

硕 士 研 究 生 读 书 报 告



题目 分类器在多数据集下的比较

作者姓名 汤建华

作者学号 21651084

指导教师 李启雷

学科专业 移动互联网与游戏开发

所在学院 软件学院

提交日期 二○一七年一月

Comparison of classifier in most of the series

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Li Qilei

By

Tang Jianhua

Zhejiang University ,P.R. China

2017

摘要

随着机器学习的广泛深入，虽然对于单个数据集的两种学习算法的比较已被仔细审查相当长的一段时间了，但对于多个数据集的多种算法比较的统计检验问题却一直被忽略了，而它对典型的机器学习研究更为重要。本文综述了目前的实践，然后从理论和实证上考察了几个合适的测试。在此基础上，我们推荐了一套简单而安全，稳定的非参数测试分类统计比较：对两种分类比较的Wilcoxon符号秩检验以及对多个数据集的多种分类比较的弗里德曼测试和相应的事后检验。后者的结果也可整齐地呈现新推出的CD（临界差）图。

**关键词**：比较研究，统计方法，Wilcoxon符号秩检验，Friedman检验，多重比较试验

Abstract

With machine learning extensively, although for a single data set of two kinds of learning algorithms have been compared carefully for a long time, but for the statistical test comparison of various algorithms with multiple data sets has been ignored, and its typical machine learning research is more important. This article reviews the current practice and then examines several suitable tests both theoretically and empirically. On this basis, we recommend a set of simple, safe and stable classification statistics comparison of nonparametric tests: Wilcoxon signed rank test for two classifications, and Freedman test and corresponding post test for multiple classifications of multiple data sets. The results of the latter can also be neatly presented by the new CD (critical difference) diagram.

**Keywords：**Comparative study, statistical method, Wilcoxon sign rank test, Friedman test, multiple comparison test

1简介

在过去的几年中，机器学习领域越来越意识到需要公布统计验证的结果。这归因于该区域的成熟和越来越多现实世界的应用程序，打开机器学习框架使得开发新的或修改现有的算法并比较它们的可用性变得更容易。

在一个典型的机器学习论文里，一个新的机器学习算法，它的一部分或一些新的处理前或处理后的步骤已被提出，并隐含了对现有算法提高产量改进性能的假设。另外，各种解决方案已被提出，目的是从失败中分辨成功。一个号码被选择用于测试数据集的测试，则算法的运行和所得模型的质量将使用适当的措施，即最常见的分类精度进行评价。剩下的步骤以及本文的主题，将对性能改进的假设进行统计地验证。

以下部分探讨了相关的理论工作和现行做法。各种研究已经解决了单一数据集两种分类器的比较问题并提出几种解决方案。他们的消息已经采取社区和过于自信配对t检验交叉验证让位于McNemar检验和5×2交叉验证。另一方面，多个数据比较多个分类器设置，这种情况更常见，尤其是一般的性能，而不是在某些特定问题中的性能在进行测试时，仍然是未开发的理论和有待各种特设程序，无论是缺乏统计的理由或使用不恰当的统计方法。为了看到实际操作中使用了什么，我们研究了国际会议近期（1999- 2003年）的机器学习程序。我们观察到，许多本来优秀创新的机器学习论文的结束是从一个矩阵得出结论，例如，比较所有麦克尼马尔对分类器进行的测试，似乎是针对多重比较的测试，如ANOVA和弗里德曼试验是有待发明。

论文的核心是用于统计测试，可能是（或已经是）多个数据组的两个或多个分类比较的研究。从形式上看，假设我们已经测试ķ学习上的N个数据集的算法. *cij*是对第i个数据组第j个算法的性能得分。它的任务是基于所述值*cij*，判断所述统计算法是否显著不同，在两个以上的算法中，哪些是性能不同的特定算法。我们不记录这些方差的得分，只假定测量结果为“可靠”；因此，我们需要对每个数据集进行足够的实验，优选，所有的算法都使用同样的随机样本进行评价。我们做出的抽样方案没有其他的假设。

在第3部分，我们将观察每个测试背后的理论假设来阐明我们的问题。虽然一些测试在机器学习文献中相当普遍，许多研究人员似乎对于实际测量什么，以及它们在哪些情况下适合一无所知。我们还将展示如何呈现多重比较的结果和整齐的spacefriendly图形。在第4部分我们将提供一些特性试验的经验见解。

**2前期工作**

实验结果的统计评估被认为是新的机器学习方法验证的主要部分已经有很长时间了。所用的试验却早已相当幼稚并且未经证实。虽然针对单一问题的一对分类器的比较程序已经提出将近十年了，但针对更多的分类和/或多个数据集的比较研究仍然是部分使用并且没有令人满意的解决方案。

2.1相关的理论工作

从该领域中引用最多的是Dietterich（1998）的一篇论文。在描述机器学习的统计分类问题后，他的重点是确定两种算法在给定的数据组进行测试，哪一种会产生更精确的分类。他考察了5个统计测试，并通过引进新制作的总结分析5×2CV t检验来得出分析结果，它克服了低估方差的问题并由此升高I型更传统的配对t检验超过通常的k折交叉验证的褶皱的错误。

多次运行情况下的算法是不恰当的，Dietterich发现McNemar检验对矩阵强大的5×2CV t检验的分类是错误的。他提醒t检验不要重复随机抽样，也鼓励交叉验证之后采用t检验。该5×2CV t检验已经由提出更强大5×2CV F检验的Alpaydın（1999）进行改进使其具有较低的I型误差和更高的功率。

Bouckaert（2003）认为理论上的自由度是不正确的，因为依赖实验和经验发现值应该被代替使用，而纳多和Bengio（2000）提出的校正重采样t检验用于调整方差正是基于所述

实例子集之间的重叠。 Bouckaert和弗兰克（Bouckaert和Frank，2004; Bouckaert，2004）也研究了机器学习实验并举一反三发现了5×2CVt检验的不满之处，还选择了校正重采样t检验。评价k折交叉验证方差问题的更普遍研究，见证了Bengio和Grandvalet工作（2004）。

没有上述研究就不能评估多个分类器的性能也不能研究多个数据集的分类器在测试时的统计适用性。对于前者，Salzberg（1997）提到的ANOVA可以作为一种可能的解决方案，但事后他描述了二项式检验与Bonferroni校正的多重比较。他本人指出，二项测试缺乏较好的非参数检验，Bonferroni校正则过于激进。比如说巴斯克斯等人（2001）和皮萨罗等人（2002年）使用ANOVA和弗里德曼试验用于单一数据集的多个模型的比较（尤其是神经网络）。

最后，对多个数据集分类的比较，根据我们现有的知识，赫尔（1994年）是第一个在信息检索与文档相关性评估中使用非参数检验来比较分类器的人（也参见Schutze等人，1995）。 Brazdil和苏亚雷斯（2000年）使用平均排名比较分类算法。目标是选择的一个优化算法，不需要进行统计学检验它们之间的显著差异。

