实。进一步，这样的科学分析方法还可以向其他类型生物推广。可以想象，到那时，动物行为领域的研究将是怎样一番前所未有的新气象。本文就是基于这样的愿景，尝试通过构建马蹄蟹体表特征与其有无追随者的统计分析模型，以期能够通过马蹄蟹体表特征较为准确地获知其追随者的有无状态。

二、数据来源和相关说明

我们的数据包含 173 个完整观测，4 个自变量和 1 个因变量。各变量含义及类型如表 2-1 所示：

表 2-1 变量含义及类型

变量	变量含义	变量类型
自变量	颜色(color)	4 水平定性变量
	背鳍(spine)	3 水平定性变量
	宽度(width)	定量变量
	重量(weight)	定量变量
因变量	有无追随者(y)	2 水平定性变量

其中颜色 4 水平分别为：1（偏浅）、2（中等）、3（偏深）、4（深），背鳍 3 水平分别为：1（都好）、2（有一个破）、3（都破），因变量 2 水平分别为：0（无）、1（有）。

三、变量间关系的初步分析

考虑到自变量中既有定性变量，也有定量变量，而因变量为二水平定性变量。因此我们采用 Pearson 卡方独立性检验探讨了各定性自变量与定性因变量的关系，采用散点图及 Pearson 相关系数检验探讨了两个定量自变量间的关系，并采用箱线图和 t 检验探讨了各定量自变量与二水平定性因变量的关系。另外，由于 Pearson 相关系数检验以及 t 检验均依赖于对总体的正态性假定，因此在对数据执行上述两个检验时，我们还对两个定量自变量进行了正态性检验。

表 3-1 是颜色(color)与有无追随者(y)两个定性变量构成的 2×4 二维列联表：

表 3-1

颜　色 / 有无追随者	1（偏浅）	2（中等）	3（偏深）	4（深）
0（无）	3	26	18	15
1（有）	9	69	26	7

由 R 对表 3-1 进行 Pearson 卡方独立性检验得到的结果如图 3-2 所示：

```
        Pearson's Chi-squared test

data:  color.y
X-squared = 14.078, df = 3, p-value = 0.002802
```

图 3-2

综合表 3-1 和图 3-2 我们可以得到两条结论：①在允许 0.28%的犯错概率下，马蹄蟹颜色与其是否有追随者之间存在相互关系；②颜色偏浅或中等的马蹄蟹较颜色偏深或深的马蹄蟹更有可能拥有追随者。

表 3-3 是背鳍(spine)与有无追随者(y)两个定性变量构成的 2×3 二维列联表：

表 3-3

背　鳍 / 有无追随者	1（都好）	2（有一个破）	3（都破）
0（无）	11	8	43
1（有）	26	7	78

由 R 对表 3-3 进行 Pearson 卡方独立性检验得到的结果如图 3-4 所示：

```
        Pearson's Chi-squared test

data:  spine.y
X-squared = 2.6018, df = 2, p-value = 0.2723
```

图 3-4

综合表 3-3 和图 3-4 我们可以得到两条结论：①马蹄蟹颜色与其背鳍破损程度间不存在显著的相互关系，但并不表明它们之间一定不存在相互关系；②背鳍都好的马蹄蟹较背鳍有一个破的马蹄蟹可能有更高机率拥有追随者，但即使有也不会高太多。

图 3-5 是由 R 绘制的 3 张图形的组合，从左至右依次为：马蹄蟹宽度(width)与重量(weight)的散点图、无追随者的马蹄蟹宽度与有追随者的马蹄蟹宽度的箱线图、无追随者的马蹄蟹重量与有追随者的马蹄蟹重量的箱线图。

