关于发展学术品牌的思考

★ 模式识别国家重点实验室 · 胡包钢

最近 NLPR 实验室重新组建模式识别基础研究团队。我作为成员之一，感慨与思考颇多。愿分享自己的一些想法，抛砖引玉，以期促进实验室更好地发展。

首先涉及到实验室内基础研究与应用研究之间的关系。以本人开展的研究对象植物为比喻，我理解实验室犹如一棵大树，基础研究反映了大树在地下的根系部分，应用研究可以理解为是地上的枝叶部分。真可谓“根深才能叶茂”！实验室能够发展多高和多好，最后还要追溯到它的根基部分。回首实验室成长历程，其进展是相当显著、有目共睹的。我接触到的许多国外知名学者都对 NLPR 的整体研究实力表示欣赏，特别欣喜的是实验室年轻老师们发表在国际顶级刊物或会议上的论文，其水平和十年前相比已不可同日而语。但是在国家重点实验室的五年一评估中，除成绩与特色外，专家组仍然指出我们在模式识别领域内基础研究方面工作不足。作为模式识别理论小组成员，我强烈感受到压力和责任，心有不甘但又十分认同专家们的意见。我认为问题产生的根本原因还是在于缺少基础研究方面的相关文化和学术积累。谈到文化，其中涉及两个基本疑问：

什么是模式识别领域中的基础研究？

如何在模式识别领域中开展有品牌的基础研究工作？

借鉴《国家“十一五”基础研究发展规划》中对“基础研究”的定义[1]，第一个问题大体可以回答为“是针对模式识别为应用背景开展的获取新知识、新原理、新方法的研究活动”。由于模式识别本身属于应用基础研究类别，有时对具体研究归属的理解更易产生差异。如我们在国际顶级刊物或会议文章发表的方法性文章可以理解为是应用基础研究，但是为什么专家们还不太认同我们在基础研究方面的工作投入。对此可以有多种解答。我个人理解我们研究中首先十分缺少对模式识别研究中新概念、新知识、新原理方面的工作。在新方法研究方面我们的工作更为突出，因为模式识别更为靠近算法层面，表明方法研究是其中十分重要的内容。但是，我们在方法“新”的层面是否足够显著，我们在追求共性的模式识别方法研究方面是否更为重视，都值得思考。我本人更偏向于要广义的理解模式识别领域中的基础研究工作，关键是要强调“原始创新”，能够成为后来工作的“源头”或“里程碑”。从发展历史看，模式识别方面研究工作中可能不必要“严格”划分基础和非基础类别。非基础工作也可能会发展出原创而基础性的新知识，新方法。前提是必须要成为创新的有心人。针对专家意见，实验室不断强化基础研究团队的举措是对的。一个科研部门要落实基础研究文化并形成若干核心人员。它能够无形中整体提高该部门的研究水准，同时使应用研究可以不间断地获得“源头活水”，确保自身发展更强、更壮。较为重要还有在向专家汇报时，我们有必要讲清哪些是基础研究工作及其学术创新。事实上，实验室在计算机视觉方面有着很好的基础理论研究积累和成果。在国际上开创脑网络组学研究方面也是独树一帜。关键是我们要讨论清楚，有针对性地改变发展中的不足。如加强模式识别中新概念、新知识方面的基础研究。

对于第二个问题，所谓“品牌”的说法，对任何人，研究组，实验室都适用。这并不需要区分你是开展基础研究还是应用研究。中国科技发展开始步入“质重于量”的阶段。“品牌”的考察应该有特定范围并是多元的。如果是基础研究，特定主题下的国际范围比较应该是必须的。研究所已经开始强调“代表性成果”考核方面的规划。这是管理上的进步。作为个人或研究团队，不能不思考我们的代表性成果是什么，能否形成品牌？

