学校代码：10200　　学号：2013012422

东北师范大学

本科毕业论文

k 折交叉验证折数确定方法初探

学生姓名：隆征帆

指导教师：马文卿　副教授

所在学院：数学与统计学院

所学专业：统计学

中国·长春

2017 年 4 月

东北师范大学本科生毕业论文扉页

独 创 性 声 明

本人郑重声明：所提交的毕业论文是本人在导师指导下独立进行研究工作所取得的成果。据我所知，除了特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。对本人的研究作出重要贡献的个人和集体，均已在文中做了明确的说明。本声明的法律结果由本人承担。

毕业论文作者签名：_____ 日期：_____

摘要

本文首先介绍了模型选择及交叉验证方法，然后通过马蹄蟹案例的分析探讨了折数在 k 折交叉验证中的重要性，接着通过实际案例初步研究了不同样本量下 k 折交叉验证预测效果稳定性与折数 k 的关系。在这一关系的研究中，得到一个重要规律：当 k 位于 $[n/i], i = 1,2,\cdots$ 附近时，k 折交叉验证预测效果标准差有局部极小值，并 i 越小，局部极小值通常越小。基于这样的规律，我们提出 k 折交叉验证折数的粗略确定方法，并用实际案例检验了这一确定方法。

关键字: 模型选择　交叉验证　k 折交叉验证　折数

Abstract

In this paper, we first introduce the model selection and cross validation method, then discuss the importance of determination to folds in the k-fold cross validation through the analysis of the shark case, and then study the relationship between the number of fold k and the test results' stability of k-fold cross validation for different sample size of dataset through the actual case. In this study of the relationship, we get an important rule: when the folds locate around $[n/i], i = 1,2,\cdots$, there will be a local minimum for the standard deviation of k-fold cross validation test-error, and usually the smaller the i, the smaller the local minimum is. Based on this rule, we propose a rough method for verifying the folds of the k-fold cross validation and verify the method with the actual case.

Key words: Model selection cross validation k-fold cross validation folds

目 录

1 模型选择与交叉验证

千百年来，人们可能做过许许多多的美梦，期待着科学的发展能够帮助人们不断逼近世间万物的真相。今天，随着互联网尤其是移动互联网的发展，信息爆炸式增长的时代已然来临，各种各样庞大的数据集正如潮水一般向我们涌来。而统计学，作为一门以数据为中心的学科，正在义不容辞地承担起这个“帮助人们逼近世间万物真相”的重任。在统计学家的世界里，有一件东西很重要，那就是建立模型（简称建模）。建模之于统计学，就像筷子之于中国人的饮食、刀叉之于英国人的饮食一样，不可或缺。统计学家正是用基于数据构建的模型来试图逼近世间万物真相的。

在与 *Norman R. Draper* 合著的《*Empirical Model-Building and Response Surfaces*》一书中，英国著名统计学家，*Box-Jenkins* 模型（即 *ARIMA* 模型）及 *Box-Cox* 变换的提出者 *George Edward Pelham Box* 这样写道[1]：虽然从本质上来说，所有的模型都是错的，但它们中的一些是有用的（原文：*essentially, all models are wrong, but some are useful*）。这一发人深省的话语告诉人们：统计推断所用模型是基于不完全信息确定的，而信息的不完全使得错误的发生在所难免，然而这并不代表统计建模是毫无意义的骗人把戏，相反，只要我们能将错误控制在一定范围之内，那么这个模型对于实际应用而言就是有用的。

事实上，信息的不完全正是统计推断赖以存在的前提与基础。试想，如果一个人拥有了未卜先知的能力，他能够获得关于未来的所有信息，那么当荧屏上出现诸如天气预报、股价预测等的画面时，他会有怎样的反应呢？笔者觉得，这位天神下凡般的人物可能会在一旁暗自嘲笑道：这帮愚蠢的人类，真是朽木不可雕也。虽然信息的不完全使得统计推断有了它存在的意义与价值，但与此同时，它也使得不同的人甚至同一个人在面对相同问题时，可能建立多种不同的模型。于是，一个现实而重要的问题就摆在了人们的面前，它就是模型的选择。

1.1 模型选择

人们似乎期待着模型选择能够实现这样的功能，即在所有可能模型中，选择出与真实情况最为接近，或换句话说，选择出误差或损失最小的模型。然而遗憾的是，这几乎是不可能的。问题在于：我们无法穷尽所有可能模型，我们并不知道所谓真实情况，我们也没有关于接近度或误差的唯一衡量标准。于是，模型选择这样的问题，就像模型建立这样的问题一样，成了一个仁者见仁、智者见智的问题。借用 *George Edward Pelham Box* 的话说就是：从本质而言，所有的模型选择方法也是错的，但它们中的一些是有用的。

到目前为止，人们发展了众多的模型选择方法，这些方法大体上可以概括为如下三种（各方法详情可参考文献[2]和文献[3]）：

一、选择样本内预测损失最小的模型：设 $(Y_1,\boldsymbol{X}_1),(Y_2,\boldsymbol{X}_2),\cdots,(Y_n,\boldsymbol{X}_n)$ 为包含 p 个自变量、1 个因变量的 n 个样品的数据，其中 $\boldsymbol{X}_i=(X_{i1},X_{i2},\cdots,X_{ip})',i=1,2,\cdots,n$，并设 $f_\alpha(\boldsymbol{X})=f_\alpha(X_1,X_2,\cdots,X_p),\alpha=1,2,\cdots,m$ 为由这些数据建立的 m 个模型，则该模型选择方法将会选择使得

$$E_{内}(\alpha)=\frac{1}{n}\sum_{i=1}^{n}loss(f_{\alpha}(\boldsymbol{X}_i),Y_i)$$

