硕 士 研 究 生 读 书 报 告



题目：分类器在多个数据集下的统计比较

Statistical Comparisions of Classifiers

over Multiple Data Sets

作者姓名  汤建华

作者学号 nb15117

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○15 年 12 月

摘要

虽然对单个数据集下两种学习算法比较的方法已被研究过相当长的一段时间。典型的机器学习研究对于在多个数据集下的统计检验问题更为重要，但却一直都却忽略了。本篇文章查阅了当前的方法然后从理论和实证上考察几个合适的测试。在此基础上，我们为分类器在多个数据集下的统计比较推荐了一套简单而安全，稳定的非参数测试分类统计比较：Wilcoxon比较了两个分类器的分级测试；弗里德曼测试相应的事后检验，涉及到更多个数据集比较分类。后者的结果也可整齐地呈现新引入的CD（临界差）图。

关键字：比较学习，统计方法，维尔康森等级测试，弗里德曼测试，多项比较测试

Abstract

While methods for comparing two learning algorithms on a single data set have been scrutinized for

quite some time already, the issue of statistical tests for comparisons of more algorithms on multiple

data sets, which is even more essential to typical machine learning studies, has been all but ignored.

This article reviews the current practice and then theoretically and empirically examines several

suitable tests. Based on that, we recommend a set of simple, yet safe and robust non-parametric

tests for statistical comparisons of classifiers: the Wilcoxon signed ranks test for comparison of

two classifiers and the Friedman test with the corresponding post-hoc tests for comparison of more

classifiers over multiple data sets. Results of the latter can also be neatly presented with the newly

introduced CD (critical difference) diagrams.

Keywords: comparative studies, statistical methods, Wilcoxon signed ranks test, Friedman test,

multiple comparisons tests

1 前言介绍

过去几年之中,机械学习社区越来越意识到对已发表结果的统计验证的需要。这种方法可以运用到很多成熟的领域：大量增加的现实生活的应用以及可用于一些现有的机器学习框架，这样可以更简单的修正已有的算法和开发新的算法，并且在彼此之间进行比较。

在典型的机器学习的论文当中，一个新的机器学习算法，它其中的一部分或者前部分或者后部分已经都被提出了。因此我们得到一个隐含的假设：这种提高可以促成现有算法的一个效率的提高。并且提出了很多解决问题的方法，希望从现有的失败中找到成功的方法。论文选择了一些数据集用于测试，然后运行算法，最后通过后使用合适的方法去评估模型的结果质量，这种是最常见的准确分类。论文的剩下的步骤是使用统计学的方法去验证表现的提高建设。

剩下的步骤是探索相关理论工作和现有方法。很多研究院强调了基于单数据集的分类器比较并且提出相应的一些解决方案。这些信息被大家采取了。从而过于自信地运用t检验交叉验证McNemar检验5×2交叉验证。另一方面，通过多数据集来比较分类器，这种情况往往更为常见，尤其是当我们去测试整体的表现而并非某一特别方面的表现的时候，这种方法仍然没有在理论上进行开发，从而遗留下很多需要完善的步骤，这些步骤要么缺乏统计基础，要么在不合适的领域运用了统计理论。因此，为了去研究在实际操作中我们应该如何去运用这种方法，我们研究了最近的国际会议的机器学习理论（1999-2003）。我们观察到需要优秀和有创新的机器学习论文都是在一些矩阵中得出的结论为结尾，比如说：麦克尼马尔比较所有成对的分类器的实验，与多对比较试验相似，比如ANOVA和弗里德曼测试还未问世。

论文的核心是统计测试的研究，这种测试可能会被（或已经被）用于两组或者多组的数据集的比较。正式的说，假设我们已经测试了在N数据集的k学习算法，在第三部分我们可以观察到每格测试背后的系统的假设。尽管有些测试在机器学习领域是比较简单的，常见的，很多研究人员普遍会忽略测试究竟是为了测量什么和他们运用于哪些情形之下。我们也会展示怎么运用一些简洁的图标去呈现多项比较的结果。第四部分我们会提供一些对于多项比较的经验看法。

2 准备研究

很长一段时间，实验结果的统计评估已被认为验证新的机器学习方法的主要部分。所用的试验却是简单和未经证实的。虽然程序一对分类器对比较单一问题已经提出了将近十年，比较研究，更多的分类和/或多个数据集仍然使用部分和不能令人满意的解决方案。