2.2测试实验：ICML论文分析

我们分析了最近国际会议关于机器学习（1999- 2003年）的五篇论文。我们的重点是，通过比较至少两个分类的文件测量其分类准确率，均方误差，AUC（Beck和舒尔茨，1986年），精度或其他一些模型的性能得分。

在这项研究中用于评价分类器性能的抽样方法和措施不直接相关。惊人的是，虽然这种分类精度仍然是唯一使用的衡量标准，但是医疗（Beck和舒尔茨，1986年；Bellazzi和祖潘，1998）和机器学习领域（教务长等，1998;兰利，2000年）却呼吁其他的措施，例如AUC也应该被采用。分类精度唯一真正的竞争是文档检索领域中使用的措施。这也是唯一允许使用单独的测试数据集，而不是使用横验证或随机抽样的领域。

我们文献最感兴趣的是分析算法之间差异的方法。研究论文通常以表格形式公布多个数据集的两个或更多个分类器的结果。我们没有记录分类器整体性能包含有多少（非正式）的声明。然而，从四分之一到一半的文献包含一些统计方法无论是确定最佳的方法还是比较它们之间的性能。

比较分类器最简单的方法是计算所有数据集的平均值，这种平均法比较幼稚且较少采用。配对t检验大概是评估差异的统计学意义唯一使用的方法。它们分为三类：只有两种方法可用于比较，第三种方法（一种新方法或基本方法）是被比较的，或者三种方法进行相互比较。尽管对多个假设重复警告测试，Bonferroni校正每年仅被用于少数ICML论文中。一个常见的​​非参数方法是计算算法执行得更好，更坏或相等的次数，次数有时是成对出现的，从而产生一个胜/平局/损失计数的矩阵，而另一种方法是计算优于其他算法的数据集的数目。有些作者喜欢只计算统计学显著的差异；为了验证这一点，他们使用各种技术对两种算法进行了审查比较。

这些数据需要谨慎使用。一些论文没有明确说明所使用的取样和测试方法。此外，它经常很难判断在一个特定取样过程中，测试或测量的质量是否等同于一般的取样过程。

1999年2000年2001年2002年2003年

论文总数 54 152 80 87 118

我们研究相关论文 19 45 25 31 54

抽样方法[％]

交叉验证，留一法 22 49 44 42 56

随机抽样 11 29 44 32 54

独立的子集 5 11 0 13 9

分数函数[％]

分类准确率 74 67 84 84 70

分类的准确性 68 60 80 58 67

精度 21 18 16 25 19

ROC，AUC 0 4 4 13 9

偏差，置信区间 32 42 48 42 19

分类器整体比较[％] 53 44 44 26 45

数据集的平均值 0 4 6 0 10

t检验来比较两种算法 16 11 4 6 7

配对t检验vs其他 5 11 16 3 7

配对t检验每个vs每个 16 13 4 6 4

胜/平局/损失数 5 4 0 6 9

显著胜/平局/损失数 16 4 8 16 6

表1：国际会议关于在1999年至2003年机器学习的一篇文献概述。报告了适用于我们研究

的相关论文数量的百分比（三线及以下）。

**3统计和实验分类比较**

综述表明没有多个数据集分类器比较的既定程序。不同的研究者采用不同的统计和常识性的技术来判断算法之间的差异是真实的还是随机的。这一部分，我们将研究几个已知和不太为人所知的统计检验，并以我们的目的研究它们真正测量什么以及它们对数据做出的安全假设是否适用。

作为起点，两个或多个学习算法被运行在一个合适的数据集并使用分类精度，AUC或其他一些措施（见表2和6中的示例）进行评价。我们不记录这多个样本的方差结果，因此对采样方案不做任何假定。唯一的要求是，该编译结果需要提供每个数据集的算法性能的可靠评价。在通常的实验装置中，这些数字来自交叉验证或反复分层随机分割到训练和测试数据集。

用于评估单个数据集和多个数据集的两种分类器差异的试验之间存在着根本差别。当测试单个数据集时，我们通常计算在重复训练和测试随机样本的平均性能和它的变化。由于这些样品通常是相关的，在设计统计程序和测试时需要格外小心，以避免方差偏估的问题。

我们的工作是对每一个数据集的多个重采样，仅仅是用于评估性能得分而不是其偏差。方差的来源是（独立的）数据集在性能上的差异，而不是（通常依赖）样品，所以升高的1型误差并不是问题。因为多个重采样不偏置得分评估，各种类型的交叉验证或留一法程序可以被使用且没有任何风险。

此外，正确的统计测试是对单一数据集与多个数据集分类器的比较是不相关的，因为我们必须首先解决前一个问题，才能解决后一个问题。因为在多个数据集运行算法时自然赋予了一个独立测量样本，这种比较在单个数据集的比较中更简单。

我们还应该强调的是，“样本量”在以下部分将参考数据集使用的数量，而不是参考每个单独组的拉伸训练/测试样本的数目或实例的每一组数。因此样本大小可以小到5，并且通常远低于30。

3.1两种分类器的比较

在多个数据集的两种分类器比较的测试中，我们将讨论两点。提醒大家目前广泛使用的t检验作为一般概念上不恰当的，统计不安全的方法。由于我们最终将推荐的Wilcoxon（1945）符号秩检验，它会呈现更多的细节。另外，更少使用的测试是符号检验，它比秩检验更弱，但也有其独特的优点。其他信息是从不同的方面对分类器的统计差异进行描述，所以测试方法的选择不仅仅是基于统计是否适当，还要依据我们打算测量的对象。

3.1.1数据集的平均化

一些机器学习论文的学者计算横跨测试数据集的分类器的平均分类精度。韦伯（2000年）的论文表示，“在不同的领域错误率是否相称是有争议的，因此是否跨域的平均错误率是非常有意义的”。如果对不同数据集的结果没有可比性，其平均值是毫无意义的。有一个不同的情况是，研究算法在一组相关问题下的比较，如对某种疾病从不同的机构或不同的文本挖掘问题的医学数据库具有类似性质。

平均数也很容易受到异常值。它们允许分类器在一个数据集的优秀性能来补偿整体性能的不良，或者相反，在一个域的完全失败可以战胜大多数公平的结果。这种行为可能是比较理想的，但一般我们更希望分类器尽可能多的表现出许多问题，这使得数据集的平均化显得不恰当的。

由于没有多少论文报告这样的平均数，我们可以认为，社会各界普遍认为他们毫无意义。因此，平均值没有被使用（也无用）于进行统计推理z轴或t检验。

3.1.2配对t检验

一种常见的测试不同数据集的两个分类器结果的差异是否是非随机的方法是计算一个配对t检验，检查其数据集性能的平均差是否显著不同于零。

C*i*1和C*i*2表示N个数据集中第i个的两种分类器性能分数，*di*表示C*i*2 -C*i*1的差值。 t统计量是根据d /σD及带有N - 1自由度的学生分布所计算得到的。