达到最小的 α 对应的模型。不妨记这个 α 为 α^*，则模型 $f_{\alpha^*}(\boldsymbol{X})$ 即为样本内预测损失最小的模型。这里 $loss(\cdot)$ 是损失函数，常见的损失函数有：$loss(x,y)=(x-y)^2$、$loss(x,y)=|x-y|$、$loss(x,y)=I(x\neq y)$、$loss(x,y)=\alpha|x-y|I(x<y)+(1-\alpha)|x-y|I(x>y)$、$loss(x,y)=c|x-y|I(|x-y|>\varepsilon)$ 等等。我们常见的一些参数估计方法如最小二乘估计、极大似然估计等都可以归入这一框架。

二、选择样本内预测损失与模型复杂度加权和最小的模型：该模型选择方法将会选择使得

$$E_{内+复}(\alpha)=\frac{1}{n}\sum_{i=1}^{n}loss(f_{\alpha}(\boldsymbol{X}_i),Y_i)+\lambda\cdot complexity(f)$$

达到最小的 α 对应的模型。不妨记这个 α 为 α^*，则模型 $f_{\alpha^*}(\boldsymbol{X})$ 即为样本内预测损失与模型复杂度加权和最小的模型。这里 $\lambda\geq 0$ 为模型复杂度相对于样本内预测损失的权重系数，$complexity(f)$ 代表模型复杂度，通常取 $complexity(f)=g(p)$，其中 $g(p)$ 是关于模型未知参数个数 p 的函数。一些模型选择准则如 AIC、BIC，以及一些参数估计方法如岭估计、$Lasso$、适应性 $Lasso$、弹性网等都可以归入这个框架。

三、选择交叉验证预测损失最小的模型：以 k 折交叉验证为例，该模型选择方法将会选择使得

$$E_{交}(\alpha)=\frac{1}{k}\sum_{j=1}^{k}loss\left(f_{\alpha}^{(-j)}\left(\boldsymbol{X}^{(j)}\right),Y^{(j)}\right)$$

达到最小的 α 对应的模型。不妨记这个 α 为 α^*，则模型 $f_{\alpha^*}(\boldsymbol{X})$ 即为 k 折交叉验证预测损失最小的模型。这里 $loss\left(f_{\alpha}^{(-j)}\left(\boldsymbol{X}^{(j)}\right),Y^{(j)}\right)$ 代表利用除第 j 折样本外的其余样本建模，并用第 j 折样本计算得到的平均预测损失。下面的 1.2 节将会详细地阐述交叉验证方法的基本思想与操作流程。

1.2 交叉验证

1974 年，明尼苏达大学统计学院创始人 *Seymour Geisser*，以及今伦敦学院大学（*University College London*）统计科学系名誉教授 *Mervyn Stone* 在两本不同杂志上发表了各自的文章，*Seymour Geisser* 的文章名为“*A Predictive Approach to the Random Effect Model*”[4]，发表在《*Biometrika*》上，而 *Mervyn Stone* 的文章则名为“*Cross-Validatory Choice and Assessment of Statistical Predictions*”[5]，发表在《*Journal of the Royal Statistical Society*》上。在这两篇文章中，两位作者首次独立系统地提出了交叉验证方法。

1975 年，*Seymour Geisser* 又在他的一篇名为“*The Predictive Sample Reuse Method with Applications*”[6]的文章中对交叉验证方法的有关应用及面临的某些问题作了详细阐释。在这篇文章

中，*Seymour Geisser* 多次强调着这样的观点：预测比参数估计更重要。他赞同卡尔皮尔逊（*Karl Pearson*）的“统计学的根本问题在预测”的观点，认为统计学应更加关注模型的预测效果而不是参数估计的效果。基于这样的观点，模型选择的首要关注点便是模型的预测效果，换句话说就是，我们应该基于模型的预测效果对模型进行选择。

交叉验证方法正是一种基于模型预测效果的模型选择方法，它的基本思想是每次从不同分割点将原始样本一分为二，取其中一部分数据用于模型的建立，这部分数据称为训练数据集（*training dataset*），而剩下的数据则用于评估模型预测效果，这部分数据称为测试数据集（*testing dataset*）。重复多次，对所有的模型预测效果加和取平均，便可以得到模型的平均预测效果。交叉验证就是基于这个平均预测效果对模型进行选择的。具体来说，交叉验证可分为完全交叉验证（*Exhaustive cross-validation*）和非完全交叉验证（*Non-exhaustive cross-validation*）[7]，1.2.1 小节将介绍完全交叉验证，1.2.2 小节则介绍非完全交叉验证。

1.2.1 完全交叉验证

完全交叉验证是在一定条件下，对所有可能的原始数据集的分割，评估一次模型预测效果，然后将平均预测效果用于模型选择的交叉验证方法。完全交叉验证的典型例子便是留 p 交叉验证（*Leave-p-out cross-validation*）[7]。它是在固定训练数据集的样本量为 $n-p$，测试数据集的样本量为 p 的条件下，穷尽所有可能的分割，对每种分割方式都评估一次预测效果，然后将平均预测效果作为模型选择的最终依据。

完全交叉验证因为穷尽了所有分割，所以一旦训练数据集的样本量得以确定，那么对同一预测模型和同一数据而言，它的交叉验证结果是固定的，这是完全交叉验证相较于非完全交叉验证的优势所在。在 1.2.2 小节中我们将会看到：对非完全交叉验证而言，由于没有穷尽所有分割，它的交叉验证结果是不唯一的，这种不唯一性将导致：两次利用同一方法可能得到不同的评价结果。

然而正所谓“金无足赤，人无完人”，虽然完全交叉验证的结果是唯一的，但它却存在着一个致命弱点，以留 p 交叉验证为例，当 n,p 较大时，特别是当 p 接近于 $[n/2]$ 时，完全交叉验证的计算将无法实现。这一点，通过利用排列组合知识便可得到。我们知道，对于样本量为 n 的数据集，要将其分割为两组，一组样本量为 p，另一组样本量为 $n-p$，则其互不相同的分法共有

$$O(n,p) \triangleq \binom{n}{p} = \frac{n!}{p!\,(n-p)!}$$

种。这样 $O(n,p)$ 就是得到完全交叉验证结果所必须的最少计算次数，由于每次计算预测效果都需要利用数据重新建立模型，因此实际的计算次数将会远远超过 $O(n,p)$.

