Computer Science and Application 计算机科学与应用, 2016, 6(3), 132-136
Published Online March 2016 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2016.63017

Hans汉斯

Comparison of Cross Validation Results of Classification Model Based on Nonparametric Method

Qizhao Xu

Yunnan University of Finance and Economics, Kunming Yunnan
Email: 357980462@qq.com

Received: Feb. 20th, 2016; accepted: Mar. 14th, 2016; published: Mar. 17th, 2016

Abstract

The true generalization error is a scientific evaluation criterion to model selection. For the classification model, the rate of miscarriage of justice, which is an excellent estimation, is based on cross validation to the true generalization error. So we compare models through the cross validation results. Cross validation results are random variables, which have its distribution. For the random variable, its distribution is very hard to detect. Therefore, based on the comparison of cross validation results, this paper designs a hypothesis testing through the nonparametric method to inspect whether a significant difference exists between two classification models.

Keywords

Cross Validation, Model Comparison, Nonparametric, Hypothesis Test

基于非参数方法的分类模型交叉验证结果比较

徐奇钊

云南财经大学，云南 昆明
Email: 357980462@qq.com

收稿日期：2016年2月20日；录用日期：2016年3月14日；发布日期：2016年3月17日

文章引用: 徐奇钊. 基于非参数方法的分类模型交叉验证结果比较[J]. 计算机科学与应用, 2016, 6(3): 132-136.
http://dx.doi.org/10.12677/csa.2016.63017

摘　要

本文主要研究了基于非参数方法的分类模型交叉验证结果比较，主要是对实例通过非参数的方法进行模型比较的假设检验，检验两分类模型是否存在显著差异。模型的真实泛化误差是一个较为科学的模型比较标准，对于分类模型而言，模型的真实泛化误差表现为分类模型的误判率，而基于交叉验证得到的结果是模型误判率的一个优良估计，可以通过交叉验证结果对模型进行比较。交叉验证结果是随机变量，存在分布，而对于此随机变量而言，其分布是很难观测的，因此，对于交叉验证结果的比较，本文通过非参数的方法进行模型比较的假设检验，检验两分类模型是否存在显著差异。

关键词

交叉验证，模型比较，非参数，假设检验

1. 引言

对于因变量为分类变量的数据进行建模，有多种建模方法。有基于统计的方法、基于机器学习的方法等等。在对数据进行建模分析时，模型对于新数据的泛化能力是评价模型好坏的一个重要标准。因此，对模型的比较，较为科学的方法应是对其泛化能力进行比较。事实上，如果已知产生数据的真实模型，那么使用不同的模型对数据进行建模，与真实模型进行比较，理论上可以得到不同模型的真实泛化误差，用以衡量、比较各模型泛化能力。然而，由于真实模型不可观测，用于建模的各个模型的真实泛化误差是不可得的，只有使用真实泛化误差的估计对各模型进行比较。

Wasseman L.等[1]指出，交叉验证是泛化误差的估计中最简单、使用最广泛的方法。对于模型间的比较，吴喜之等[2]使用 k 折交叉验证结果的均值直接进行比较。而使用统计学频率学派的观点看，交叉验证结果作为真实泛化能力的估计量本身是一个随机变量，而对不同模型的交叉验证结果进行比较，实际上是对不同随机变量的比较。因此，较为合理的方法不是直接比较交叉验证结果的均值，而是使用假设检验的方法，对交叉验证不同结果的位置参数进行比较，科学的排除随机因素的影响。高红[3]指出分类器的分类错误率是不可得的，只能被估计出来，并且其估计，即交叉验证结果与其折数、测试集的选取有关。Fushiki, T. (2011) [4]使用了多次的 K 折交叉验证估计了模型的预测误差。不同模型交叉验证结果应有其自身的分布，对一些特殊的模型理论上应可以推导出精确分布，而实际中，这是很困难的或不可能做到的，为了排除随机因素的影响，更为合理的方法是进行多重的交叉验证得到交叉验证结果的观测数据，并采用不依赖于分布假设的非参数方法进行分析。Conover, W. J. (2012) [5]提出了多种非参数检验的方法，其中，对本文引用的符号检验理论方法做出了详细说明。

吴喜之[6]指比较成对数据要满足假定，每一对数据或者来自同一个或者可比较的类似的对象。对于两个模型建模效果的比较，由于每一对数据是同一个交叉验证的测试集数据，满足此假设，可以将每一对数据相减后，利用符号检验等进行两模型效果比较即可。