在我们的论文中，t检验有3个弱点。第一个是对称性：只有当数据集的差异是对称的时候t检验才有意义。在这种观点下，使用配对t检验用于比较一对分类器来计算数据集的平均值将变得没有意义。平均差值d等于两种分类器的平均得分的差值，d = C2 - C1。这种形式的t检验和两个平均值的比较之间的唯一区别（如以上讨论的那些）就是在分母里面直接使用t检验无关样品：配对t检验减少由数据组之间的方差（或者换一种方式，由分类器之间的协方差）引起的标准误差σD。

韦伯（2000）通过计算几何平​​均值接近对称性问题的相对比例(∏*i c*1 *i* /*c*2 *i* )1/*N*。由于它等于*e*1/*N* ∑*i*(ln*c*1 *i* −ln*c*2 *i* )，这个统计在本质上和普通平均值是一样的，不同之处在于它比较的是对数得分。这种转变的效果因此相当值得怀疑。昆兰（1996）计算的算术平均值的相对比例；由于倾斜分布，在没有进一步操作时这些不能用于t检验。补偿不同问题的复杂性的一个更简单的方法是通过平均得分来分配差异，

*di* = (*c*1 *i* −*c*2 *i* )*/*[(*c*1 *i* +*c*2 *i* )/2]。

关于t检验的第二个问题是，除非样本大小足够大（〜30个数据集），配对t检验需要比较的两个随机变量之间的差别才是正态分布。我们的问题本质上并没有给出任何常态的条款，并且数据集的数量通常远小于30。讽刺的是，柯尔莫哥洛夫 - 斯米尔诺夫和类似的测试用于测量小样本的正态分布并没有用，也就是说，它们是不可能检测异常并提醒不要使用t检验。因此，我们只要满足正态分布就能使用t检验，因为我们有小样本，但小样品也禁止我们检查出分布形状。

C4.5 C4.5 +m 差值 排名

成人（样品） 0.763 0.768 +0.005 3.5

乳腺癌 0.599 0.591 -0.008 7

乳腺癌威斯康星 0.954 0.971 +0.017 9

CMC 0.628 0.661 +0.033 12

电离层 0.882 0.888 + 0.006 5

光圈 0.936 0.931 -0.005 3.5

肝病 0.661 0.668 +0.007 6

肺癌 0.583 0.583 0.000 1.5

淋巴 0.775 0.838 +0.063 14

蘑菇 1.000 1.000 0.000 1.5

原发肿瘤 0.940 0.962 +0.022 11

大黄 0.619 0.666 +0.047 13

投票 0.972 0.981 +0.009 8

酒 0.957 0.978 + 0.021 10

表2： C4.5与 m = 0的AUC和C4.5与m调整到最佳的AUC的比较。该表的右手列说明了通常不会被公开在实际论文中的计算。

第三个问题是，t检验仅表示数据集的平均化，受到离群的影响，它歪斜检验统计并通过增加估计的标准误差降低测试的功率。

3.1.3 Wilcoxon符号秩检验

Wilcoxon符号秩检验（秩，1945）是一种非参数替代配对t检验，它列出了每个数据集的两个分类器的性能差异，忽视标志，并对正和负差异的行列进行比较。

*di* 还是表示N个数据集中第i个的两种分类器的性能分数的差值。根据其绝对值排名；为避免平局对平均行列进行分配。R +表示第二种算法优于第一种算法时，数据集行列的总和，而R-则相反，表示第一种算法优于第二种算法时，数据集行列的总和。*di* = 0的行列被平分到总和之中；如果其中有奇数，它就被忽略：

T表示总和中较小的一个，T = MIN（R +，R-）。大多数书籍上的一般统计数据包括：关于T 和高达25（有时更多）的N的确切临界值的一张表。对于大量的数据集，统计

分布大致正常。由于α= 0.05，如果z小于-1.96，零假设是可以被拒绝的。

让我们以一个过程进行举例说明。表2表示C4.5与m（在一片叶子中的最小数目）设定为零的AUC和C4.5与m调整到最佳的AUC的比较。对于后者，由于m∈{0，1，2，3，5，10，15，20，50}，通过5倍内部交叉验证训练可以计算出AUC。实验是根据UCI资源库和二进制类属性的14个数据集展开进行的。我们原来用的是昆兰的C4.5代码，装配与整合成机器学习系统橙色（Demsar和祖潘，2004）接口，这给我们提供了交叉验证程序，调整参数分类和评分功能。我们正试图拒绝零假设因为这两种算法表现同样出色。

有两个数据集的分类器表现效果一样（肺癌和蘑菇）；如果其中有奇数，我们将忽略它。行列式从最低到最高绝对差进行分配，相等的差值（0.000，±0.005）则分配平均到行列式中。

正差行列之和R + = 3.5+ 9+ 12+ 5+ 6+ 14+ 11+ 13+ 8+10 + 1.5 = 93和负差行列之和R- = 7 + 3.5 + 1.5 = 12。根据魏氏测试精确临界值的表，对于一个置信水平α= 0.05和N = 14的数据集，如果总和较小的数小于或等于21，则分类器之间的差异是显著的。因此我们拒绝零假设。

Wilcoxon符号秩检验比t检验更合理。但它只是定性地假定差异的对称性：较大的差异仍然需要更多的计算，这或许是可取的，但绝对幅值被忽略。从统计来看，Wilcoxon符号秩测试更安全，因为它不需要假设正态分布。此外，异常值（很多数据集的非常好/坏的表现）对魏氏效果的影响比对t检验更低。

Wilcoxon检验假定连续差异*di*，因此，它们不应该被舍弃一个或两个小数，因为这将减少测试的功率由于更多的平局。

当满足配对t检验的假设，Wilcoxon符号秩检验比配对t检验功率更小。另一方面，当假设被违反，Wilcoxon检验甚至可以比t检验更强大。

3.1.4获胜，失败和平局的计数：符号检验

一种流行的方式来比较分类器的整体性能是当一种算法全部获胜时计算数据集的数量。当多个算法进行比较时，两两比较通常建立在一个矩阵中。

有些作者在统计推断中也使用这些计数，用二项式检验的一种形式，被称为符号检验（Sheskin，2000; Salzberg，1997）。如果这两个算法被比较，根据零假设的假定，相当于每个算法都应该赢得大约N个数据集中的N / 2。胜的数目是根据二项式分布进行分配的；获胜的关键数字可以在表3中找到。对于更多数量的数据集，胜的数量服从零假设分布，它依据N（N / 2，√N/ 2），其允许使用的z测试：如果胜的数目至少是N / 2 +1.96√N/ 2（或者对于一个拇指的快速规则，N / 2 +√N），则这种算法对p<0.05明显更好。由于平局比赛支持零假设，在两个分类器之间我们不应该低估而是分裂他们；如果其中有一个奇数，我们再次忽略它。

