数据科学研究型人才培养的思考与实践

石 川[1]，王 啸[1]，杨 成[1]，李清勇[2]，吴 斌[1]

（1. 北京邮电大学 计算机学院，北京 100876；

2.北京交通大学 计算机学院，北京 100044）

摘 要：分析数据科学研究型人才的培养定位与层次，结合具体实践过程提出数据科学人才的能力要求和培养方案，最后通过典型案例介绍对不同学生的具体培养过程，说明取得的效果。

关键词：数据科学；大数据专业；研究型人才

0 引 言

随着大数据时代的到来，企业与社会亟需数据科学人才。2015年《国务院关于印发促进大数据发展行动纲要的通知》里指出“信息技术与经济社会的交汇融合引发了数据迅猛增长，数据已成为国家基础性战略资源，大数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响”。同时指出要加强专业人才培养，要“创新人才培养模式，建立健全多层次、多类型的大数据人才培养体系。鼓励高校设立数据科学和数据工程相关专业，重点培养专业化数据工程师等大数据专业人才。大力培养具有统计分析、计算机技术、经济管理等多学科知识的跨界复合型人才”。教育部于2016年在我国高等学校设置数据科学和大数据技术本科专业，随后国内有700多所高校相继设立了该专业。首批毕业生开始进入企业或者攻读研究生，这对缓解大数据人才缺口具有重要意义，也为培养高水平数据科学科研人才和技术专家奠定基础。如何培养高水平的数据科学人才是教学科研和人才培养工作重点关注的内容。

1 数据科学研究型人才的培养定位与层次

高水平数据科学人才简单来说就是：顶天立地。顶天：能够深入研究数据科学核心问题，发表数据挖掘、人工智能等领域的一流会议论文，例如CCF A/B类会议；立地：能够解采用创新性方法决企业实际业务难题，提升生产效率，并总结经验撰写论文或申请专利。

高水平数据科学人才培养主要面向“211”级别高校的学生，他们具有一定的动手能力和较强的探索欲，这类学生是大学生的主体。本科生主要掌握计算机的基础知识，具备数据科学的基本能力。经过本科阶段的培养，硕士和博士研究生已经具备基本素质，但是要经过严格、系统的培养才能成为高水平数据科学人才，不同研究领域的学生应具备的能力要求和培养过程也存在较大区别。

总体来讲，数据科学研究型人才的培养层次可以分为以下3方面：技术人才，应用人才，跨学科人才。①对于技术人才，重在培养学生在数据科学领域的科学思维与科研技能，基于一套科学合理的科研逻辑，建立一套提出问题、分析问题、解决问题及科研写作的研究体系。②对于应用人才，重在培养学生在数据科学实际应用中解决问题的技能，培养学生结合实际应用，理解数据，理解业务，解决实际场景中的问题。③对于跨学科人才，则是面向多学科、多领域融合目标对学生进行培养，学生不仅需要计算机技能，还要结合其他领域的知识，解决不同领域与学科中数据分析处理的问题。