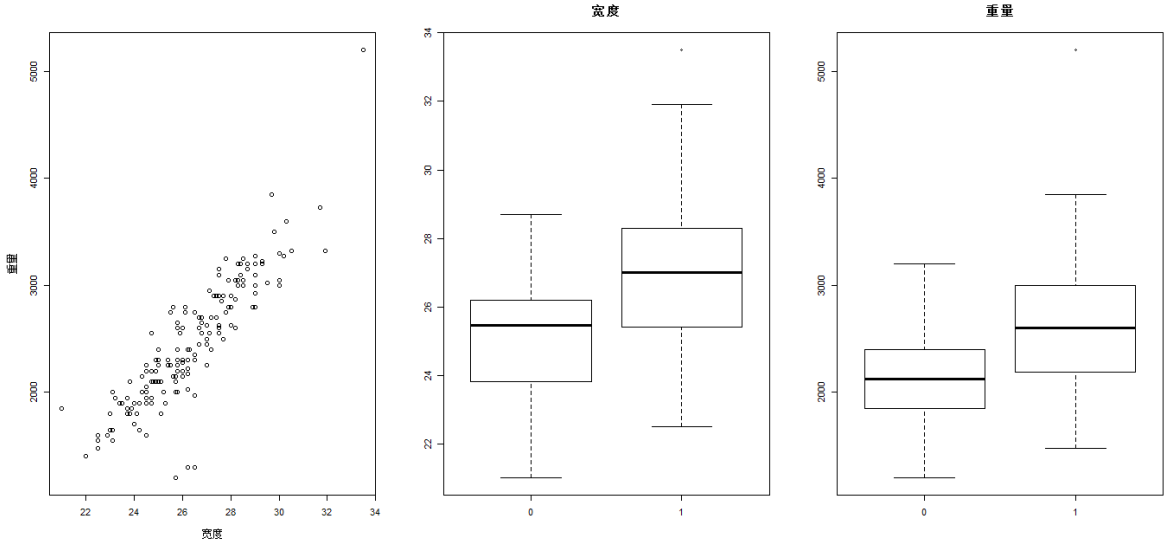

图 3-5

从图 3-5 中我们可以得到如下直观性结论：①马蹄蟹宽度和重量两个变量间存在明显线性关系；②无追随者的马蹄蟹宽度的平均水平明显低于有追随者的马蹄蟹；③无追随者的马蹄蟹重量的平均水平明显低于有追随者的马蹄蟹。为了验证上述结论，我们分别进行了 Pearson 相关系数检验和 t 检验，并在进行这些检验之前对宽度和重量两个变量做了正态性检验。

图 3-6 是由 SPSS 软件绘制的两个定量自变量（宽度、重量）的 Q-Q 图：

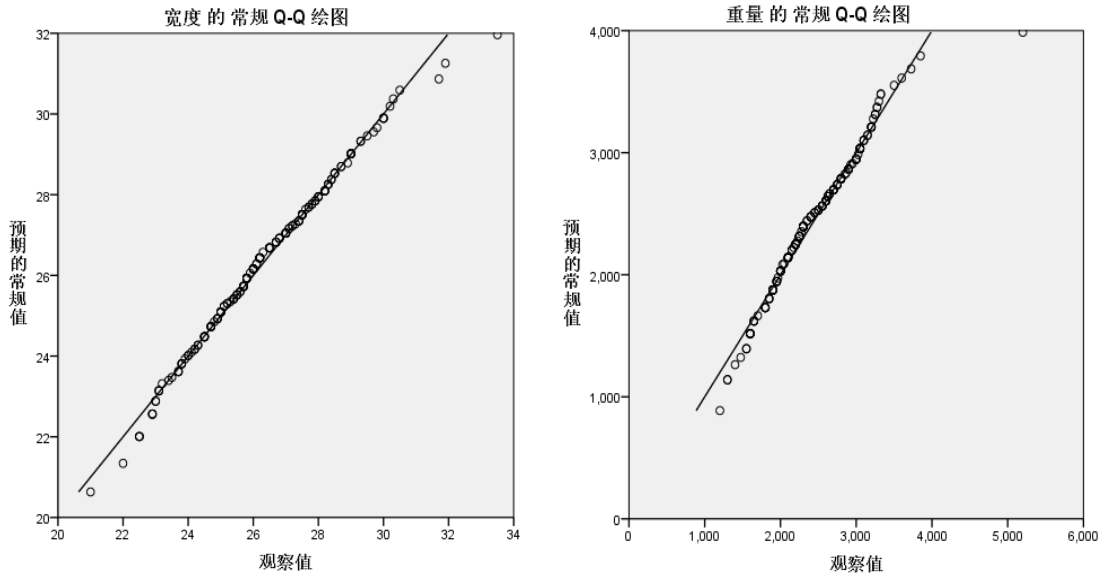

图 3-6

虽然我们永远无法证明数据服从某个特定分布，但从 Q-Q 图中可以看出：至少两个定量自变量的真实分布不会偏离正态分布太远，因而我们可以采用基于正态总体的 t 检验。