表3：α= 0.05，α= 0.10双尾符号检验的临界值。如果在至少Wα数据集时表现更好，那么其中一种分类器显著优于另一种

由表2的例子，选自14个数据集中的11个C4.5 +m较好（当两个分类器打成平局时，数量也是两个数据集之一）。根据表3当p <0.05时这种差异很显著。

这个测试没有假定分数或差异的任何对称性，也没有假定正态分布，因此适用于任何数据（只要观察，即数据集是独立的）。另一方面，它比Wilcoxon符号秩检验弱得多。根据表3，符号检验不拒绝零假设，除非一种算法几乎总是优于其他。

有些作者喜欢只有显著获胜和失败的计算，对每个数据集采用统计检验可以得到其中的意义，例如Dietterich的5×2CV。推理这种做法的背后是因为“有些输赢是随机的，这不应该算数”。如果统计测试可以区分随机和非随机的差异，这将是一个有效的观点。然而，如果零假设是正确的，统计检验只测量实验结果的不可能性，而不是零假设的可（不）行性。

为了便于讨论，假设我们比较了一千个不同数据集的两种算法。在每一个情况下，算法A比算法B要好，但差异并不显著。事实就是对于每个单一的情况，两种算法之间的差异归因于机缘巧合，但其中一种算法怎么可能在1000个独立实验中每次运气都这么好？

与民间信仰相反，只计算显著获胜和失败，不作更多测试是相当不可靠的，因为它绘制了P <0.05时计算什么和不计算什么的任意门槛。

3.2多种分类器的比较

上述测试没有被设计用于推理有关多个随机变量的装置。然而许多机器学习论文的作者却出于这一目的而使用它们。这种可疑程序的一个常见的​​例子就是通过所有21配对t检验和报告的结果，如“A算法被发现比B和C显著好，并且算法A和E均比D显著好，但对于其他配对却没有显著差异 ”来比较7种算法。当进行了这么多测试，由于随机巧合零假设在一定比例上被拒绝，所以列出它们没有什么意义。

多重假设检验是公认的统计问题。通常的目标是控制家庭明智误差，使得在任何的概率比较中第一类型的错误至少出现一次。在机器学习文献中，Salzberg（1997年）提到了多重测试的一个通用的解决方案即Bonferroni校正，并指出它通常是非常保守且微弱的，因为它设定假设的独立性。

统计数据提供了更有力的专业程序，用于测试多重方法之间的显著差异。在我们看来，最有趣的两个是众所周知的方差分析和非参数对应的弗里德曼测试。后者，特别是其相应的

Nemenyi事后测试是不太为人所知，并且有关它们的文献很少；出于这样的原因，我们将对它们进行更细致的介绍。

3.2.1方差分析

用于测试两个以上相关样品之间差异的常见统计方法是反复测量方差分析（或者被试内ANOVA）（费舍尔，1959）。“相关样品”还是表示相同的数据集测定的分类器性能，在训练和测试集中优选使用相同的分割。当所有分类器执行相同并且只观察到随机差异的时候，零假设可用于测试。

方差分析将总的可变性分配到不同分类器，不同数据集和剩余（误差）之间的可变性。如果分类器之间的可变性显著大于误差可变性，我们可以拒绝零假设，并得出结论：分类器之间存在一些差异。在这种情况下，我们可以用一个事后检验找到哪种分类器真正不同。对于方差分析的很多测试，对我们来说最适宜的两个是Tukey检验（杜克，1949），用于比较所有分类器彼此和Dunnett检验（Dunnett法，1980），用于控制所有分类器的比较（例如，比较基本分类器和一些改进建议，或用现有的几种方法比较新提出的分类器）。这两个过程都除以数据集的剩余方差来计算两个分类器之间差异的标准误差。为了得到分类器之间的成对比较，性能中相应的差异除以标准误差，并与临界值进行比较。这两个过程与t检验是如此相似，所不同的是Tukey和Dunnett法得到的临界值较高，以确保至多5％的机会使得其中一种成对差异被错误地认定是显著地。

不幸的是，方差是基于它们在分析机器学习算法性能时最有可能违反的假设。首先，方差分析假定采样从正态分布进行绘制。在一般情况下，也不能保证一系列问题的精度分布是常态的。事实上，即使分布不正常，这只是一个小问题，很多统计学家不会反对使用单因素方差分析，除非分配是明显地双峰（哈密尔顿，1990）。重复测量的ANOVA在第二和更重要的假设中是球形（在通常的方差分析中一个属性类似同性质的方差，这需要随机变量具有相等方差）。因为学习算法和数据集的这种性质并非是理所当然的。违反这些假设对事后测试甚至会造成更大的影响。因此ANOVA似乎并不是典型的机器学习研究的一个合适的综合性试验。

由于我们对参数检验保留意见，将不再一一介绍方差分析和更详细的事后检查，尤其是这些试验是众所周知的，并在统计文献中做了描述（扎尔，1998; Sheskin，2000）。

3.2.2 Friedman检验

弗里德曼试验（弗里德曼，1937年，1940年）是重复测量一个非参数等效方差分析。它对每个数据集的算法进行单独排名，表现最好的算法排名第1，第二好的算法排名第2**. . .**如表6所示，为防止平局（如虹膜，肺癌，蘑菇和原发肿瘤），要平均等级分配。

表示k个算法对第i个的N个数据集的第j个等级。 Friedman检验比较算法的平均排名， 。零假设规定所有的算法都是等效的，所以它们的排名R j也是相等的，弗里德曼统计 是根据 的 自由度，当N和k足够大时（作为一个经验法则，N> 10且k> 5）来进行分布的。对较小数目的算法和数据集，准确临界值已计算（扎尔，1998; Sheskin，2000）。

伊曼和达文波特（1980）指出，弗里德曼的 是保守不可取地，并派生了更好的统计 ，这是根据该F分布具有k-1和（k-1）（N-1）自由度来进行分布的。临界值的表可在任何统计书中找到。

至于两分类器的比较，当满足方差分析的假设时，（非参数）弗里德曼试验在理论上比（参数）方差分析功率更少，但这不是他们不满足假设的原因。弗里德曼（1940）实验相比ANOVA和他对56个独立问题的测试表明，这两种方法大体上一致（表4）。当其中一个方法在P <0.01时有显著性，另外一个方法在至少P <0.05时有显著性。只有在两种情况下ANOVA方差分析发现显著差异而弗里德曼没有发现显著差异，但是在第4种情况下则恰恰相反。

如果零假设被拒绝，我们就可以着手进行事后检验。当进行单因素方差分析及所有分类器彼此进行比较时，Nemenyi试验（Nemenyi，1963）与Tukey检验相似。如果相应的平均行列相差至少临界差， ，则两个分类器的性能明显不同，其中临界值 是被√2整除后的学生化的范围统计。