现在我们暂且只考虑 $O(n,p)$，当 $n=1000$，$p=8$ 时，$O(n,p)\approx 2.41\times 10^{19}$，完成这样的计算量需要多长时间呢？目前世界上速度最快的超级计算机是我国自主研发的神威太湖之光，它的持续计算性能为 $93.015\ PFlops$[8]，即每秒能执行 93.015×10^{7} 次浮点运算。因此保守估计，对于一个量为 1000 的数据集应用留 8 交叉验证，神威太湖之光超级计算机至少要 24 小时不间断地工作约 300 天，而如果应用留 9 交叉验证，这一数字将会变成 91 年。可以想象，如果使用普通计算机，当 n,p 较大且要求训练数据集样本量至少不小于测试数据集样本量时，执行一次留

p 交叉验证所需要的时间相较于人类平均寿命而言将是一个天文数字。这一致命弱点严重限制了留 p 交叉验证在实际模型选择中的应用，因此目前完全交叉验证中应用较多的还是留 p 交叉验证的一个特例，即留 1 交叉验证。

在开始 1.2.2 小节前，还有一点需要说明。那就是：留 1 交叉验证，与非完全交叉验证中的 k 折交叉验证的一个特例，即 n 折交叉验证是等价的。事实上，如果严格来说，k 折交叉验证中的特例 n 折交叉验证不属于非完全交叉验证，但为了叙述的整体性起见，这里仍将 n 折交叉验证划归于非完全交叉验证范畴，而将它的同义词留 1 交叉验证划归于完全交叉验证。

1.2.2 非完全交叉验证

非完全交叉验证是对部分可能的原始数据集的分割，评估一次模型预测效果，然后将平均预测效果用于模型选择的交叉验证方法。它与完全交叉验证的区别就在于它不会穷尽对原始数据集的所有可能分割。非完全交叉验证的典型例子有：k 折交叉验证（*k-fold cross-validation*）、$Holdout$ 方法、重复随机子抽样验证/蒙特卡罗交叉验证（*Repeated random sub-sampling validation / Monte Carlo cross-validation*）等[7]。

k 折交叉验证是将原始数据集随机划分为 $k\ (2 \le k \le n)$ 个等大小（或近似等大小）的子数据集，然后依次选择第 i 个数据集作为测试数据集，其余子数据集作为训练数据集（$i = 1,2,\cdots,k$），每次利用训练数据集建立模型以及相应的测试数据集评价预测效果，由此得到平均预测效果，并将平均预测效果作为模型选择依据的交叉验证方法。

k 在 k 折交叉验证中被称为折数（$folds$），当原始数据集样本量 n 不是折数的整数倍，即 $k[n/k] \neq n$ 时，k 个子数据集只能近似等大小，此时应如何分配各个子数据集的样本量呢？对于这一问题，鲜有文献提及，笔者认为：此时应选择使得“各个子数据集的样本量的方差”达到最小的分配方式，下面的定理将给出该分配方式。

定理 当原始数据集样本量 n 不是折数 k 的整数倍，即 $k[n/k] \neq n$ 时，选择使得“各个子数据集的样本量的方差”达到最小的分配方式等价于：将原始数据集分为 k 个子数据集，其中 $n - k[n/k]$ 个子数据集样本量为 $[n/k]+1$，其余 $k - n + k[n/k]$ 个子数据集样本量为 $[n/k]$，这里 $[n/k]$ 表示 n/k 的整数部分

证明 设 k 个子数据集的样本量的方差为 S^2，第 i 个子数据集的样本量为 N_i，各子数据集平均样本量为 $\overline{N} = (1/k)\sum_{i=1}^{k} N_i$，则根据方差的定义有：

$$S^2=\frac{1}{k}\sum_{i=1}^{k}(N_i-\overline{N})^2$$

我们知道，无论各子数据集的样本量为多少，原始数据集样本量始终等于各子数据集样本量的总和，即 $\sum_{i=1}^{k}N_i=n$，所以有：

$$S^2=\frac{1}{k}\sum_{i=1}^{k}\left(N_i-\frac{n}{k}\right)^2$$

至此，寻找使得“各个子数据集的样本量的方差”达到最小的分配方式的问题就转化为了如下规划问题：

$$\min\sum_{i=1}^{k}\left(N_i-\frac{n}{k}\right)^2$$

$$\text{s.t.}\ \sum_{i=1}^{k}N_i=n,\ N_i\in N^+,n,k\ \text{为常数}$$

我们容易知道：当原始数据集样本量 n 是折数 k 的整数倍，即 $k[n/k]=n$ 时，只需取 $N_i=n/k$，便有 $S^2=0$，因此当 $k[n/k]=n$ 时，$N_i=n/k$ 便是上述规划问题的解。而现在原始数据集样本量 n 不是折数 k 的整数倍，即 $[n/k]\neq n/k$，因此 $N_i=n/k$ 不再是上述规划问题的解。

由于 $[n/k]$、$[n/k]+1$ 是距离 n/k 最近的两个正整数，因此我们猜想当所有 N_i 取 $[n/k]$ 或 $[n/k]+1$ 时，可以使得 $\sum_{i=1}^{k}(N_i-n/k)^2$ 达到最小值。在验证这个猜想之前，我们先证明：所有 N_i 取 $[n/k]$ 或 $[n/k]+1$ 的取法是唯一的。

