对计算机实验环节不规范描述的识别与编改*

张 枸[1)] 游中胜[2] 汤兴华[1)] 孙 凡[1)]
1)西南大学期刊社,400715;2)重庆师范大学编辑出版中心,400047:重庆

摘 要 分析计算机论文中实验环节常见的不规范描述问题及其原因,给出编辑识别与修改的方法和建议。

关键词 计算机论文;实验环境;实验样本;实验结果

Distinguishing and revising non-standard experiment descriptions in computer science papers // ZHANG Xun, YOU Zhongsheng, TANG Xinghua, Sun Fan

Abstract Non-standard experiment descriptions in computer science papers are discussed, and some suggestions are proposed.

Key words computer science paper; experimental environment; experimental sample; experimental result.

First-author's address Journal Press of Southwest University, 400715, Chongqing, China

实验方法和程序在科技论文中占据着重要地位,实验的结果能直观地反映出相关方案的优劣,是对论文所提出方案有效性的重要佐证[1-2]。迄今为止,编辑工作者对物理、农林、医学等领域论文中实验部分出现的问题进行了探讨[3-5],但对于计算机领域的相关问题却鲜有报道。与所有实验学科一样,计算机专业的科学研究往往也包含实验环节,对其过程的描述同样需要准确性和明晰性,目的是让读者能够重复实验,清楚地了解作者的研究工作,同时为进一步研究打下基础;然而,在已发表的计算机论文中,却常常忽略了实验环节的重要性,出现一些描述不规范的问题。

1 常见问题

1.1 实验环境的描述

1)描述不完整。就像生命科学、化学等学科的实验方法一样,算法的运行实验应该具有可重复性,因此对于实验环境的描述应该尽可能详细。计算机实验环境的描述非常重要,它关系到实验结果是否可靠、正确;但在目前的计算机论文中,关于硬件环境和软件环境的描述常常显得很随意,不准确、不具体,软、硬件环境包含的项目不明确。例如,有的文章只给出了硬件环境,有的文章只给出了软件环境,有的文章虽同时给出了软、硬件环境,但是描述却存在不规范之处。

例 1 [6] 以上结果均在配置为 Intel 双核 CPU, 2.40 GHz 主频,8 GB 内存,显卡为 NVDIA GeForce 9800 的机器上运行得到。

例 1 的实验环境仅仅包含了硬件环境,却忽略了软件环境,而其参考文献中却包含了软件环境的明确描述。例如在一篇题为《Flow and changes in appearance》的文献中,明确提到"The building was modeled using AutoCAD…"。

2)描述不准确。

例 2 [7] 机器配置为:Intel P4 2.6,1 GB,120 GB 硬盘。

例 3 [8] 实验平台选取为测试环境:Pentium® 4, 3.0G Processor,512 M 内存,Microsoft Window XP Professional, IDA Pro 5.0, Matlab7.1, Visual Studio.net 2003。

例 2 中未明确交代各参数的含义,只能猜测 2.6 指的是 CPU 频率, 1 GB 则是指内存容量。同样,例 3 中对于 CPU 的频率的描述是"3.0 G",显然缺少了频率单位 Hz,而且对于内存容量只写了"512 M",也缺少了字节单位"B"。

1.2 实验样本的描述

关于实验样本,主要有以下 2 类:1)由作者自己制造的样本数据(包括作者自己采集的样本数据或编程生成的样本);2)其他研究者或机构提出的数据。

对于第 1 类数据,即自制数据,如果文章中不将样本数据提供出来,则读者无从知晓真实的实验样本,也无法重复作者的实验过程。这种情况在涉及到计算机图像处理的文章中尤为常见。文章中往往给出几幅原始图像,但是对图像并不交代出处。

对于第 2 类数据,许多文章在引用此类数据时也比较混乱。如例 4 ~ 6 中引用的均是 UCI(University of California Irvine,加州大学尔湾分校)机器学习数据库中的数据集,但是却用了 3 种不同的形式对其进行说明。

例 4 [9] 为进一步测试算法 DSMAEC 的性能,我们从 UCI 数据集中选取 3 个数据集(breast cancer, iris, glass)对其进行测试。

例 5 [10] 正文 下面通过 UCI ML① 的 MON K1 数据集进一步说明。

脚注:①http://archive.ics.uci.Edu/ml/

*国家社会科学基金资助项目(10XTQ007);教育部人文社科基金资助项目(10YJA860011);中央高校基本科研业务费资金资助项目(XDJK2013C122);中国高校科技期刊研究会基金资助项目(GBJXC1270)

例 6 [11]　正文　真实数据集 2:Forest Covertype 数据集[18]。(为了与本文相区别,将例子中的文献序号用黑体表示——编者注。)

参考文献:

[**18**]　Newman D J, Hettich S, Blake C L, et al. UCI repository of machine learning databases. http://archive.ics.uci.edu/ml/