2. 理论说明

2.1. 数据收集

对于分类模型的比较，模型对应的真实泛化误差表现为模型的真实误判率，使用交叉验证的方法，所得到的测试集的误判率就是真实误判率的一个良好估计。一个分类模型对应的测试集误判率本身是一个随机变量，与折数、测试集的选取有关，对于模型M_i，定义测试集误判率如(1)式，

$$\varepsilon_t^i = \frac{\#\left(y_{tj} \neq f_{tj}^i\right)}{m} \tag{1}$$

其中，ε_t^i 为第 i 个分类模型在第 t 个测试集上的误判率，y_{tj} 为第 t 个测试集第 j 个观测的真实因变量数据，f_{tj}^i 为第 i 个模型对第 t 个测试集第 j 个因变量的预测值，$\#\left(y_{tj} \neq f_{tj}^i\right)$ 为第 i 个模型在第 t 个测试集上的误判个数，m 为测试集观测点个数。

对第 i 个分类模型进行多次 k 折交叉验证，由于每次所选取的测试集不同，就得到一组 ε_t^i 数据，可以视为分类模型测试集误判率的样本数据。

2.2. 交叉验证及非参数检验概述

交叉验证是一种判断模型好坏的重要方法，其一般过程为首先拿原数据集中的一部分数据作为训练集进行建模，再用另一部分在训练模型时没有用到的数据集作为测试集，得到模型的泛化误差，进而对模型进行比较。k 折交叉验证是把原数据集分为 k 份，每一次建模使用数据集的 $k-1$ 份数据，用剩下的 1 份数据作为测试集得到每次建模的泛化误差。n 重 k 折交叉验证是将随机把数据分成 k 份的过程重复 k 次，得到多个 k 折交叉验证测试集的泛化误差，即将 k 折交叉验证重复 n 次。

对于非参数检验，百度百科给出的定义是“在总体方差未知或知道甚少的情况下，利用样本数据对总体分布形态等进行推断的方法”。非参数检验方法不涉及或很少涉及对数据分布等假定，在数据集满足一定分布假定的条件下，非参数检验方法没有参数方法高效，而对于一些复杂的数据集，前者相较于后者而言有较高的普适性，可用于分析这些较为复杂的数据类型。

2.3. 随机森林分类与支持向量机分类的对比分析

本文对于后文实例数据将使用随机森林分类模型及支持向量机分类模型进行建模，并对两模型建模效果进行比较。随机森林分类模型与支持向量机分类模型均为基于算法的模型，没有关于分布的假定，适合处理本文实例中的复杂数据类型的数据。随机森林是基于基于树的组合方法，通过在随机的选取有放回再抽样的样本建立决策树模型，而对于每一个决策树的建模，随机选取变量作为每一节点的分割变量，通过产生的树模型投票得到分类结果；支持向量机分类模型是把空间中的不同类型的点使用超平面进行分割，而使得此超平面与各类点的距离最大，而不涉及数据分布的一种方法。两种方法都是基于算法出发，对数据集的假定较少，因此很难通过经典的统计方法进行模型比较，而是通过交叉验证等方法比较其分类效果。

2.4. 两个分类模型交叉验证结果的非参数检验

对两个模型交叉验证测试集的误判率，对象是相同测试集，对两个模型的误判率，每一对误判率数据对应着同一个测试集，因此，可以使用成对数据检验方法。对于每一个分类模型在 n 重 k 折交叉验证中，对每一重 k 折交叉验证可以得到 k 个测试集误判率，一共可以观测 nk 个测试集误判率数据。对于两个模型比较，由于是成对数据，故比较两组误判率数据相减后所得的差，检验其中位数是否为零。由于不知道分布，因此，较为适用的方法是对误判率差数据使用符号检验其中位数是否为零。

3. 实例分析

本文使用 r 软件自带的鸢尾花数据集进行建模分析。对鸢尾花数据集分别使用支持向量机分类模型，以及随机森林分类模型进行建模，此处，直接套用 r 软件 e1071 包及 randomForest 包直接进行建立模型，对两个模型的建模效果进行比较。对每一模型建模使用 10 重的 5 折交叉验证。对于支持向量机分类在鸢尾花数据的 10 重 5 折交叉结果如表 1 所示(每一行是一次 k 折交叉验证)。

对于随机森林分类在鸢尾花数据的 10 重 5 折交叉验证结果如表 2 所示(每一行是一次 k 折交叉验证)。

表 2 与表 1 数据对应数据相减，可以得到两模型测试集误判率差数据，如表 3 所示。

Table 1. The results of cross validation based on support vector machine
表 1. 支持向量机交叉验证结果