2.1 相关理论研究

Bouckaert（2003）认为，理论上的自由度是不正确的，由于依赖实验和经验发现值之间应该使用，而纳多和Bengio（2000）提出的校正重采样t检验用于调整基于所述方差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举一反三，发现了5×2CVt检验不满意和选择了校正重采样t检验。有关的一般的工作估计k折交叉验证方差的问题，详见Bengio和Grandvalet工作（2004）。

上述研究处理评估多个分类和性能都不研究的统计数据时分类器测试了多个数据集的适用性。对于前者的情况下，Salzberg（1997）中提到的ANOVA为一体的可能的解决方案，但事后描述了二项式检验与Bonferroni校正多重比较。作为Salzberg本人注意到，二项式检验缺乏更好非参数检验和邦费罗尼的功率修正过于激进。 V'azquez （2001）和皮萨罗等（2002），例如，使用ANOVA和弗里德曼试验多个模型的比较（尤其是，神经网络）上一个单一的数据集。最后，对分类在多个数据集的比较，赫尔（1994年）提出，谁使用非参数检验在信息检索比较分类器的第一与文档的相关的评估。 Brazdil和苏亚雷斯（2000）使用的平均排名比较分类算法。追求选择的一个不同的目标优化算法，它们不具有统计学检验它们之间的差异的显着性。

2.2 实际测试：ICML论文分析

我们研究了机器学习最近几次国际会议的程序分析论文（1999- 2003年）。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/召回或其他一些模型的性能得分。用于评价分类器的性能的抽样方法和措施不直接相关的研究。得出的结论是惊人的，虽然这种分类精度通常仍唯一的衡量标准使用，尽管从医疗（Beck和舒尔茨，1986年的声音; Bellazzi和祖潘，1998年）和机器学习领域（1998;兰利，2000年）敦促其他的措施，例如AUC，使用效果较好。对分类的唯一真正的竞争精度是在文档检索的区域中使用的措施。这也是唯一的地方，其中数据的丰度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。

更引起我们兴趣的是论文的方法是算法之间的差异分析。所研究论文发表了两个或更多个分类器的结果在多个数据集，通常以表格形式。我们没有记录有多少人包括（非正式）有关声明说明分类器的整体性能。然而，从一个季度的论文中至少是有一半包括一些统计方法要么是确定最佳的方法要么是比较它们之间的性能。最简单的方法来比较分类是计算平均值的所有数据集这种方法比较简单和较少应用。配对t检验即将是使用的唯一方法去评估差异的统计学意义。它们分为三类：只有两种方法被进行了对比。一种方法（一种新的方法或基方法）是与其他的比较，或者所有方法之间相互比较。尽管存在多个假设重复警告测试，Bonferroni校正每年被用于少数ICML论文。一个常见的非参数方法是计数的算法执行得更好，更坏或同等表现次数;数量是有时成对从而导致胜/平/输计数的一个矩阵，而另一种方法是计算的在其上算法优于所有其他数据集的数目。有些作者喜欢只计算有统计学显著的差异; 为了验证这一点，他们使用各种技术的两种算法上面进行了审查比较。

这个数字需要谨慎去使用。一些论文没有明确说明取样和测试方法。此外，我们会经常很难决定是否一个特定取样过程中，测试或测量的质量等同于没有。

1. 统计和试验分类器的比较

综述表明，没有既定程序对多个数据集进行比较的分类。不同的研究者采用不同的统计和常识性的技术来决定算法之间的差异是否是真实的还是随机的。在本节中，我们将研究一些已知和鲜为人知的统计检验，并研究它们是否适合我们从和的他们真正衡量其安全点就他们让攻方数据的假设目的。

作为起点，两个或多个学习算法被运行在一组合适的数据组，并使用分类精度，AUC或一些其它测量（见表2和6中的示例）进行了评价。我们不记录这些成果在多个样本的方差，因此承担任何关于采样方案。唯一的要求是，编译的结果提供的算法上的每个数据集的性能的可靠的估计。在平时的实验装置，这些数字来自交叉验证或反复分层随机拆分到训练和测试数据集。

有用来评估对单个数据集两个分类和在多个数据集的差异之间的差的测试之间的根本区别。当在单个数据集的测试，我们通常计算均值的性能和它的方差超过重复训练和测试上随机样本的实例。由于这些样品通常是相关的，很多护理的需要设计的统计程序和测试，避免出现偏差的偏估计。