2 数据科学人才的能力要求与培养方案

文章编号：　　中图分类号：G642

根据数据科学家的能力要求[1]，高水平数据科学人才要具有综合能力。总体能力要求如图1所示。

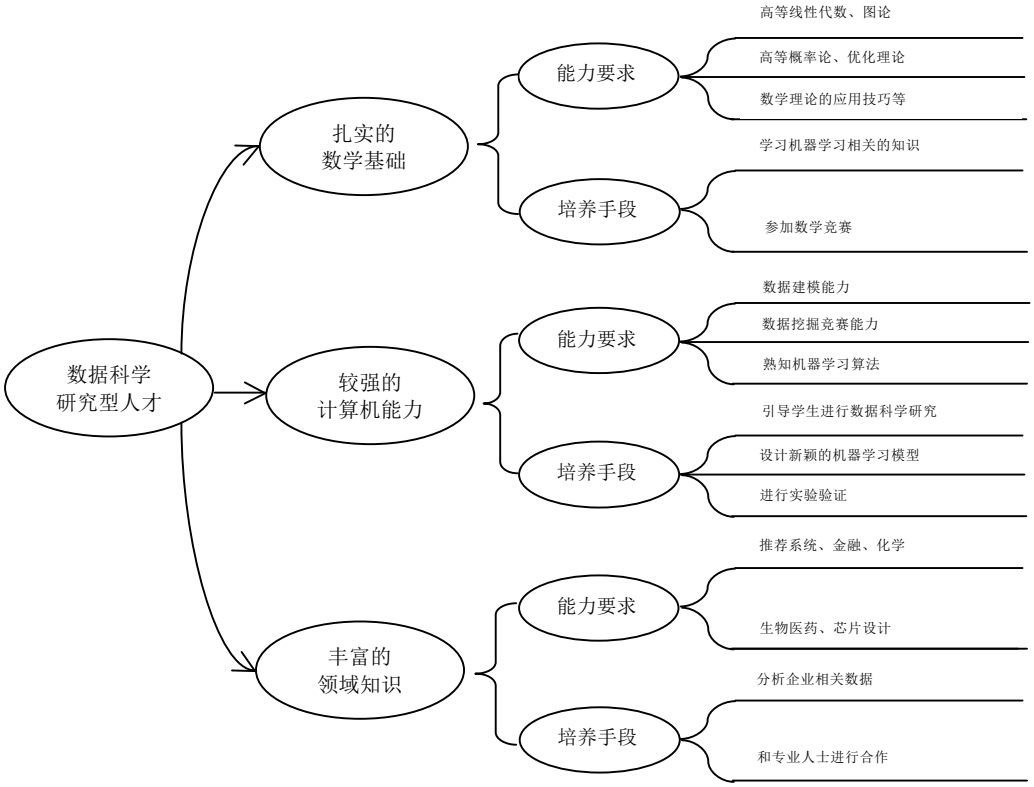

图 1 数据科学型人才能力要求与培养手段

1）扎实的数学基础。

掌握基本的数学方法和工具；具有较强的数学建模能力，能够将实际问题建模为数学问题。数据科学领域紧密相关的数学知识包括线性代数、概率论、优化理论和图论。虽然学生在大学阶段都学习过这些数学知识，但是尚缺乏在实践中运用与进一步理解。对实际问题的建模与设计算法都需要这些数学知识，因此非常重要。研究生阶段没有完整的时间重新学习这些知识，可以通过学习机器学习和人工智能相关的知识，真正理解与运用这些数学基础知识。也可以通过鼓励学生参加数学建模竞赛，培养学生利用数学解决实际问题的能力。

2）较强的计算机能力。

能够熟练编程，熟悉常用算法，能用算法解决实际问题。从事数据科学的研究与开发需要较强的计算机能力。大学计算机类的本科生已经掌握基础的编程能力和基本的计算机应用能力，但是这些能力只是用确定的算法解决简单的问题，还不具备解决实际问题的能力。要培养学生利用人工智能算法（主要是机器学习）创造性解决复杂真实问题的能力。对于这方面能力的培养，主要是通过引导学生进行数据科学方面的研究达到：将真实的有价值的数据科学问题建模为机器学习（数学）问题；设计新颖的机器学习模型创造性解决该问题；进行实验验证并撰写论文或申请专利。也可以通过鼓励学生参加数据挖掘竞赛，锻炼学生的动手能力。

3）丰富的领域知识。

掌握丰富的领域知识，能够将领域知识转化为计算机语言，将领域知识融入到算法设计中。领域知识是和要解决的问题紧密相关的，虽然我们无法提前掌握所需要的全部领域知识，

但是可以通过快速学习并掌握解决问题的基本方法来快速理解一个新领域。通过分析数据可以快速掌握数据的特点和行业基本知识；与专业人士合作也是快速掌握领域知识的有效方法。