1. 双尾Nemenyi测试的临界值
2. 双尾邦费罗尼-Dunn检验的临界值；包括控制分类器在内的分类器数量

表5：弗里德曼试验后事后检验的临界值

当所有的分类器与控制分类器进行比较时，我们可以在多重假设检验中代替Nemenyi测试使用一般程序用于控制家庭明智错误，如Bonferroni校正或类似程序。虽然这些方法一般比较保守，并且功率很小，但它们在这种特殊情况下比Nemenyi测试功能更大，因为后者调整临界值用于k（k - 1）/ 2次比较，而与对照组相比我们只做了k – 1次比较。

比较第i个和第j个分类器的检验统计量所使用的方法是

z值被用于从正态分布表中找到相应的概率，然后与一个适当的α进行比较。试验的不同之处在于他们调整了α值来补偿多重比较。

Bonferroni-Dunn检验（邓恩，1961年）通过除以α得到比较的次数（一般是k- 1）来控制家庭明智的错误率。另一种方法是使用Nemenyi测试中的同一方程来计算相同试验的CD值，但使用的临界值为α/（k - 1）（为方便起见，它们列于表5（b））。Nemenyi试验和邓恩试验表格之间的比较表明，当所有分类器只和控制分类器而不是和它们本身进行比较时，事后检验的功率要大得多。因此，当我们实际上只测试一个新提出的方法是否比现有方法更好时，我们不应该进行两两比较。

对于单步骤邦费罗尼 - 邓恩过程的对比，升压和降压过程对假设的差异依次进行了测试。我们通过P1，P2，..来表示P值，使得P1≤P2≤...≤Pk-1。Holm（1979）和霍赫贝格（1988）想出了这种最简单的方法。他们将每一个Pi与α/（k-i）进行比较，但不同的是测试的顺序。霍尔姆的降压过程从最显著的p值开始。如果p1小于α/（k - 1），相应的假设被拒绝，我们就可以比较p2和α/（k - 2）的大小。如果第二个假设也被拒绝，就继续第三个测试，依次进行下去。直到某个零假设不能被拒绝，所有剩余的假设也将被保留。Hochberg的升压过程则正好相反，将最大的p值与α相比，第二大的p值与α/ 2比较，依次进行下去，直到它可以拒绝一个假设。所有较小的P值也可以拒绝这些假设。

霍梅尔过程（霍梅尔，1988年）对于计算和理解比较复杂。首先，对于所有的k = 1 .. j我们需要找到最大的j值使得 ，如果这样的j值不存在，我们可以拒绝所有的假设，否则我们就拒绝所有 的假设。

霍尔姆过程比Bonferroni - 唐恩功率更大，并且对于所测试的假设没有额外的假定。Bonferroni-Dunn检验的唯一好处在于它更容易描述和想象，因为它对所有的比较都使用了相同的CD值。反过来，霍赫贝格和霍梅尔的方法比霍尔姆的方法拒绝了更多的假设，但他们在一些情况下可能超过规定的家庭明智错误，因为它们是基于仍在调查中的西门斯猜想。据报道（荷兰，1991），已增强的方法之间的差异实际上相当小，因此更复杂的霍梅尔方法

并没有比简单的霍尔姆方法提供很大的优势。

虽然我们只对弗里德曼测试的事后测试使用了很多程序，但是当不同的类型用于多个假设的测试时，它们通常可以用于控制家庭明智的错误。也存在其他类似的方法，以及一些可以代替控制家庭明智的误差的方法，来控制错误拒绝零假设的数量（假发现率，FDR）。后者对于机器学习算法是不适合的，因为它们的评价要求研究者决定可接受的错误发现率。例如谢弗（1995）提出了更完整的正式描述所有这些的方法和讨论。

有时弗里德曼测试报告了显著差异，但事后检验却无法检测它。这是由于后者的功率太低。没有其他的结论比一些算法可以在这种情况下发现差异。在我们的实验中这种现象只发生在少数情况下。

程序记录在表6的数据中，比较了四种算法：C4.5与m固定为0和cf（置信区间）固定为0.25，C4.5与m在5倍内部交叉验证，C4.5与cf在5倍内部交叉验证，最后C4.5固定为两个参数，对它们的值进行所有的组合。参数m设定为0，1，2，3，5，10，15，20，50以及cf设定为0，0.1，0.25和0.5。

它们的平均等级为算法提供了一个公平的比较。平均来看，C4.5 +m和C4.5 + m+ cf排在第二位（分别是2.000和1.964），另外C4.5和C4.5 + cf排在第三位（分别是3.143和2.893）。 Friedman检验检查了测得的平均等级与零假设预期的平均秩R j = 2.5是否存在显著差异：

C4.5 C4.5 + m C4.5 + cf C4.5 + m + cf

成人（样品） 0.763（4） 0.768（3） 0.771（2） 0.798（1）

乳腺癌 0.599（1） 0.591（2） 0.590（3） 0.569（4）

乳腺癌威斯康星 0.954（4） 0.971（1） 0.968（2） 0.967（3）

CMC 0.628（4） 0.661（1） 0.654（3） 0.657（2）

电离层 0.882（4） 0.888（2） 0.886（3） 0.898（1）

虹膜 0.936（1） 0.931（2.5） 0.916（4） 0.931（2.5）

肝脏疾病 0.661（3） 0.668（2） 0.609（4） 0.685（1）

肺癌 0.583（2.5）0.583（2.5） 0.563（4） 0.625（1）

淋巴 0.775（4） 0.838（3） 0.866（2） 0.875（1）

蘑菇 1.000（2.5）1.000（2.5） 1.000（2.5） 1.000（2.5）

原发性肿瘤 0.940（4） 0.962（2.5） 0.965（1） 0.962（2.5）

大黄 0.619（3） 0.666（2） 0.614（4） 0.669（1）

投票 0.972（4） 0.981（1） 0.975（2） 0.975（3）

酒 0.957（3） 0.978（1） 0.946（4） 0.970（2）

平均排名 3.143 2.000 2.893 1.964

表6： C4.5与m = 0和C4.5与优化的参数m和/或cf的AUC比较。括号中的排名用于弗里德曼测试的计算，通常不会发表在实际论文中。

根据四个算法和14个数据集， 是根据4-1 = 3和（4-1）×（14-1）= 39个自由度的F分布来分配的。 F（3,39）对于α= 0.05的临界值是2.85，所以我们拒绝零假设。

进一步的分析取决于我们打算研究什么。如果没有单独的分类器，我们使用Nemenyi检验进行配对比较。临界值（表5（a））为2.569而相应的CD值是 。既然算法的最好和最差的性能之间的差异较小，我们可以得出结论，事后测试的功率不够大到足以检测算法之间的任何显著差异。

当P = 0.10，CD为 。我们可以找出两组算法：纯C4.5的性能明显比C4.5 + m和C4.5 + m + cf要差。我们不知道C4.5 + cf属于哪一组算法。如果说它同时属于两种算法则没有统计意义，因为主体不能同时来自两个不同的群体。正确的统计表明，实验数据不充足，以至于不能得出任何关于C4.5 +cf的结论。