设样本量等于 $[n/k]$ 的子数据集有 $a_1\ (0\le a_1\le k)$ 个，样本量等于 $[n/k]+1$ 的子数据集有 $a_2\ (0\le a_2\le k)$ 个，则它们满足如下的二元一次方程组：

$$\begin{cases}a_1[n/k]+a_2([n/k]+1)=n\\ a_1+a_2=k\end{cases}$$

从而 $a_1=k-n+k[n/k]$，$a_2=n-k[n/k]$。因此上述猜想可以更加精确地表述为：当原始数据集被分成样本量为 $N_1=[n/k],N_2=[n/k],\cdots,N_{a_1}=[n/k],N_{a_1+1}=[n/k]+1,N_{a_1+2}=[n/k]+1,\cdots,N_k=[n/k]+1$ 的 k 个子数据集时，$\sum_{i=1}^{k}(N_i-n/k)^2$ 取到最小值。下面我们将证明上述猜

想是正确的，为了叙述方便起见，我们令 $r=[n/k]$，$s=n/k$，$t=[n/k]+1$。

当原始数据集被分成样本量为 $N_1=r,N_2=r,\cdots,N_{a_1}=r,N_{a_1+1}=t,N_{a_1+2}=t,\cdots,N_k=t$ 的 k 个子数据集时，有

$$\sum_{i=1}^{k}\left(N_i-\frac{n}{k}\right)^2=\sum_{i=1}^{k}(N_i-s)^2=a_1(r-s)^2+a_2(t-s)^2\triangleq f^*$$

现在假设 f^* 不是 $\sum_{i=1}^{k}(N_i-n/k)^2$ 的最小值，则必存在满足条件 $\varepsilon_1+\varepsilon_2+\cdots+\varepsilon_k=0$ 的整数 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_{a_1},\varepsilon_{a_1+1},\varepsilon_{a_1+2},\cdots,\varepsilon_k$，使得原始数据集被分成样本量为 $N_1=r+\varepsilon_1,N_2=r+\varepsilon_2,\cdots,N_{a_1}=r+\varepsilon_{a_1},N_{a_1+1}=t+\varepsilon_{a_1+1},N_{a_1+2}=t+\varepsilon_{a_1+2},\cdots,N_k=t+\varepsilon_k$ 的 k 个子数据集时有

$$\sum_{i=1}^{k}(N_i-s)^2=\sum_{i=1}^{a_1}(r+\varepsilon_i-s)^2+\sum_{i=a_1+1}^{k}(t+\varepsilon_i-s)^2$$

$$=a_1(r-s)^2+a_2(t-s)^2+2(r-s)\sum_{i=1}^{a_1}\varepsilon_i+2(t-s)\sum_{i=a_1+1}^{k}\varepsilon_i+\sum_{i=1}^{k}\varepsilon_i^2$$

$$=f^*+2(r-s)\sum_{i=1}^{a_1}\varepsilon_i-2(t-s)\sum_{i=1}^{a_1}\varepsilon_i+\sum_{i=1}^{k}\varepsilon_i^2$$

$$=f^*+2(r-t)\sum_{i=1}^{a_1}\varepsilon_i+\sum_{i=1}^{k}\varepsilon_i^2=f^*-2\sum_{i=1}^{a_1}\varepsilon_i+\sum_{i=1}^{k}\varepsilon_i^2<f^*$$

从而有

$$g(\boldsymbol{\varepsilon})=g(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_k)\triangleq-2\sum_{i=1}^{a_1}\varepsilon_i+\sum_{i=1}^{k}\varepsilon_i^2<0$$

$$\varepsilon_1+\varepsilon_2+\cdots+\varepsilon_k=0\text{ 且 }\forall i,\varepsilon_i\in Z$$

由条件 $\varepsilon_1+\varepsilon_2+\cdots+\varepsilon_k=0$ 我们有：

$$g(\boldsymbol{\varepsilon})=-2\sum_{i=1}^{a_1}\varepsilon_i+\sum_{i=1}^{k}\varepsilon_i^2=\sum_{i=1}^{a_1}\left(\varepsilon_i^2-\varepsilon_i\right)+\sum_{i=a_1+1}^{k}\left(\varepsilon_i^2+\varepsilon_i\right)=\sum_{i=1}^{a_1}\varepsilon_i(\varepsilon_i-1)+\sum_{i=a_1+1}^{k}\varepsilon_i(\varepsilon_i+1)$$

然而利用二次函数性质我们容易得到：$\varepsilon_i(\varepsilon_i-1)\geq 0$，$\varepsilon_i(\varepsilon_i+1)\geq 0$ 在 $\varepsilon_i\in Z$ 时是恒成立的，这意味着 $g(\boldsymbol{\varepsilon})\geq 0$ 在 $\varepsilon_1+\varepsilon_2+\cdots+\varepsilon_k=0$ 且 $\varepsilon_i\in Z$ 的条件下恒成立，这显然与 $g(\boldsymbol{\varepsilon})<0$ 自相矛盾。

上述自相矛盾的结果告诉我们：关于“f^* 不是 $\sum_{i=1}^{k}(N_i - n/k)^2$ 的最小值”的假设是错误的。因此 f^* 是 $\sum_{i=1}^{k}(N_i - n/k)^2$ 的最小值，即当原始数据集被分成样本量为 $N_1 = [n/k], N_2 = [n/k], \cdots, N_{a_1} = [n/k], N_{a_1+1} = [n/k] + 1, N_{a_1+2} = [n/k] + 1, \cdots, N_k = [n/k] + 1$ 的 k 个子数据集时，$\sum_{i=1}^{k}(N_i - n/k)^2$ 取到最小值。

综上，当原始数据集样本量 n 不是折数的整数倍，即 $k[n/k] \neq n$ 时，选择使得“各个子数据集的样本量的方差”达到最小的分配方式等价于：将原始数据集分为 k 个子数据集，其中 $n - k[n/k]$ 个子数据集样本量为 $[n/k] + 1$，其余 $k - n + k[n/k]$ 个子数据集样本量为 $[n/k]$，这里 $[n/k]$ 表示 n/k 的整数部分。

$Holdout$ 方法是一种最简单的交叉验证方法，它是一种将原始数据集随机分成两份，一份作为训练数据集，另一份作为测试数据集，用训练数据集建模，用测试数据集得到预测效果，然后基于这个预测效果进行模型选择的交叉验证方法。$Holdout$ 方法中的训练数据集的样本量是任意的，但一般要求训练数据集样本量不小于测试数据集样本量。