4）沟通交流和总结表达能力。

在研究过程中和学生充分沟通交流，共同确定研究问题，探索解决方案。要求学生将研究工作撰写成学术论文，并投寄到领域顶级会议或期刊，积极参与宣传工作。这些活动可以培养学生的沟通交流和总结表达能力。

3　培养过程

培养过程的总体流程如图2所示。

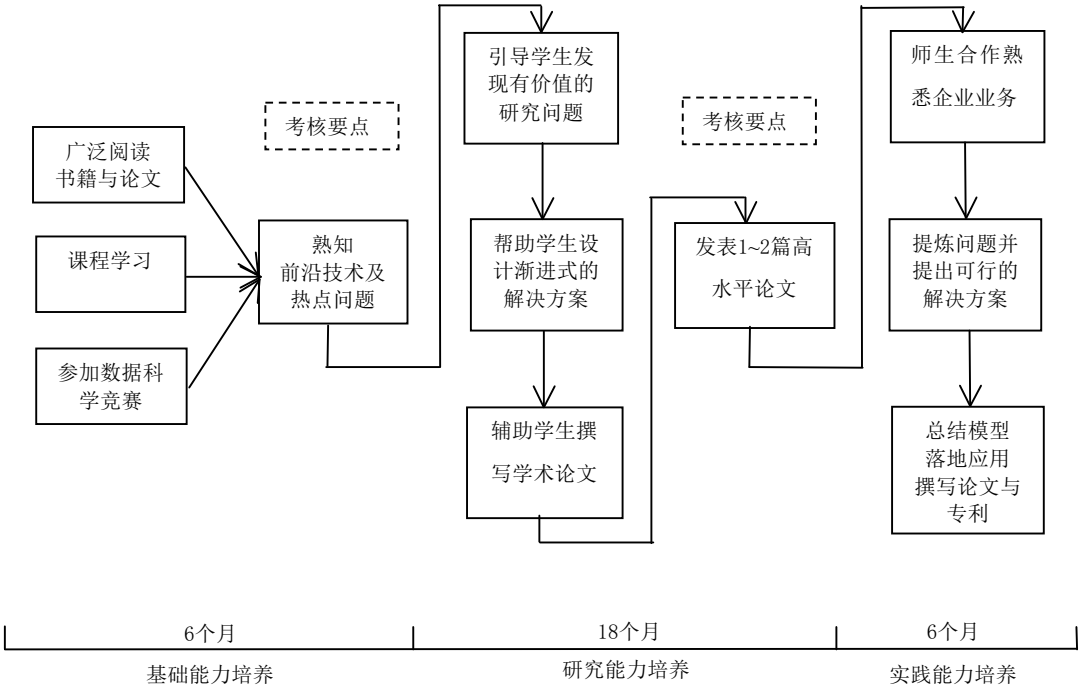

图 2 数据科学型人才培养过程的总体流程

1）基础能力准备。

对刚上研究生的研一学生而言（也可以是保研的大四学生），需要半年左右的时间培养数据科学的基础能力。主要的方式是学习数据挖掘和机器学习的基础知识，并学会灵活应用。一些难度适中的书籍和课程包括但不限于：李航的《统计学习技术》[2]和Andrew Ng的机器学习课程[3]。要求学生对常用的机器学习算法能够真正理解，能够推导公式，并能够根据实际问题作出改进；与此同时，要求学生阅读数据挖掘和机器学习领域的最新论文，了解前沿技术和热点问题，为科研做好准备。鼓励学生参加合适的数据挖掘竞赛，通过竞赛灵活掌握、运用机器学习方法，真正懂得如何用技术解决实际问题。