图 3-7 从上至下依次为 SPSS 软件给出的 Pearson 相关系数检验结果、SPSS 软件给出的无追随者马蹄

蟹宽度/重量均值与有追随者马蹄蟹宽度/重量均值 t 检验结果。

相关性

		宽度	重量
宽度	Pearson 相关性	1	.887**
	显著性（双尾）		.000
	N	173	173
重量	Pearson 相关性	.887**	1
	显著性（双尾）	.000	
	N	173	173

**. 在置信度（双测）为 0.01 时，相关性是显著的。

独立样本检验

		列文方差相等性检验		平均值相等性的 t 检验						
									差值的 95% 置信区间	
		F	显著性	t	自由度	显著性（双尾）	平均差	标准误差差值	下限	上限
宽度	已假设方差齐性	2.820	.095	-5.731	171	.000	-1.76037	.30716	-2.36669	-1.15406
	未假设方差齐性			-6.082	149.243	.000	-1.76037	.28942	-2.33227	-1.18848
重量	已假设方差齐性	4.656	.032	-5.492	171	.000	-464.57178	84.59737	-631.56141	-297.58216
	未假设方差齐性			-5.888	153.027	.000	-464.57178	78.90218	-620.44995	-308.69362

图 3-7

图 3-7 的检验结果有力地验证了上面我们得到的关于宽度和重量两个变量的 3 条直观性结论。（注：虽然上面的 t 检验是双边 t 检验，但在给定的显著性水平下，双边 t 检验通过意味着单边 t 检验一定通过）

我们将上述变量间关系的初步分析结果汇总如下：

①在允许 0.28%的犯错概率下，马蹄蟹颜色与其是否有追随者之间存在相互关系；

②颜色偏浅或中等的马蹄蟹较颜色偏深或深的马蹄蟹更有可能拥有追随者；

③马蹄蟹颜色与其背鳍破损程度间不存在显著的相互关系，但并不表明它们之间一定不存在相互关系；

④背鳍都好的马蹄蟹较背鳍有一个破的马蹄蟹可能有更高机率拥有追随者，但即使有也不会高太多；

⑤马蹄蟹宽度和重量两个变量间存在显著线性关系；

⑥无追随者的马蹄蟹宽度的平均水平显著低于有追随者的马蹄蟹；

⑦无追随者的马蹄蟹重量的平均水平显著低于有追随者的马蹄蟹。

四、模型选择

根据变量类型特点，我们选择 10 种模型对上述数据进行了分析。表 4-1 至表 4-6 分别是利用 5 折、6 折、7 折、8 折、9 折、10 折交叉验证得到的 10 种模型 4 项预测效果指标均值的统计表（注：所有统计表均按 Youden 指数降序排列，表中排名前 3 的模型均加粗表示）。

表 4-1

5折交叉验证结果				
	1-TPR	FPR	Youden	错分率
决策树	0.2268	0.4831	0.2901	0.3176
随机森林	0.2202	0.5057	0.2741	0.3235
Adaboost	0.1743	0.5774	0.2483	0.3176
Bagging	0.2572	0.5400	0.2029	0.3588
logit	0.0875	0.7144	0.1981	0.3294
AIClogit	0.0602	0.8226	0.1173	0.3529
BIClogit	0.0687	0.8251	0.1062	0.3588
probit	0.0296	0.9200	0.0504	0.3765
AICprobit	0.0190	0.9733	0.0076	0.3882
BICprobit	0.0190	0.9867	-0.0057	0.3941

表 4-2

6折交叉验证结果				
	1-TPR	FPR	Youden	错分率
随机森林	0.2167	0.5286	0.2547	0.3333
Bagging	0.2575	0.5031	0.2394	0.3571
决策树	0.2556	0.5247	0.2197	0.3452
logit	0.0989	0.7027	0.1984	0.2976
Adaboost	0.1724	0.6328	0.1948	0.3333
probit	0.0552	0.8056	0.1392	0.3036
AIClogit	0.1310	0.7311	0.1379	0.3333
BIClogit	0.0807	0.8344	0.0849	0.3512
AICprobit	0.0631	0.8740	0.0629	0.3274
BICprobit	0.0238	0.9792	-0.0030	0.3631