重复随机子抽样验证又名蒙特卡罗交叉验证，它是一种每次将原始数据集随机分成两份，一份作为训练数据集，另一份作为测试数据集，用训练数据集建模，用测试数据集得到预测效果，重复多次得到平均预测效果，然后基于这个预测效果进行模型选择的交叉验证方法，其中每次分割时两个子集的样本量是随机的，但一般要求训练数据集样本量不小于测试数据集样本量。因此蒙特卡罗交叉验证可以看作是 $Holdout$ 方法的多次组合，它类似于 k 折交叉验证，但与 k 折交叉验证最大的不同在于：蒙特卡罗交叉验证每次使用的测试数据集可能存在重复数据，而 k 折交叉验证每次使用的测试数据集是互不相同的。因此，相比于 k 折交叉验证，蒙特卡罗交叉验证的优势在于训练数据集样本量与测试数据集样本量之比不依赖于折数，但它的劣势在于对数据的利用不够充分，因为某些数据将不会出现在测试数据集中，而另外一些数据则会重复出现在测试数据集中。

可以看到，上述非完全交叉验证方法克服了完全交叉验证可能面临的“即便当今最先进计算机也无法解决”的计算瓶颈，但这种舍弃掉部分计算量的作法也带来了新的问题。以 k 折交叉验证为例，随机将样本分成 k 份共有

$$D(n,k) = \frac{n!}{k! \times ([n/k]+1)!^{n-k[n/k]} \times [n/k]!^{k-n+k[n/k]}}$$

种分法，然而实际操作时，我们只会选择其中一种分法，然后基于这种分法计算 k 折交叉验证预测效果，这就给模型选择带来了不确定性。但这种不确定性的存在是伴随着计算机计算能力的局限产生的，换句话说：正是计算能力的有限导致了非完全交叉验证方法的产生，而任何一种非完全

交叉验证方法，由于其无法穷尽所有的样本分割方式，它在应用于同一原始数据集以及同一预测模型时，其得到的预测效果必然具有不确定性。因此笔者认为：对于实际工作者而言，关键是在一些非完全交叉验证方法中选择不确定性最小的那种方法。这一观点也将是本文下面研究和探讨 k 折交叉验证折数确定方法的基本指导思想。

2 折数与 k 折交叉验证

2.1 折数确定在 k 折交叉验证中的重要性

笔者曾在统计案例分析的课程实践中分析过一个关于马蹄蟹追随者状态和马蹄蟹体表特征的数据集，这个数据集包含 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）。当时为了研究马蹄蟹体表特征与其有无追随者之间的关系，进而达到通过马蹄蟹体表特征预测马蹄蟹有无追随者的目的，笔者针对该数据集构建了 10 种统计模型。表 2-2 至表 2-7 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 项交叉验证预测效果指标的统计表（注：所有统计表均按 Youden 指数降序排列，表中排名前 3 的模型加粗显示）。表中各项指标的含义如下：

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。

表 2-2

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 2-3

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 2-4

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 2-5

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 2-6

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 2-7

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

从这个案例中，我们可以直观地看到：不同的折数导致了不同的模型选择结果，并且可以预见这是在运用 k 折交叉验证方法进行模型选择时经常会遇到的问题。那么在实际工作中，当我们在面对不同折数导致的不同的模型选择结果时，该如何抉择呢，是加权平均各折数的交叉验证结果，还是选择其中一个较稳定的折数的交叉验证结果？依笔者的观点来看，我们应选择其中一个较稳定的折数的交叉验证结果。原因有如下几点：第一，如果选择加权各折数的交叉验证结果，会遭遇“到底哪些折数参与加权”的新问题，并且各折数的权重难以确定；第二，各折数的交叉验证结果的稳定性可通过方差来刻画，虽然多数情况下无法求得这个总体方差，但可以使用样本方差进行估计；第三，相较于做多个折数的交叉验证然后加权平均的做法，选择其中一个较稳定的折数做交

叉验证，对计算机资源的耗费要小得多，实际操作起来也更加简单方便。基于以上观点以及不同折数会导致不同模型选择结果的事实，折数在 k 折交叉验证中的重要性就不言而喻了。2.2 节将会通过个案初步探讨不同数据量下 k 折交叉验证结果的稳定性与折数 k 的关系，进而为根据样本量和计算成本确定 k 折交叉验证最优折数提供依据。

2.2 k 折交叉验证结果的稳定性分析

对固定的折数 k，什么才是影响 k 折交叉验证结果稳定性的主要因素呢？是不同的建模方法，不同的数据特征，还是不同的数据量，亦或者是其他因素呢？笔者分析认为，对固定的折数 k，k 折交叉验证结果之所以出现不确定性，是因为随机将样本分成 k 份的方法不止一种（除非 $k=n$），而实际应用 k 折交叉验证时，我们只会选取其中一种分法。这一点笔者曾在 1.2.2 小节中有所提及，在那里，笔者由排列组合知识得到了分法数目的数学表达式，即

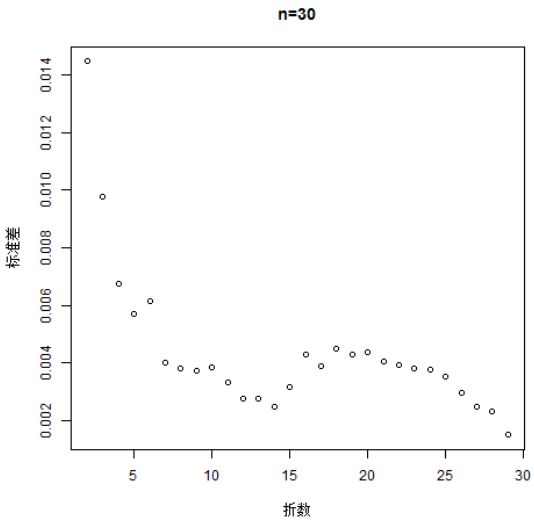
$$D(n,k)=\frac{n!}{k!\times([n/k]+1)!^{n-k[n/k]}\times[n/k]!^{k-n+k[n/k]}}$$