在我们的工作，从每一个数据集的多个重采样仅用来评估业绩得分，而不是它的方差。方差的来源是超过（独立）的数据组中的性能的差异，而不是在（通常依赖）的样品，所以升高的1型误差不是问题。因为多个重采样不偏置计分估计，各种类型的交叉验证或离开一法程序，可以使用没有任何风险。

此外，对于在单一数据集进行比较正确的分类统计检验的问题是不相关的，在某种意义上多个数据集，我们首先必须解决前一个问题，以解决后者的比较。自运行算法对多个数据集自然给出的独立测量的样品，这种比较比对单个数据集的比较更简单。

我们还应该强调的是，“样本量”在以下部分将参考使用的数据集的数量，而不是从每个单独的组，或在各组的实例数绘制训练/测试样本的数目。因此样本大小可以是小到5，并且通常远低于30。

3.1比较两个分类器

在测试的两种分类到多个数据集的比较讨论，我们将提出两个点。我们要警惕目前广泛使用的t检验的一般概念上不恰当的，统计上是不安全的。由于我们将最终推荐的Wilcoxon（1945）符号秩检验，它会呈现更多的细节。另外，更是很少使用的测试是符号检验比Wilcoxon检验较弱，但也有其独特的优点。其它邮件将是从不同的方面进行分类，所以测试的选择之间的所述统计度量的差异应不仅基于统计是否适当，而且就我们打算测量。

3.1.1平均超过数据集

机器学习论文一些学者计算整个测试数据集分类的平均分类精度。在口头上ofWebb（2000年），“这是值得商榷的不同域中的错误率是否相称，因此无论是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。不同的情况下，是研究中，这些算法进行对一组相关的问题，如对某种疾病从不同机构或具有类似性质的各种文本挖掘问题医学数据库相比较。

平均数也很容易受到异常值。它们允许分类的上一个数据集优异的性能，以弥补整体表现不好，或者相反，在一个域中的彻底失败可以凌驾于公平的结果对大多数人。有可能是在其中这样的行为是期望的，而在一般我们可能更希望表现良好在尽可能多的问题成为可能，这使得平均超过数据组不合适的情况下，分类器。

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，平均值也未使用的（也有用）用于与Z-或t-检验统计推断。

3.1.2配对t检验

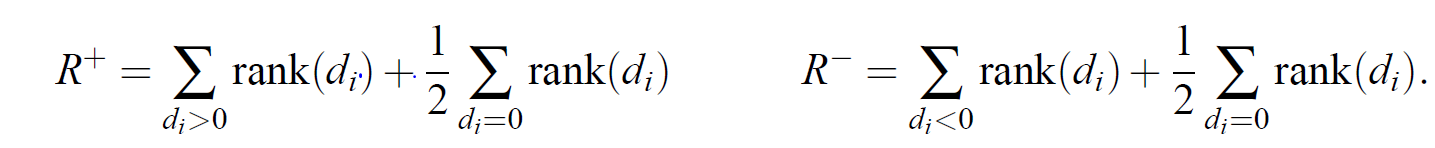
测试在各种数据集两个分类'的结果之间的差是否是非随机的常见方法是计算一个配对t检验，检查在其性能上比数据组的平均差值是否显著不同于零。让文C1 i和C2I是两个分类性能得分上的第i个选自N个数据集，并让二是差C2I-c1i。t统计量计算为d/ SD和是根据与N-1个自由度的学生分布。

第二个问题与t检验是，除非样本大小足够大（大于30的数据集），配对t检验需要比较的两个随机变量之间的差异是正态分布。我们的问题的性质不给予正常的任何条款和数据集的数量通常远低于30。讽刺的是，柯尔莫哥洛夫 - 斯米尔诺夫和类似的测试进行测试分布的常态没有能力对小样本，也就是，它们不可能检测异常和警告不要使用t检验。因此，对于使用t检验，我们需要正态分布，因为我们有小样本，但小的样品也禁止我们检查分布形状。

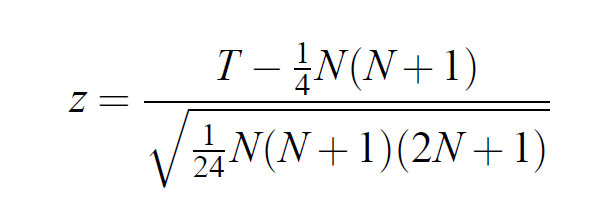
第三个问题是，t检验是，正如平均超过数据集，受异常值而歪斜检验统计量，并通过增加所估计的标准误差降低测试的功率。