如果未曾思考，或讲不清楚，如何能够得到同行认可？下一个关键问题是如何定义和判断“品牌”工作。说来这也是个模式识别中的典型问题。除同行定性评议之外，还可以采用让数据说话的方式来综合判断。如借用学术界经常应用的“里程碑(Stone)”术语。所谓“里程碑式”的工作是指在相关主题研究中，有关工作（表现形式可以是文章、专利、软件、数据库等）是不能不提到的。源头的“里程碑”和后来发展中的“里程碑”还是有显著差异的。对于多数为发展中的“里程碑”工作，识别的方式可以是应用主题关键词在 GoogleScholar 一类搜索工具中查询，有关工作应该进入前五名。该考察方式优点是克服应用绝对引用数量无法在冷门与热门主题之间实现相对比较的局限性。它会更鼓励研究工作要追求影响力以至“源头式”研究，而不是热门课题。缺点是考察结论受到关键词的选取而会发生变化。但是，该结论的有效性在同行评议中是可以辨别出来的。建议其中初始的关键词是由被评审者来提供。希望能够看到更多对学术“品牌”多元化评估的讨论。当该方面定义清楚后，全方位思考“如何发展”也是十分重要的。

下面我会谈谈我对有关品牌发展方面的理解和体会。

首先无论是团队还是个人需要明确设定国际品牌发展目标。这种品牌设定方式是始于我从事植物生长建模研究中的经历。该领域通常是以模型方法命名区分流派研究的。经过十多年共同努力与不断创新，目前中法两国研究人员共创的“GreenLab（中文名称为青园）模型”方法已经成为植物生长建模领域研究中的国际品牌之一，并为所有参与研究者分享。在品牌发展过程中，每个人的具体学术贡献也是明确的。建立共赢和鼓励个人冒尖的运行机制也是团队间长久合作发展并共建品牌的重要保证之一。回顾该研究历程使我更加认同中国古代学者《礼记 · 中庸》之言：“凡事预则立，不预则废”。

其次是基础研究要有顶层设计。目的是形成“兼有特色和系统性”的研究成果。学术品牌讲究“特色”，强调的是研究工作要“与众不同”。“系统性”表明不要用“打一枪换一个地方”的方式来从事研究，好的研究成果必然是建立在一定时间积累上的。顶层设计中最要思考的问题是“So what(那又怎么样)?”。或者自问：我的预期研究成果是否重要，能否得到同行认可，如何走的更远？当我进入机器学习研究工作后，逐步条理出顶层设计规划：围绕一个基本科学问题：“学什么”（或学习目标选择），重点在两个基本研究主线：“增加数据驱动模型的透明度与基于信息理论的机器学习”。我在国内外学术交流时会谈到规划方面内容。这不仅是介绍我们研究特色，更是考察我能否讲清楚预期“新概念”的“原创性”（如模型透明度研究中提出带广义约束的建模方法）。另一方面如何通过长远发展规划来摆正研究工作心态，避免“短平快”与减少“外界干扰”，对于基础研究是十分重要的。

再一点是要具备学术自信。只要融入世界潮流，中国人是有很强竞争力的。随着研究的推进，这几年我在研究组内提出的品牌目标是“共同发展可以进入模式识别教科书中的新知识”。我也试图引导同学们要明白世界上知名学者的多数学术贡献和地位起源于博士期间的工作。借鉴韩愈《师说》中观点“弟子不必不如师”，我理解博士论文更应表现为“学有专攻，弟子理应超过师”。如 1996 年在 ACM SIGMOD 上发表的一篇称为 BIRCH 聚类方法的文章 [2]，到目前为止 GoogleSholar 引用数为 3514 次。而该文第一作者是大陆留学美国的女博士生 ZHANG Tian [2]。该方法特别适合于大数据处理，因此从学术到应用上的重大影响力是可想而知的。文章的内涵更多是思想与算法层面，其中仅有九个计算公式。又如我们课题组张晓晚同学，她在最近被 IEEE TKDE 接收发表的一篇论文中 [3]，首次对 ROC 曲线中给出拒识区间的几何解释。我理解这也是一种“新知识”型品牌，并会成为未来模式识别教科书中的内容。上述事例表明不要将基础研究工作与建立品牌看得过难，关键是要有志向和行动。我本身数学基础并不好，但在