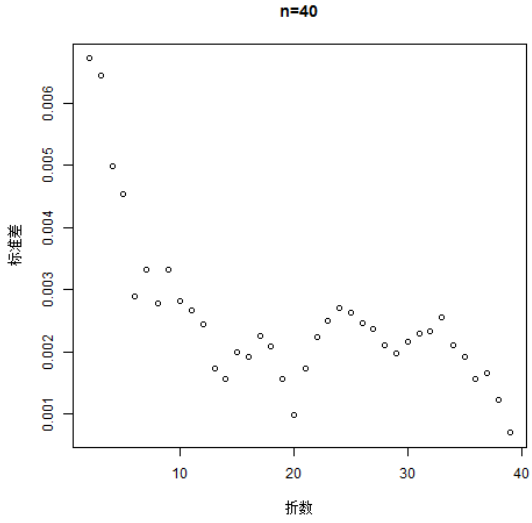
。从表达式中我们可以清楚地看到，对固定的折数 k，$D(n,k)$ 只与原始数据集的样本量 n 有关。这提醒我们：对固定的折数 k，原始数据集的样本量 n 才是影响 k 折交叉验证结果稳定性的主要因素。然而遗憾地是：直接绘制 $D(n,k)$ 的函数图像是困难的，特别是当 n 较大时，$n!$ 将超过计算机所能存储的数值范围。面对这一问题，笔者曾尝试通过拆分分子分母的阶乘各项，然后各项先作除法后作乘法的方式来克服这一困难，但事与愿违，新的问题接踵而至，作除法后很多项接近于 0，导致了严重的计算误差。

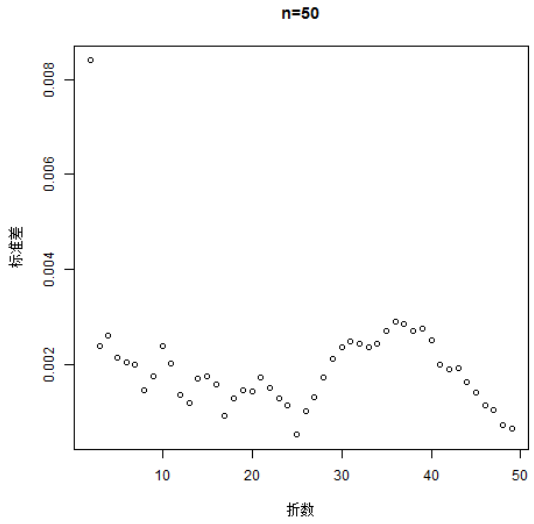
鉴于理论探讨的困难性，笔者决定从实际案例入手来探讨样本量与 k 折交叉验证结果稳定性关系。笔者曾在实用回归分析的课程实践中分析过中国石油股票 y_t 与中国石化股票 x_t 在 2015 年 261 个交易日的收盘价数据，当时笔者利用 2015 年 1 月 4 日到 2015 年 11 月 1 日共 201 个交易日的数据构建了中国石油股票收盘价差分序列与中国石化股票收盘价差分序列的回归方程为：$y_t-y_{t-1}=0.001+1.728668(x_t-x_{t-1})$，并利用该模型及中国石化股票收盘价预测了 2015 年 11 月 2 日到 2015 年 12 月 31 日共 43 个交易日中国石油股票收盘价，预测误差绝对值平均约为 0.127，最大预测误差绝对值约为 0.528，最小预测误差绝对值约为 0.004，预测误差绝对值标准差约为 0.119. 从以上预测误差我们知道，对该数据集的差分序列，线性回归模型是一个较好的预测模型。

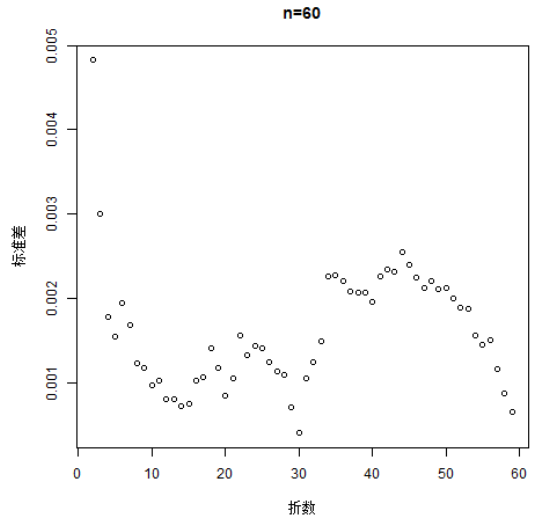
本文下面就将对该数据集前 $n\,(n=4,\cdots,100)$ 个差分序列应用线性回归模型，并计算不同样本量、不同折数、不同分法下的 k 折交叉验证预测效果，以估计不同样本量、不同折数下的 k 折交叉验证预测效果标准差，进而为通过样本量确定最优折数提供依据。

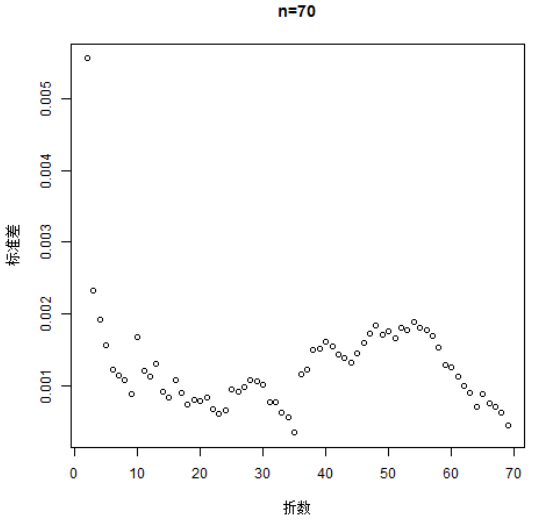
下图显示了 $n=30,\cdots,100$ 时，各折数与交叉验证预测效果标准差估计的散点图，其中交叉验证预测效果的标准差用于衡量交叉验证预测效果的稳定性，需要说明的是：这里的标准差是对总体标准差的样本估计，样本量为 30.