3.1.3 Wilcoxon符号秩和检验

TheWilcoxon符号秩检验（秩，1945）是一种非参数替代配对t检验，居中的两个分类为每个数据集的性能上的差异，无视标志，并且为正，比较的行列消极的差异。让二再次是对第i两个分类器的性能分数之间的差异从N个数据集。的差异，根据其绝对值排名;平均等级分配的情况下联系。令R+是行列为在其第二算法胜过所述第一数据集的总和，和R-等级的总和为相反。di=0被平分的总和之中;如果有其中的奇数，忽略。



令T的款项，T= MIN（R +，R-）的较小值。大多数书籍上的一般统计数据包括对于T高达25（有时更多）准确临界值对于N的表。对于数据集，统计数量是正项分布的，假设x=0.05，若z小于-1.96，否假设会被拒绝。



让我们举例说明在一个实施例的过程。表2示出了用于C4.5具有m与米调整到最佳的AUC AUC的比较（在一个实施例叶的最小数目）设定为零，C4.5。对于后者，AUC已经计算与训练样例5倍内部交叉验证平方米{0,1,2,3,5,10,15,20,50}。实验是从UCI库二进制类属性14的数据集进行。我们原来用的昆兰的C4.5代码，配有集成入机器学习系统橙色（Demsar和祖潘，2004），这为我们提供了交叉验证程序，类调整参数，和计分函数的接口。我们正试图拒绝零假设这两种算法同样表现出色。

有在其上的分类器表现同样（肺癌和蘑菇）两个数据集;如果有其中的奇数，我们将忽略之一。行列从最低分配到最高绝对差，以及相等的差异（0.000，±0.005）被分配平均行列。

行列用于正差之和R += 3.5+ 9 +12 +5+ 6 +14 +11+ 13+ 8 +10 +1.5 =93和行列为负差之和等于R-=7 +3.51.5= 12。根据用于所述魏氏的测试精确的临界值的表中，以a =0.05和N=置信水平14数据集，所述分类器之间的差异是显著如果和的较小等于或小于比21。因此，我们拒绝零假设。

TheWilcoxon符号秩检验比t检验比较明智的。它假定公度的差异，但只能定性：有较大的差异还是更个性化，这可能是需要的，但绝对幅度被忽略。从统计上看，测试是更安全的，因为它不假定正态分布。此外，异常值（在几个数据集特别好/坏的性能）对魏氏效果比对t检验更低。

**References**

E. Alpaydın. Combined 5×2 F test for comparing supervised classification learning algorithms.

Neural Computation, 11:1885–1892, 1999.

J. R. Beck and E. K. Schultz. The use of ROC curves in test performance evaluation. Arch Pathol

Lab Med, 110:13–20, 1986.

R. Bellazzi and B. Zupan. Intelligent data analysis in medicine and pharmacology: a position statement.

In IDAMAP Workshop Notes at the 13th European Conference on Artificial Intelligence,

ECAI-98, Brighton, UK, 1998.

Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation.

Journal of Machine Learning Research, 5:1089–1105, 2004.

C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. URL

http://www.ics.uci.edu/mlearn/MLRepository.html.

R. R. Bouckaert. Choosing between two learning algorithms based on calibrated tests. In T. Fawcett

and N. Mishra, editors, Machine Learning, Proceedings of the Twentieth International Conference

(ICML 2003), August 21-24, 2003, Washington, DC, USA. AAAI Press, 2003.

R. R. Bouckaert. Estimating replicability of classifier learning experiments. In C Brodley, editor,

Machine Learning, Proceedings of the Twenty-First International Conference (ICML 2004).

AAAI Press, 2004.

R. R. Bouckaert and E. Frank. Evaluating the replicability of significance tests for comparing

learning algorithms. In D. Honghua, R. Srikant, and C. Zhang, editors, Advances in Knowledge

Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May

26-28, 2004, Proceedings. Springer, 2004.

P. B. Brazdil and C. Soares. A comparison of ranking methods for classification algorithm selection.

In Proceedings of 11th European Conference on Machine Learning. Springer Verlag, 2000.