表 4-3

7折交叉验证结果				
	1-TPR	FPR	Youden	错分率
决策树	0.2285	0.4514	0.3201	0.3095
随机森林	0.1957	0.5035	0.3008	0.3095
Bagging	0.2127	0.5115	0.2758	0.3214
Adaboost	0.1400	0.5901	0.2699	0.2917
logit	0.1324	0.6251	0.2425	0.3095
AIClogit	0.1324	0.6715	0.1961	0.3274
BICprobit	0.0751	0.7338	0.1911	0.3155
BIClogit	0.1419	0.6689	0.1891	0.3333
probit	0.0941	0.7318	0.1741	0.3274
AICprobit	0.0846	0.7477	0.1677	0.3274

表 4-4

8折交叉验证结果				
	1-TPR	FPR	Youden	错分率
AIClogit	0.0950	0.4052	0.4998	0.2024
BIClogit	0.0449	0.5199	0.4352	0.2024
logit	0.1022	0.4793	0.4185	0.2321
AICprobit	0.0860	0.5545	0.3595	0.2440
probit	0.0777	0.5827	0.3395	0.2500
决策树	0.2173	0.4756	0.3071	0.2976
BICprobit	0.0449	0.6911	0.2640	0.2560
随机森林	0.2373	0.5057	0.2570	0.3333
Adaboost	0.1849	0.6242	0.1909	0.3333
Bagging	0.2761	0.5820	0.1419	0.3869

表 4-5

9折交叉验证结果				
	1-TPR	FPR	Youden	错分率
随机森林	0.2090	0.4291	0.3619	0.3041
决策树	0.2198	0.4581	0.3221	0.3158
Bagging	0.2366	0.4728	0.2906	0.3275
logit	0.2920	0.4566	0.2514	0.3392
probit	0.2834	0.4788	0.2377	0.3392
BICprobit	0.2421	0.5280	0.2299	0.3333
Adaboost	0.1550	0.6230	0.2220	0.3216
BIClogit	0.2421	0.5586	0.1993	0.3450
AICprobit	0.2408	0.5780	0.1812	0.3509
AIClogit	0.2725	0.5697	0.1578	0.3684

表 4-6

10折交叉验证结果				
	1-TPR	FPR	Youden	错分率
logit	0.3256	0.3456	0.3288	0.3294
probit	0.3347	0.3456	0.3197	0.3353
随机森林	0.1787	0.5219	0.2994	0.3176
BICprobit	0.2651	0.4597	0.2752	0.3412
AIClogit	0.2616	0.4700	0.2683	0.3412
BIClogit	0.2901	0.4597	0.2502	0.3529
AICprobit	0.2630	0.4982	0.2387	0.3529
Adaboost	0.1428	0.6347	0.2225	0.3176
Bagging	0.2470	0.5356	0.2174	0.3647
决策树	0.2391	0.5586	0.2022	0.3588

表 4-1 至表 4-6 各项指标的含义说明如下：

①1-TPR：TPR 全称 True Positive Rate，指 y 实际为 1，预测为 1 的概率，因此 1-TPR 反映犯“实际为 1，预测为 0”这一错误的概率，这里不妨称其为犯第一类错误的概率；

②FPR：FPR 全称 False Positive Rate，指 y 实际为 0，预测为 1 的概率，因此 FPR 反映犯“实际为 0，预测为 1”这一错误的概率，这里不妨称其为犯第二类错误的概率；

③Youden：Youden 指数是 TPR 与 FPR 之差，它与犯上述两种错误的总概率成反比，用来反映犯两类错误的总大小；

④错分率：指预测错误的个体占总个体的比例，用来反映总的预测精度。

综合表 4-1 至表 4-6 来看，从控制两类犯错概率的角度出发，随机森林模型是一个相对较好的选择。另外随机森林模型在建模过程中对自变量进行了选择，其错分率相较其它模型而言也不高。因此本文后面

将对全部数据构建随机森林模型，并给出各变量在随机森林模型的重要程度以及该模型的样本内预测精度。关于表 4-1 至表 4-6，我们还有如下以下几点需要说明：

①由于使用了交叉验证方法，因此严格说来，上述 4 个指标前都应加上“平均”二字；

②6 个广义线性模型（logit 回归全模型、probit 全模型，logit 回归 AIC 模型、probit 回归 AIC 模型，logit 回归 BIC 模型、probit 回归 BIC 模型）的预测都涉及到阈值的确定。表中给出的这 6 个模型的预测指标，都是在确定好各模型最优阈值的前提下得到的；

③6 个广义线性模型最优阈值的确定都是基于最大化平均 Youden 指数（最小化平均两类错误总量）实现的，所谓平均 Younden 指数之“平均”，则是因为最优阈值的确定过程使用了 10 折交叉验证方法；