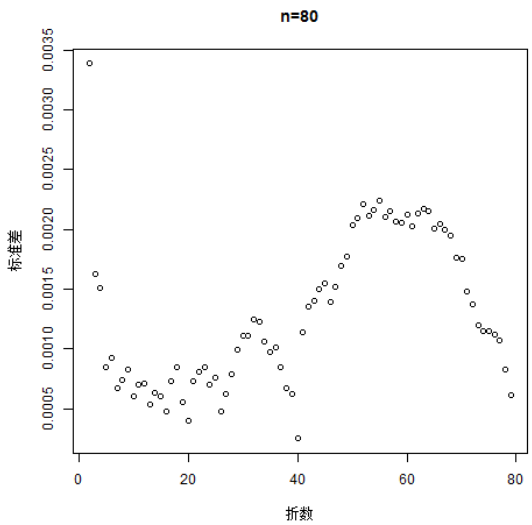

n=30
标准差
折数


n=40
标准差
折数


n=50
标准差
折数


n=60
标准差
折数


n=70
标准差
折数


n=80
标准差
折数

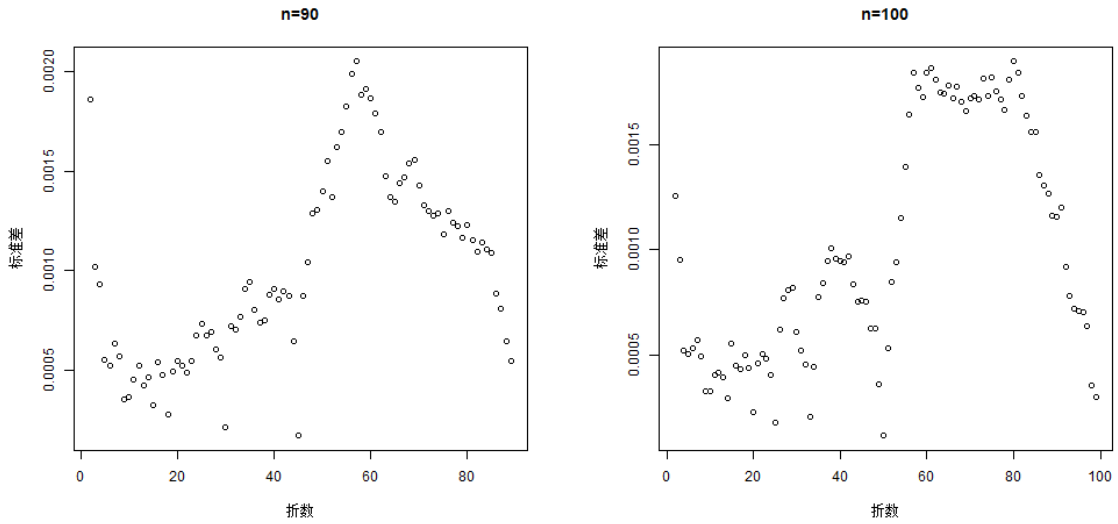


图 2-8 折数与交叉验证预测效果标准差估计的散点图（n=30,40…,100）

通过对图 2-8 的仔细观察，笔者发现：在标准差估计误差允许范围内，当 k 位于 $[n/1],[n/2],[n/3],[n/4],\cdots$ 附近时，标准差会达到一个局部极小值，且通常标准差在 $[n/i]$ 的值随 i 的递增而递增。

3 折数确定方法的探讨及检验

3.1 折数确定方法的探讨

如果不考虑计算成本，从 k 折交叉验证预测效果稳定性角度出发，那么毫无疑问我们应该选择 n 折交叉验证，即留 1 交叉验证。然而随着样本量 n 的增大，计算成本将随之增加。1.2.2 小节在讲述非完全交叉验证时笔者指出：类似 k 折交叉验证这样的非完全交叉验证方法的产生是迫于计算资源的有限而产生的。因此我们在探讨折数确定方法时必须考虑计算成本，这里我们将计算成本用允许的最大折数 k_{max} 代表。通过 2.2 节对个案的探讨，我们发现了重要规律，即当折数使得原始数据集尽量等分为样本量为 $i, i=1,2,\cdots$ 的子数据集时，k 折交叉验证预测效果标准差有局部极小值，并且这样的子数据集的样本量越小，局部极小值通常越小。基于这样的规律，我们提出 k 折交叉验证折数粗略确定方法如下（n 为原始数据集样本量，k 为确定的折数）：

1. 当 $k_{max} \geq n$ 时，选择 $k=n$;
2. 当 $[n/(i+1)] \leq k_{max} < [n/i]$ 时，选择 $k=[n/(i+1)]$，$i=1,2,\cdots$

即如果原始数据集样本量 n 不大于允许最大折数 k_{max}，选择 n 折交叉验证，否则看允许最大折数 k_{max} 是否大于等于 $[n/2]$，若是，选择 $[n/2]$ 折交叉验证，否则继续看允许最大折数 k_{max} 是否大于等于 $[n/3]$，若是，选择 $[n/3]$，如此继续下去，直到找到离 k_{max} 最近且位于 k_{max} 左侧的 $[n/i]$ 为止。