收集数据之前的另一种可能的假说有可能通过调整参数来提高C4.5的性能。验证这一点最简单的方法是运用邦费罗尼-Dunn检验来计算CD值。在表5（b）中我们发现，四种分类器的临界值q0.05 是2.394，所以CD值是 。 C4.5 + m +cf的性能明显比C4.5（3.143- 1.964 = 1.179> 1.16）要好，但C4.5 + cf则没有（3.143- 2.893 = 0.250 <1.16），尽管C4.5 + m略低于临界差，但接近它（3.143 - 2.000 = 1.143≈1.16）。我们可以得出结论：实验表明固定m似乎有帮助，尽管我们没有检测到固定cf是否有显著提高。

对于其他测试，我们必须计算并对相应的统计和p值进行排序。该标准误差是

Holm程序拒绝第一和第二个假设是因为相应的P值比调整的α值小。第三个假设不能

被拒绝；如果有更多假设，我们也将保留它们。

Hochberg程序从底部开始。如果不能拒绝最后一个假说，就检验倒数第二个，直到可以拒绝它以及所有小于P值（最顶层的那一个）的假设。

最后，霍梅尔过程发现J = 3不满足k = 2的条件，J的最大值是2，最开始的两个假设可以被拒绝，因为它们的P值都低于α/ 2。

所有的降压和升压过程表明C4.5 + cf + m和C4.5 + m与C4.5存在显著差异，尽管邦费罗尼-Dunn检验发现C4.5和C4.5 + m非常相似。

3.2.3考虑实验的多次重复

在我们的实验中，我们对培训/测试集的重复过程使用了已测量和平均化的AUCs。例如，表6中的每个细胞代表了五倍交叉验证的平均值。我们还需要考虑方差或个人褶皱的结果吗？

考虑到多次观察每个细胞的方差分析和弗里德曼试验表明观测是独立的（扎尔，1998）。这种情况不适用在这里，因为多个随机样本在训练数据中会重叠。我们不知道有任何统计检验可能会考虑到这一点。

3.2.4结果的图形演示

当多个分类器进行比较时，事后测试的结果可以呈现在一个简单的图中。图1展示了表6中数据的分析结果，图中的顶部线条代表我们绘制方法的平均行列轴。轴可以转动，以便

最低（最好）的行列是正确的，因为我们认为在右侧的方法为好。

相互比较所有的算法，然后对差异不明显的算法进行连接（图1（a））。图表上方还展示了临界差。

如果使用Bonferroni-Dunn检验的方法与对照相比，我们可以标记间隔中的一个CD值到控制算法的平均等级的左边和右边（图1（b））。任何算法与该区域以外的等级是显著不同于对照的。其他事后测试的类似图表需要绘制出用于测试和比较的相应顺序，从而对每个分类器和指定程序做出不同调整的临界区间，这很容易变得混乱。

对于另一个例子，图2以图形表示了Mladenic和Grobelnik（1999）研究的来自雅虎层次的五个结构域的关键字预测的特征得分措施的比较。分析表明，信息增益的表现显著差于确凿证据，交叉熵TXT和赔率比例似乎有相同的性能。这些数据还不足以断定多重信息文本与信息增益或词频是否有相同表现，同样也不能断定词频与多重信息文本是否相等或者比三种方法更好。

1. Nemenyi测试中所有分类器相互比较。对没有显著差异（在P = 0.10）的分类器的连接

（b）邦费罗尼-Dunn检验中一种分类器与其他分类器的比较，所有分类器与等级外标记的间隔距离显著不同（P <0.05）于对照。

图1：表6中数据事后测试的可视化。

图2：召回的各种特征选择措施的比较；对Mladenic和Grobelnik论文（1999年）结果的分析

**4测试的经验比较**

我们实验中观察到所述试验的两个属性：可复制性和拒绝零假设的可能性。我们进行实验来回答诸如“哪些统计检验最有可能得到正确的结果”或“哪种测试具有类型1 /类型2的最低错误率”的问题将是毫无意义的，因为所提出的参考测试假定了不同的对称性，所以从不同的方面来比较分类器。因此拒绝或不拒绝零假设的正确答案，不能很好地确定，并且在一定程度上与选择的测试有关。

4.1实验装置

我们在实验中对10个随机抽取的数据集的学习算法进行反复比较，并返回到测试中记录P值来检验研究试验的行为。

4.1.1数据集和学习算法

我们根据实验对几种常用的学习算法及它们的变化：C4.5，C4.5与m和C4.5与固定最佳精度的cf，另一种学习算法用橙色来执行（类似原来的C4.5特征），朴素贝叶斯学习者使用连续的黄土概率模式（克利夫兰，1979年），朴素贝叶斯学习者结合连续属性概率使用法耶兹 - 伊拉尼的离散（法耶兹和伊拉尼，1993年）和最邻近离散（K = 10，邻与高斯核调整权重）。

根据UCI机器学习库（布雷克和梅尔茨，1998年）我们已编制了四十个真实数据集的样本，我们已经使用了数据集的离散类型并且避免人工数据集像和尚的问题。因为没有分类器对所有可能的数据集都最佳，在我们的模拟实验中，研究人员希望展示特定算法的独特优势并由此选择数据组对应的汇编。我们通过使用10倍交叉验证提前对所有数据集的分类器进行分类精度的测量。当比较两个分类器时，随机选择十个数据集样品使得数据集i被选择的概率正比于 ，其中d是这些数据集在分类精度上的差值（正或负），k表示偏差，通过它我们可以调节不同分类器之间的差异。尽管k = 0时数据集的选择是随机的均匀分布，k的值越高，我们更有可能选择有利于特定学习方法的数据集。需要注意的是选择数据设置与了解他们的成功（如预先估计）只是一个模拟，而研究人员将根据其他标准选择数据集。用在实际评估中描述的程序算法将被视为作弊。

我们决定避免专门用于检测统计测试而构建的“人工”分类器和数据集，例如由Dietterich（1998）所使用的。在这样的实验过程中，需要对真实世界的数据集和学习算法做出一些假设，而人工数据和算法则建立在模仿现实世可控的方式下。在我们的例子中，我们将以失败的概率对一组（可能是虚构的）数据集构建两个或多个分类器，这样我们观察1型和2个错误率的统计试验就可以知道正确的假设。

不幸的是，我们不知道我们对现实世界的假设是什么。什么程度下分类精度（或其他成功的措施）是不对称的？它们的正常分布是怎么样（ab）的？方差如何均匀？此外，如果我们要做某些假设，统计理论应该明确我们正在建立的实验结果。由于我们使用的统计测试在理论上很好理解，我们不需要检测现实世界的数据是否符合他们的假设。换句话说，从理论我们知道，对小样本的t检验（即对一个小数目的数据集）需要正态分布，所以通过构建人工环境产生非正态分布，我们可以使t检验失败。然而真正的问题是，现实世界的分布对t检验工作是否足够正常。