④由于编程时设定了统一的随机种子点，对不同模型使用交叉验证时，交叉验证的各折样本完全一致，因此不同模型间的各项指标具备可比性。

五、数据建模

基于交叉验证的结果，我们从 10 种模型中选择了随机森林模型，下面我们将利用全部数据构建随机森林模型并给出模型的相关分析结果。

图 5-1 是利用 R 对全部 173 个观测构建随机森林模型的结果

```
Call:
 randomForest(formula = 有无追随者 ~ ., data = shark, importance = T)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 32.95%
Confusion matrix:
   无 有 class.error
无 29 33   0.5322581
有 24 87   0.2162162
```

图 5-0-1

随机森林模型利用 Bootstrap 方法对原始样本进行了 500 次重抽样，然后对每个 Bootstrap 样本进行变量选择并构建决策树，因而总共构建了 500 棵决策树，随机森林模型的最终结果即为这 500 棵决策树的算术平均值。这里，每次重抽样都会有部分原始观测未进入 Bootstrap 样本，这些样本点被称为 Out-of-bag 样本点，它们将自动成为测试集用以估计模型的平均预测误差。图 5-1 中 error rate 的 OOB 估计就是由这些样本点计算得到的。从图 5-1 中我们得知：由 500 个测试集得到的随机森林模型的平均错分率约为 32.95%，平均犯第一类错误概率(1-TPR)约为 27.5%，平均犯第二类错误概率(FPR)约为 45.3%。

表 5-2 展示了利用随机森林模型对训练集的预测结果与实际观测结果的混淆矩阵。

表 5-2

实际观测 / 预测结果	0（无追随者）	1（有追随者）
0（无追随者）	61	2
1（有追随者）	1	109

由表 5-2 可知随机森林模型的样本内预测精度约为 98.3%，样本内犯第一类错误概率约为 1.8%，样本内犯第二类错误概率约为 1.6%.

图 5-3 展示了各自变量在随机森林模型中的相对重要程度

	无	有	MeanDecreaseAccuracy	MeanDecreaseGini
宽度	9.756889	4.2476352	10.041103	30.615978
重量	3.665955	0.6990094	2.892078	27.310702
颜色	12.436065	7.6483263	14.484690	10.232287
背鳍	3.913056	-3.6791328	-0.341523	4.889972

图 5-3

图 5-3 告诉我们各自变量在随机森林模型中的相对重要程度依次为：颜色>宽度>重量>背鳍，并且去掉背鳍这个变量不仅不会降低模型预测精度，反而会提升模型预测精度。

六、结论和建议

虽然我们利用随机森林模型得到的样本内预测精度高达 98%，但这说明不了模型的真实预测效果。说明模型真实预测效果的指标应是其在测试集上的预测精度，但我们可以看到随机森林模型在测试集的平均错分率高达 32.95%。事实上，从上面的 k 折交叉验证结果我们可以知道，上述 10 种模型的预测精度均低于 80%，且普遍性地低于 70%，而这样的预测精度显然难以令人满意。造成这一局面的原因，我们分析认为可能有如下几点：

①还存在更加适合该数据集的模型，而这些模型未被我们考虑；

②所收集的观察数据不足以刻画马蹄蟹追随者的状态，还存在更多影响马蹄蟹追随者状态的变量未在我们考虑范围之内。

③这些变量信息已足以刻画马蹄蟹追随者的状态，但数据存在较大误差，影响了分析结果和预测效果。

为此我们建议重新考虑影响马蹄蟹追随者状态的可能因素，收集更多新的数据，并考虑更多的可能模型，以期实现更为准确的分析与预测。

上述 10 种模型的分析预测效果虽然都不甚理想，但通过上述分析我们依然得到了一些具备一定价值的结论，这些结论有：

①马蹄蟹的宽度和重量是影响马蹄蟹追随者状态的重要因素，一般而言，较宽和较重的马蹄蟹容易拥有追随者。

②马蹄蟹的颜色也是影响马蹄蟹追随者状态的重要因素，一般而言，颜色较浅的马蹄蟹容易拥有追随者。

③马蹄蟹背鳍的破损程度对马蹄蟹追随者状态的影响不大，但它仍有可能与其它因素交互影响马蹄蟹追随者的状态，这种交互影响作用仍有可能是显著的。