3.2 折数确定方法的检验

为了检验上述 k 折交叉验证折数粗略确定方法的有效性，笔者针对上述折数确定方法编写了计算机程序，输出了当原始数据集样本量 $n = 5,6,\cdots,100$ 时，由该方法得到的最优折数 $\hat{k}_n$，并将这些最优折数与 2.2 节所用案例得到的稳定性排名前三的较优折数 k_{n1}, k_{n2}, k_{n3} 作了对比分析。表 3-1 列出了各个样本量下，$\hat{k}_n$ 与 k_{n1}, k_{n2}, k_{n3} 的数值。（注：考虑到实际案例是对总体标准差的估计，而估计具有不确定性，应此笔者认为只要 $\hat{k}_n$ 等于任意一个排名靠前的实际较优折数，即可视为判断无误，结果一列标注“×”的即为判断有误）

表 3-1

样本量	kn1	kn2	kn3	kn(估计)	结果	样本量	kn1	kn2	kn3	kn(估计)	结果
5	5	4	3	5		29	14	13	10	14	
6	6	5	4	6		30	14	13	12	15	×
7	7	6	5	7		31	15	10	9	16	×
8	8	7	6	8		32	16	15	14	16	
9	9	8	7	9		33	16	17	11	16	
10	10	9	8	10		34	16	17	11	17	
11	11	10	9	11		35	16	17	18	18	
12	12	11	10	12		36	18	17	16	18	
13	13	12	11	13		37	18	17	20	18	
14	14	13	12	14		38	19	18	13	19	
15	15	14	13	15		39	19	10	14	20	×
16	16	15	14	16		40	20	14	19	20	
17	17	16	15	17		41	20	19	7	20	
18	18	17	16	18		42	14	20	19	14	
19	19	18	17	19		43	20	14	15	14	
20	20	19	18	20		44	14	20	8	15	×
21	20	19	18	20		45	15	16	17	15	
22	11	9	10	11		46	12	11	15	15	
23	20	18	19	12	×	47	12	17	10	16	×
24	12	20	9	12		48	16	10	8	16	
25	19	20	12	12		49	17	16	10	16	
26	13	12	10	13		50	17	13	18	17	
27	11	10	9	14	×	51	13	17	18	17	
28	14	13	12	14		52	13	17	18	17	

样本量	kn1	kn2	kn3	kn(估计)	结果	样本量	kn1	kn2	kn3	kn(估计)	结果
53	18	11	17	18		77	19	9	13	19	
54	18	17	19	18		78	19	11	16	20	×
55	19	18	9	18		79	20	19	13	20	
56	18	14	13	19	×	80	20	16	13	20	
57	19	10	20	19		81	16	13	20	20	
58	20	15	19	19		82	16	20	10	20	
59	20	15	18	20		83	14	12	16	17	×
60	14	15	13	20	×	84	12	9	17	17	
61	12	15	11	20	×	85	17	12	10	17	
62	13	15	14	16	×	86	17	15	16	17	
63	13	15	12	16	×	87	10	11	17	17	
64	8	13	16	16		88	14	11	18	18	
65	8	13	16	16		89	9	15	13	18	×
66	14	11	17	16	×	90	18	15	9	18	
67	13	16	14	17	×	91	18	13	14	18	
68	17	14	16	17		92	13	15	10	18	×
69	17	18	9	17		93	9	8	19	19	
70	18	20	19	18		94	16	9	12	19	×
71	18	12	9	18		95	19	16	12	19	
72	12	11	18	18		96	16	12	11	19	×
73	19	18	10	18		97	16	12	14	19	×
74	15	19	18	18		98	14	20	12	20	
75	15	13	19	19		99	20	9	14	20	
76	19	11	15	19		100	20	14	10	20	

综合来看，3.1 节探讨得到的折数确定的粗略方法具有一定的有效性，由 2.2 节实际案例检验得到的该折数确定方法的正确率达到 78.3%，虽然这一结果还不甚理想，但就实际应用而言已经不差了，至少比从主观上任意确定折数要好。

4 结语

本文利用定性分析和定量分析相结合的方法并借助计算机初步研究了 k 折交叉验证折数的确定方法，在研究过程中，通过对 k 折交叉验证预测效果稳定性的分析发现了重要规律，并依据这条规律提出了 k 折交叉验证折数确定的粗略方法。但数据显示，这一方法的效率还稍显不足，因而笔者衷心期待能够有更多的学者投入到交叉验证方法的深入研究中，为实际工作者更好应用交叉验证方法进行模型选择铺平道路。

5 参考文献

[1] George E. P. Box, Norman R. Draper, Empirical Model-Building and Response Surfaces, Wiley, 1987.

[2] 李航，《统计学习方法》，北京，清华大学出版社，2012.

[3] 吕晓玲 宋捷 主编，《大数据挖掘与统计机器学习》，北京，中国人民大学出版社，2016.

[4] Seymour Geisser, A Predictive Approach to the Random Effect Model, Biometrika, 1974.

[5] Mervyn Stone, Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society,1974.

[6] Seymour Geisser, The Predictive Sample Reuse Method with Applications, Journal of the American Statistical Association, Vol 70, No. 350 (Jun., 1975), pp. 320-328.

[7] Wikipedia, Cross-validation (statistics): https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[8] 百度百科，神威 •太湖之光超级计算机：http://baike.baidu.com/link?url=Vftt9demltaOa4H2UjtReU1VCzLZT1xPDBGkCIIuA8b4K-kKgN7E_E0nuil0_jrg1faG9wud6B0FCJPm6aU-w4CT-Tk81Vgn6OMFHlcuREq4Qn-cWS5SlmCXtd1yBURqGa0w_vs94zXLk41xIU5MYuth08GyptqrwJ8sO1nqTDzWexlUb4kQjHPmzN1DgPLCr2B9v4CoW-ytwBXFf5iiUWCWcUEH7GDe8DgZxZYiIfpTH0t9w8q8VD4It13nrCjCuVGojVonNQCoMXjI6DlnEVCS7lILQJg_bHaNf1Y088G8KJ2Wbe-0sUrYN-uDkvQqW1cIaURRgPbxCsjfGR7o2_

[9] Norman Matloff, The Art of R Programming, William Pollock, 2011.

[10] Robert I. Kabacoff, R in Action(second edition), Manning Publications Co, 2015.