我们不能直接测试假设吗？正如已经提到的t检验的描述中，像柯尔莫哥洛夫 - 斯米尔诺夫检验对小样本的常态测试是不可靠的，他们极不可能检测到异常。即使我们确实有进行适当的测试，他们只计算（ab）常态分布，非均匀性差异的程度等等，而不是计算样品是否适合t检验。

我们的目的是用现实世界的学习算法和数据集未修饰的形式防止人为设定它们之间的差异，使它们故意错误分类一定比例的例子。它可以通过选择数据集的方法得到补偿：通过调整偏差来影响数据集的选择，以此来调整学习算法之间的差异。以这种方式，我们对现实世界的数据集和算法展开实验，然后观察统计表现不同程度的分类器之间的差异。

4.1.2功率和可复制性的措施

形式上，统计测试的功率被定义为测试将（正确地）拒绝零假设的概率。由于我们对什么是真正虚假的标准与所选择的试验（基于我们想要测量的分类器之间的差异）有关，我们只能观察到拒绝零假设的概率，这与功率无关。

我们从两个方面着手。首先，我们设置显著性水平为5％，并观察一千个实验中有多少个特定测试拒绝零假设。这种做法的缺点是，它只能观察到p = 0.05（这大概是我们感兴趣的）左右时的统计行为，它可能会错过一个更大的图片。因此，我们通过“功率”试验的另一项措施也观察到了P的平均值：值越低，越有可能在设定的置信区间拒绝零假设。

评估功率测试的两种测量方法使得可复制的两个措施相关。 Bouckaert（2004）提出了一个定义可用于拒绝零假设的计数。他所定义的可复制的概率是两个实验具有相同的一对算法将产生相同的结果，也就是说，这两个实验同时接受或拒绝零假设，并为这个概率设计了最优无偏差估计

其中ei是n个实验中第i个实验的结果（如果零假设被接受，ei为1，如果零假设被拒绝，ei为0），I表示指示函数，如果它的参数为真，I为1，否则为0。 Bouckaert还介绍了一种简单的方法来计算R（e）：n个实验中如果假设被接受的次数为p，而假设被拒绝的次数为q，R（e）等于 。当p= q = n / 2时出现最小值R=0.5，当p或q为0时出现最大值R=1.0。

这种测量方法的缺点是，当分类器之间只有轻微差异时，统计测试将显示出低复制性。当比较不同功率的两种测试时，结果更接近所选择α的那一个通常被视为不可靠。

当功率以p的平均值来估计时，可复制性自然是通过他们的方差来定义。 p的方差介于0到0.25之间；后者发生在一半的p值等于零，而另一半p值等于一。为了与Bouckaert的R（e）比较，我们定义相对p的方差的可复制性为

当我们在实验过程中使用这种可复制性的测量方法时，会出现的一个问题是当偏差k增加时，数据集选择的变异性减小，p的方差也是如此。这种效应的大小取决于数据集的数目。通过实验结果判断，我们收集的40个数据集足够大，可以保证所使用的k值（参照图4.c左边图形；如果选择的变异性下降，p的方差不能保持恒定）几乎不受到这种变异性的影响。

对可复制性定义的描述是相关的。因为I（ei = e j）等于 ，我们

可以改写R（e）为

这里很容易证明

事实上，Bouckaert公式是关于R（e）和Σi（ei-e）2 /（n-1）的最佳无偏差估计式，被用于总体方差的最佳无偏差估计。

4.2两种分类器的比较

我们已经对两个分类器的比较做了四次统计测试：绝对差异与相对差异的配对t检验，Wilcoxon检验和符号检验。如上所述，该实验对十个数据集选择进行了1000次随机运行。

图3左手侧的曲线图显示了当比较C4.5-cf，朴素贝叶斯分类器和kNN（说明当一个函数规模倒置，测试功率增加时曲线上升）时，作为偏差k的一种功能返回到测试的P的平均值。该图的右手侧曲线反映了α= 5％时假设被拒绝的实验次数。为了证明功率（如我们测量的）和Bouckaert可复制性的测量方法之间的关系，我们增加了右轴，显示R（e）对应的拒绝假设的数量。

需要注意的是k = 0时的实验数目，并不对应于零假设被拒绝50％。k设置较低并不意味着两种算法比较时应该产生同样效果，但只有那样我们不（人工）偏向数据集的选择，使它有利于其中之一。因此，k = 0的测试反映出我们收集的数据集在完全随机选择的情况下拒绝零假设的数量。

t检验的两种方差得到类似的结果，和相对差异略微但一直较弱的试验。Wilcoxon符号秩检验提供了较低的P值，它比t检验在几乎所有情况下更容易拒绝零假设。从理论上已知的符号检验，则比其他测试要弱得多。

可复制的两种测量方法给出了完全不同的结果。根据R（p）判断（图4左手侧曲线图），Wilcoxon检验表现出P值的最小变化。作为对照，Bouckaert的R（e）（图4右手侧曲线图）表明Wilcoxon检验最不可靠。然而，这些图表的曲线形状以及图3的右轴清楚地表明当P值接近0.05时，测试不可靠（根据R（e）），所以Wilcoxon检验似乎不可靠，是因为其较高的功率使它比其它测试更接近p= 0.05。

表示k设置为15时七个分类器的比较，对角线以下的数字显示平均P值和相关的可复制的R（p），对角线以上的数字表示α= 5％时零假设被拒绝的实验次数以及相关的R（e）。该表再次表明，Wilcoxon检验相对其他测试几乎总是返回低P值，更多的时候拒绝零假设。根据R（p）的测量，Wilcoxon检验还具有最高的可复制性 R（e），在另一方面，也比其他测试更喜欢使用远离临界0.05的P值。

总体而言，已知参数测试比非参数测试更有可能拒绝零假设，除非违反了他们的假设。我们的研究结果表明，后者的确发生在机器学习研究跨越数据集的集合进行的算法比较。因此我们建议采用Wilcoxon检验，除非满足t检验假设，或者是因为我们有很多的数据集，或者是因为我们有理由相信跨越数据集的性能测量是正态分布。作为第三个选择的符号检验，因为太弱一般是没用的。

R（e）的低数值表明，我们应该在尽可能多的合适数据集下展开试验以保证结果的可靠性（特别是当分类器之间只有轻微差异时）。

4.3多种分类器的比较

对于多个分类器的比较，数据集的样品被​​选择的概率根据C4.5和带有法耶兹 - 伊拉尼离散化的朴素贝叶斯分类器的分类精度的差异计算得到。选择这两个分类器并没有特别的理由；我们已验证这种选择对结果没有实际影响。