1.3　实验结果的描述　论文的实验结果往往用图表来表达。一张内容精彩翔实、形式简洁自明的图表或照片,一个经反复论证的公式或数学模型,其内容远远胜过文字表述[1]。这方面的主要问题如下。

1)图表的标注与图表的描述欠缺一致性。实验结果往往是用图和表来表示的,图表题和内容一般附有英文对照,以利于海外读者阅读[2]。这本无可厚非,但是某些期刊对图表的标注采用的是英文术语,而在正文部分的相应描述却用的是中文术语甚至不加描述,两者不一致。

图 1 为《异构无线传感器网络的转发连通覆盖方法》[12]一文中的活跃节点的覆盖度。对于此图,只能从此英文图题"Number of active sensors for covering area"与中文图题"覆盖集中的活跃节点数"得到纵坐标的中文含义是"活跃节点数",但是在正文中却没有其他任何关于横坐标"Deployed sensors number"的说明。

2)实验结果描述不准确。表 1 是文献[10]作者作了分类实验后得到的分类精度表。

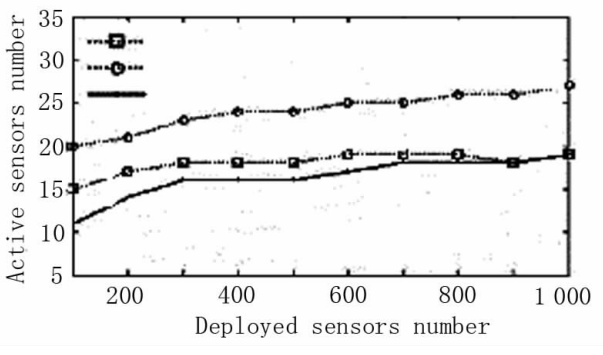

图 1　活跃节点的覆盖度

表 1　各数据集信息以及基于不同特征选择算法的 J4.8 的分类精度

Dataset	# of Features	# of Samples	FSI	IN T	IG	Relief F
Syndata	200	150	88.3	51.3	51.9	67.5
$Monk_1$	7	124	94.6	92.1	76.2	88.9
$Monk_2$	7	169	91.7	92.5	80.2	84.7
$Monk_3$	7	122	90.6	90.1	87.7	90.0
Wine	13	178	90.3	92.3	82.9	79.6
Zoo	17	101	89.3	90.6	91.6	91.0
Lung-cancer	56	32	74.2	68.3	60.7	59.3
Cmc	9	1 473	58.4	60.8	50.1	53.4
vehicle	18	846	78.7	73.1	62.6	65.5

联系上下文可知:表 1 中第 1 列的含义是实验数据集合名称,第 2 列是实验集合特征数,第 3 列是实验集合样本数,而第 4 ~ 7 列才是分类精度;但是此精度并没有标注单位。通过看正文中这样的描述"不难发现,FSI 在大多数据集上都具有较高的分类精度,其中不少已经接近或大于 90%",方才消除了歧义,确定此表中的分类精度是百分数。

2　编改方法和建议

2.1　实验环境描述的规范化　到底实验环境可不可以省略?若不能省略又如何描述?笔者认为:

1)软件环境不宜省略。软件环境包括操作系统平台、程序编译平台、建模工具等,是作实验的软件基础,只有把软件环境交代清楚,才能使重复实验成为可能。完善的软件环境应包括操作系统类型和版本,以及其他实验软件比如建模工具、编译平台等。

2)硬件环境能不能省略应区别看待。任何一段确定的代码,在不同计算机上运行时都能够返回相同的结果,而这一结果与运行速度无关;因此,对于某些只需要得到运行结果的实验,可以忽略硬件环境,如对于仅需要得到聚类精度的聚类实验。但是,如果实验统计数据中还包含了代码运行时间,那么最好给出几项重要的硬件参数,包括 CPU 类型和频率,内存类型、频率和容量,硬盘类型、转速和容量。若为图像处理实验,则最好给出显卡型号。

在描述硬件环境时应当力求做到严谨和准确。首先,对于计算机术语的描述,应力求严谨,例如中央处理器(central processing unit)可简称为 CPU,如果用 Processor 就不够严谨。其次对于参数的描述,应做到准确;对于 CPU 频率以及内存容量等,不要遗忘标注单位[13]。

综上所述,例 3 可以改为:"实验平台硬件环境为:Pentium®4,3.0 GHz CPU,512 MB 内存。实验平台软件环境为:Microsoft Window XP Professional,IDA Pro 5.0,Matlab 7.1,Visual Studio. net 2003。"