理解了理论研究创造知识并具有长远影响力之后也开始了在模式识别基础理论相关方面的工作。去年在 arXiv 上，我作为第一作者合作发表了一篇关于二值分类贝叶斯、非贝叶斯误差与条件熵上下界显式表达关系方面的文章[4]。在信息理论被称为“Bible(圣经)”的教材书中[5]，有关 Fano 下界是第二章的内容，可见其基础性。不同于应用一维边缘概率分布中不等式方式推导的特定解，我们以二维联合概率分布中优化方式导出的显式解更具一般性。非贝叶斯误差(多数分类器为此类别)也被首次引入信息论的界分析中。这项工作的缺点是局限于二值分类问题。但是“安心定志”与“持之以恒”可以帮助研究工作更上一层楼。

上面讨论反映观念可能是第一位的。这并不排除讨论改变科研体制上现有问题的重要性和紧迫性。对此我总是想，前人面临的条件只能比我们差，他们为什么还能做出高水准的研究工作？关键是要追求真正有价值的东西，如乔布斯一样用品牌去影响世界。对于应用研究，我认同“微创新”的说法，对于基础研究，追求原始创新将至关重要。针对模式识别基础研究，我们是否应该防止掉入“增量改进已有方法式研究(Incremental improvements on well-established methods)”范式的陷阱[6]。当应用数值仿真计算验证新方法有效性已经成为模式识别领域发表文章的一般套路时，我们应该更多的自问一下“新方法是否带来了新概念或新知识”？“它为什么优于现有方法，优于的条件是什么，可否理论证明”？只有好的学术价值观方能产生出有真正价值的研究成果。

近几年来大数据主题与深度学习方法兴起，它们既对模式识别研究提供了强有力的驱动力，又提出了巨大挑战。实验室在其中能够扮演什么样的角色，未来几年后可能的学术品牌会是什么？ 是应用系统方面还是理论方法方面？怎样在大数据与深度学习方面能够发展出更有原创性思想的研究工作？相信这些思考和讨论对集体和个人发展都会有帮助。这里我们又可以将实验室喻为一艘大船，只有齐心协力才能更快到达彼岸。

此篇杂谈的结尾是个人建议。前段时间，看到电视节目“中国好歌曲”，特别喜欢和感慨这样的平台，它既鼓励原创歌曲，又能推出真正有才华的新人。为此我建议实验室能否在每个年度开展一次基础研究品牌评选工作，鼓励非基础研究团队的个人或课题组递交半页纸描述的作品参加网上初评（基础研究小组人员参加否可以讨论）。之后由基础研究小组负责安排全实验室范围内的作品答辩。包括学生参与投票，最后以得票前三名作品为获奖者。提议由实验室颁发证书和额外的科研经费奖励（非个人奖金）。此活动目的是为有志于开展基础研究的人打开大门。更为重要的是在实验室内建立起良好的基础研究方面的创新文化，让我们共同追求学术品牌，为实现更高国际声誉的 NLPR 实验室而自豪。

（初稿 2014 年 4 月 14 日，
终稿 2014 年 5 月 5 日）

参考文献

[1] 中国科学技术部，《国家“十一五”基础研究发展规划》，2006, 10, 30. http://www.most.gov.cn/kjgh/kjfzgh/200708/t20070824_52690.htm

[2] Zhang, T., Ramakrishnan, R. and Livny, M., “BIRCH: an effcient data clustering method for very large databases”, ACM SIGMOD, 1996

[3] Zhang, X.-W. and Hu, B.-G., “A New Strategy of Cost-Free Learning in the Class Imbalance Problem”, preprint on IEEE TKDE, 2014.

[4] Hu, B.-G. and Xing, H.-J., “A New Approach of Deriving Bounds between Entropy and Error from Joint Distribution: Case Study for Binary Classifcations”, preprint on arXiv:1303.0943, 2013.

[5] Cover, T.M. and Thomas, J.A., “Elements of Information Theor”, (2nd eds.), John Wiley & Sons, 2012.

[6] LeCun, Y. “A New Publishing Model in Computer Science”.

http://yann.lecun.com/ex/pamphlets/publishing-models.html