结果展示在图5中。当两种算法很相似时（k的值较小），非参数弗里德曼测试再次显示比参数方差分析更强。当两种算法之间有较大差异时（在我们的实验装置中k在10附近），方差分析赶上了弗里德曼测试，并且两个试验得到类似的结果。

两个分类器比较的功率统计测试。左：P值作为偏差（k）的函数。右：零假设被拒绝（左轴）和Bouckaert R（右轴）的次数。

两个分类器比较的可复制性试验：基于R（p）的方差（左）和Bouckaert的R（e）（右）。

两个分类器比较的试验：平均p值和R（p）（对角线以下），零假设被拒绝的次数和

R（e）（对角线以上）。

方差分析和弗里德曼测试的对比

当通过R（p）测量时，弗里德曼试验的可复制性比方差分析更高，因为由可比较的

R（e）测定时两种测试具有类似的功率。总之，可复制性有时候似乎比两个分类器比较测试的可复制性要小一些。因此，常识表明，当比较多个分类器时，在尽可能多的数据集中展开测试显得更重要。

显示出参数Tukey和非参数Nemenyi测试之间的比较。我们计算他们拒绝C4.5-cf和朴素贝叶斯分类器等价以及C4.5-cf和kNN等价的次数（朴素贝叶斯分类器和kNN的比较包括之前的测试一点也不有趣，因为零假设很少被拒绝）。左边的两个曲线图表示，实验的选择是基于两种算法在图表之间的差异，而对于右边的两个图表，我们使用C4.5-cf和测试的另外六个分类器的平均水平之间的差异。在所有情况下，我们比较了这七个算法，但只在图形上呈现了成对拒绝的数量。非参数检验比参数检验更频繁地拒绝零假设。

我们不显示P值和相应的可复制性，因为由于测试的不同顺序它们不能总是被计算或被比较。

对事后检验比较与控制分类器进行了比较，如图6中使用了相同的两个选择的数据集，当差别较大，所有测试的功率有可比性，而当差异变小时，参数测试被拒绝的数量似乎

落后了（当对算法进行其他组合时，我们已观察到同样的模式）。非参数测试的顺序是按照理论预期的，有趣的是，霍尔姆和霍赫贝格测试给出几乎相同的结果。

多分类器比较的功率统计测试。偏差被定义为图表上的两个分类器在性能上的差异（左）或C4.5-cf和所有其它分类器之间的差异（右）。每个图上的左标给出假设被拒绝的次数，而右边的刻度给出相应的R（e）。

这些实验似乎再次使非参数检验比参数检验有利，虽然在两个分类器比较的情况下并不总是令人信服。因为Friedman检验在理论和实践上的优势（便于计算和解释，以等级而不是可疑平均值的形式来描述分类器的整体性能的能力），弗里德曼试验应优先于方差分析。相应的非参数事后测试给出类似的结果，因此由研究者决定稍微强大的霍梅尔测试是否比相对简单得多的霍尔姆测试更值得其计算的复杂性。

多种分类器与对照组比较的功率统计检验。偏差被定义为图表上的两个分类器在性能上的差异（左）或C4.5-cf和所有其它分类器之间的差异（右）。每个图上的左标给出假设被拒绝的次数，而右边的刻度给出相应的R（e）。

**5.结论**

我们对过去国际会议关于机器学习论文的分析表明，许多作者认为他们提出的算法应该对一系列问题进行比较并且该结果可用于得出一般性结论。然而，没有黄金标准作出这种比较，而且所进行的测试经常存在可疑的统计基础并导致无根据的和未经证实的结论。

当使用被偏方差估计缠绕的单一的数据集来做比较，是因为数据集样本例子之间的相关性，在多个数据集的比较中，方差来自数据集之间的差异，它们通常是独立的。因此我们的设置免于增加的1类型错误，这在单一数据集的测试程序中很普遍。多个数据集测试的问题则完全不同，甚至在一定意义上是互补的：对不同数据集的测量通常是不适应的，其分布的正态性和方差的同质性是有问题的。

我们从理论和实践上分析了可用于多个数据集的两个或多个分类器比较的三个熟悉的统计检验：参数测试（配对t检验和ANOVA），非参数检验（采用Wilcoxon和弗里德曼试验）和非参数检验，它假定结果免于对称性（符号试验）。在理论部分，我们通过一个典型的机器学习数据专门讨论了可能违反的测试假设。基于已知的试验统计特性和我们对机器学习数据的知识，我们的结论是，非参数检验应优先于参数检验。

我们观察到在多个现实世界的分类器和数据集中提出的统计数据的行为。我们通过偏差数据组的选择来改变分类器之间的差异，并测量拒绝零假设的可能性和测试的可复制性。我们的确发现，非参数检验更可能拒绝零假设，这暗示由于离群的存在和违反参数测试的假设证实了我们对他们理论的疑虑。经验分析也表明测试的可复制性可能是一个问题，因此实际的试验应该在尽可能多的数据集中进行。

在经验研究中，我们没有提供关于类型1/2错误率的分析。主要原因是拒绝或不拒绝零假设的正确结果，并没有明确定义并取决于我们打算测量的算法之间的差异。此外，实验中我们知道正确的假设都需要人工数据集和规定概率和错误分布的分类器。为此，我们需要

对现实世界分布做出一些假设；这些假设正好是我们在测试中摆在首位的。

总体而言，非参数检验即秩和Friedman试验都适合我们的问题。他们是合适的，因为他们假设了一些但是有限的对称性。他们比参数检验更安全，因为它们不假设正态分布或方差的同质性。这样一来，它们可以被应用到分类精度，误差率或任何其他的分类器评价测量，甚至包括模型尺寸和次数计算。经验结果认为他们比研究的其他测试更强。当比较一对分类器时后者尤其正确。

当多个分类器比较时，我们已提出一个事后分析的视觉代表。 CD图“空间友好”，因此适合纸张长度有限的情况下，但他们呈现的算法顺序，它们之间差异的幅度（以行列而言）和观察到的差异比文本或纯数字形式更明显。

有统计人员认为不应该进行显着性检验，因为它们经常被误用，无论是误解或过分强调结果（科恩，1994；施密特，1996；Harlow和Mulaik，1997）。我们的立场是，统计检验对所公布结果的有效性和非随机性提供了一定的保证。如果这是真实的，他们应该被正确执行，并小心所得结论。另一方面，统计测试不应该是赞成或反对出版工作的决定的因素。该算法的其他优点是超出把握的统计测试也应该考虑，甚至可能超过预测能力之内的纯改进。

参考文献

1. JamesCopestake. Ope rationalizing Micro-finance: Women and Craftwork in Ifugao,Upland Philippines [J] ? Human Organization,2007, (60):212-224
2. 王磊.人工神经网络原理及其应用 [J].科技咨询[3]tursjacky AWS VPC”
3. Yi Sun,Yuheng Chen,Xiaogang Wang,Xiaoou Tang. Deep Learning Face Representation by Joint Identification-Verification(Nips2014)