2.2　实验样本描述的规范化

1)对于作者自制的数据样本,建议提供链接以便读者获取。

2)若是引用其他研究者或机构提供的数据库,则最好能以参考文献的形式给出数据出处,如例 6 所示。

应特别注意,有的数据库发布者在数据引用方面有其特殊要求,此时应充分尊重发布者的意愿。UCI 机器学习数据库主页(http://archive.ics.uci.Edu/ml/)上专门有这样一句话:"For information about citing data sets in publications, please read our citation

policy."(若要在出版物中引用相关数据集,请参阅我们的引用办法。)对于打开"Citation Policy"链接,有这样的描述:"If you publish material based on databases obtained from this repository, then, in your acknowledgements, please note the assistance you received by using this repository. This will help others to obtain the same data sets and replicate your experiments. We suggest the following pseudo-APA reference format for referring to this repository: Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science."这段话明确规定了当出版物中使用了本数据库中的数据时,应在文中如何给予交代,而且这段话也充分说明了注明数据出处的3项重要意义:首先是对数据提供者的尊重与感谢;其次是帮助读者能方便地获取实验数据;最后是便于读者可以重复实验。

2.3 图表的规范化 中文科技期刊的服务的对象主要为国内读者,将图表的相关项目全部用英文标注而在正文部分关于图表的描述中对英文标注不加说明是非常不合适的;因此,在中文能表达清楚相关项目的意义时,宜尽可能使用中文,如果要对图表进行英文标注,也需在对图表的意义进行描述时给出说明。同时,该加单位的项目务必把单位补充完整,以免引起歧义。

3 结束语

实验环节对于计算机学科至关重要,对实验环节的准确描述对于论文价值的体现有着举足轻重的作用。本文指出了计算机论文实验环节中常见的不规范描述,并提出了相应的编改方法和建议,为计算机学科论文表述的进一步规范化和标准化提供参考。

4 参考文献

[1] 王晓琪. 科技论文的表现规范与编撰技巧[J]. 编辑学报,1995,7(1):40-43

[2] 熊家国,陈红叶,张志枉. 论实验研究型论文信息结构的编辑控制原则[J]. 编辑学报,2003,15(4):245-246

[3] 王书亚,谭颖波,王云亭. 医学科研论文中实验设计和统计分析内容的正确表述[J]. 中国科技期刊研究,2005,16(1):116-118

[4] 左小青,林琳. 科技论文有关热分析实验中物理量的规范化表示[J]. 中国科技期刊研究,2007,18(3):517-519

[5] 程红,李莉. 学术期刊正交试验类稿件的审读方法[J]. 编辑学报,2012,24(5):450-452

[6] 焦少慧,杨刚,HENG Pengán,等. 基于时变纹元的真实感草地枯萎模拟[J]. 软件学报,2010,21(9):2224-2236

[7] 孙扬, 赵翔, 唐九阳,等. 一种多变元网络可视化方法[J]. 软件学报, 2010,21(9):2250-2261

[8] 孔德光,谭小彬,奚宏生,等. 提升多维特征检测迷惑恶意代码[J]. 软件学报, 2011,22(3):522-533

[9] 潘晓英,刘芳,焦李成. 密度敏感的多智能体进化聚类算法[J]. 软件学报, 2010,21(10):2420-2431

[10] 王博,黄九鸣,贾焰,等. 适用于多种监督模型的特征选择方法研究[J]. 计算机研究与发展,2010, 47(9):1548-1557

[11] 张晨,金澈清,周傲英. 一种不确定数据流聚类算法[J]. 软件学报,2010,21(9):2173-2182

[12] 温俊,蒋杰,方力,等. 异构无线传感器网络的转发连通覆盖方法[J]. 软件学报,2010,21(9):2304-2319

[13] 关旭. 几个易混通信速率的量和单位辨析[J]. 编辑学报,2011,23(5):403-404

(2013-01-10 收稿;2013-03-25 修回)

注意纠正百分率表达式错误

在科技期刊审读中发现,相当数量科技期刊中的百分率表达式存在错误,应注意纠正。为了方便并排除专业学科的影响,我们选用具有通用性的出勤率 η 计算的例子。其中出勤人数为 n,应到人数为 N。

第1种错误表达式是

$$\eta(\%) = \frac{n}{N} \times 100\%, \quad (1)$$

第2种错误表达式是

$$\eta(\%) = \frac{n}{N} \times 100, \quad (2)$$

第3种错误表达式是

$$\eta = \frac{n}{N} \times 100。 \quad (3)$$

其中:式(1)等号左侧的"(%)"是多余的;式(2)不仅等号左侧的"(%)"多余,而且等式不成立;出勤率可以用小数来表示,但式(3)将等号右侧的数值扩大了100倍,错误是显然的。

出勤率 η 正确的表达式为

$$\eta = \frac{n}{N} \times 100\%, \quad (4)$$

或者为

$$\eta/\% = \frac{n}{N} \times 100, \quad (5)$$

或者为

$$\eta = \frac{n}{N}。 \quad (6)$$

式(4)和式(5)算得的"率"为百分数,式(6)算得的"率"为小数。

(同 任)