0.066667	0.066667	0.033333	0.066667	0.066667
0	0.033333	0.033333	0.033333	0
0.033333	0.033333	0.033333	0.066667	0
0	0.066667	0	0	0.033333
0.1	0	0.1	0.1	0.066667
0.066667	0.033333	0.033333	0.033333	0.033333
0.033333	0.066667	0.066667	0.066667	0.066667
0	0.033333	0.066667	0.033333	0.033333
0.066667	0	0	0.033333	0.033333
0.033333	0.033333	0.066667	0.033333	0.033333

Table 2. The results of cross validation based on random forest
表 2. 随机森林交叉验证结果

0.066667	0.066667	0.033333	0.066667	0.133333
0	0.033333	0.033333	0.033333	0
0.033333	0.066667	0.033333	0.1	0.033333
0.066667	0.1	0	0.033333	0.033333
0.066667	0.033333	0.066667	0.066667	0.033333
0.033333	0.033333	0.033333	0.033333	0.033333
0.1	0.066667	0.066667	0.066667	0.1
0.033333	0.033333	0.066667	0.066667	0.033333
0.066667	0	0.066667	0.033333	0.033333
0.033333	0.1	0.066667	0.033333	0.033333

Table 3. The differences of test set's error rate
表 3. 测试集误判率差

0	0	0	0	0.066667
0	0	0	0	0
0	0.033333	0	0.033333	0.033333
0.066667	0.033333	0	0.033333	0
−0.03333	0.033333	−0.03333	−0.03333	−0.03333
−0.03333	0	0	0	0
0.066667	0	0	0	0.033333
0.033333	0	0	0.033333	0
0	0	0.066667	0	0
0	0.066667	0	0	0

对于表 3 数据，由于无法对数据分布做出任何假设，因此，使用对分布不用做出任何假设的符号检验方法对其检验其中位数是否为零，又由于从原始数据看来，表 2 数据大于表 1 数据，即对表 3 数据做假设检验，$H_0: M_0 \le 0, H_1: M_0 > 0$。如果原假设成立，数据中大于中位数的个数与小于中位数的个数应该差不多，设 S^+ 为大于 M_0 的数据个数，S^- 为小于 M_0 的数据个数，$K = \min\left(S^+, S^-\right)$，可以证明，$K \sim B\left(n', 0.5\right)$，其中，$B\left(n', 0.5\right)$ 表示二项分布，$n' = S^+ + S^-$

对于表 3 数据，K 的观测值为 $S^- = 5$，取当前值和更极端值概率 $P\left(S^- \le 5\right) = 0.03178406$，小于给定显著性水平 0.05，拒绝原假设，可以认为 SVM 模型的建模效果较好，测试集误判率低于随机森林模型。

4. 结论

本文在传统的使用交叉验证的方法来判断模型结果的基础上，创新之处在于使用非参数的方法构造较为科学的假设检验，检验两模型交叉验证结果是否存在显著差异，较科学的说明如果拒绝两模型不存在显著差异的原假设，则两模型的差异不仅仅是由于随机因素，如测试集的选择，折数的不同所引起的，而是确实一个模型表现较好，更为适合对一组数据集的建模分析。

对于本文的一些不足之处，体现在比较两模型的过程中仅仅使用了符号检验的方法，如果实际建模过程中，可以对交叉验证结果的分布做出假设，可以使用一些效率更高的、使用信息更多的非参数方法或甚至在可以推导出交叉验证结果精确分布或渐进分布的情况下，使用参数方法进行检验，使得检验精确性更高。同时，本文所使用的检验方法只适用于两个模型的比较，对于多个模型的比较需要寻找其他的非参数方法，而不是简单的套用两模型比较的检验方法。

致　谢

感谢这篇论文所涉及到的各位学者。本文引用了数位学者的研究文献，如果没有各位学者的研究成果的帮助和启发，我将很难完成本篇论文的写作。感谢我的老师与同学，在我写论文的过程中给予我了很多你问素材，还在论文的撰写和排版等过程中提供热情的帮助。由于我的学术水平有限，所写论文难免有不足之处，恳请各位老师和学友批评和指正！

参考文献 (References)

[1] Wasseman, L. (2000) Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, **44**, 92-107. http://dx.doi.org/10.1006/jmps.1999.1278

[2] 吴喜之. 复杂数据统计方法[M]. 北京: 中国人民大学出版社, 2012.

[3] 高红. 基于交叉验证的错误率估计分析[J]. 科技信息, 2011(25): I0149.

[4] Fushiki, T. (2011) Estimation of Prediction Error by Using *K*-Fold Cross-Validation. *Statistics & Computing*, **21**, 137-146. http://dx.doi.org/10.1007/s11222-009-9153-8

[5] Conover, W.J. (2012) Practical Nonparametric Statistics. *Technometrics*, **14**, 977-979.

[6] 吴喜之. 非参数统计[M]. 北京: 中国统计出版社, 2013.