2）研究能力培养。

到了研一下学期，就要开始进入数据科学领域的研究工作，需要一年半左右的时间培养学生研究能力，完整进行1~2项研究，发表1~2篇高水平学术论文或申请专利。

（1）确定问题。经过近半年的论文阅读，学生对研究的领域已经有所了解，一些学生也开始有些研究想法。对于有研究想法的学生，教师可以引导其深入思考，自主发现有价值的研究

问题[4]；对于没有研究想法的学生，教师可以让其继续阅读思考，也可以指定有望做出成果的研究问题。从事数据科学研究，问题是最关键的，教师要引导学生发现有价值的理论或实际问题。一方面，可以通过大量阅读最新的前沿论文，发现发展趋势，找到研究问题；另一方面，可以通过工业界的实际反馈，找到技术难点。找到有价值的真问题是成功研究的基础，但也不能花费太多时间，以免学生产生急躁情绪或自我怀疑。时间大概1~3个月为宜，在这个过程中，教师要全程参与，和学生充分沟通，并且真正懂得前沿，能够发现问题。

（2）解决问题。确定好问题后，要找到创新性的解决方法，可以设计渐进式的解决方案。先考虑基本方法是怎样的，存在什么不足，如何改进。针对不同问题，采用合适的解决方法。最好采用最新技术，并且根据问题特点，进行针对性改进。这一阶段对学生能力要求较高，产生的结果差异也很大。有些学生可能1~2个月就能完成，有些学生可能要用4~5个月，甚至不能完成。教师要帮助学生设计模型，找到卡壳问题，提出解决思路；也要鼓励学生多向其他同学学习，多向专业人士请教。

（3）论文撰写。做好工作后，要撰写成学术论文，尽力投寄到好的期刊和会议上发表。学生要根据论文要求，定义好问题，写明白方法，并完善实验。这一阶段对学生能力提出了很高要求，花费的时间差异较大，写作能力较差的学生，培养更为困难，培养周期也更长[5]。一般学生要写3~4篇论文，经过2~3年的训练才能比较独立地写作。在这一阶段，教师的作用很关键，要判断学生工作的水平，选择合适的期刊和会议。教师可根据计算机学科的特点，鼓励学生投寄高水平的国际会议；但是不可盲目选择难度太大的会议，因为屡投不中会损害学生的积极性。刚开始做研究的学生，写作水平普遍达不到高水平论文的要求，教师往往要花费很大的精力指导学生写作，甚至亲自重写论文。从培养学生的角度，应该让学生花更多的时间自己撰写与修改论文。

3）实践能力检验。

到了研究生二年级下学期，可以派驻学生参与企业实习，在实际工作中培养实战能力。经过上述培养，学生能够完成1~2项研究，发表了高水平论文（CCF B类以上，即中国计算机学会推荐的B类期刊和会议），并且具有较强的动手能力和求知欲，就可以派驻到企业实习。要求学生用3~6个月解决企业实际难题，并能够总结工作成果发表论文或申请专利。实习的企业要有业务难题和真实的数据，不是只布置简单的工程开发任务。

①学生到企业后，先用1~2个月的时间熟悉企业业务和数据，并提炼出要解决的问题。这个过程需要学生、企业和教师紧密合作并相互配合。学生要快速理解企业业务，分析真实数据，找到痛点问题[6]；企业要充分配合，积极合作；教师要帮助学生提炼问题，提出解决思路。②然后，用1~3个月的时间设计模型并验证。学生要设计并实现可行的解决方案；企业要判断方案的合理性和有用性，并提供相应的支持；教师也要对方案提出建议和指导。③最后，用1~2个月总结模型，落地应用，撰写论文或申请专利。

4　案例分析及培养效果

在过去的5年时间里，实验室每年都有2~3名学生按照这种培养模式，成为了优秀的数据科学人才。在这期间培养的10余名学生（主要是硕士生），发表数据科学领域的高水平论文（CCF A/B类）40余篇；参与了阿里AIR、腾讯犀牛鸟、美团北斗等企业合作项目，发表CCF A/B类应用论文20余篇，申请专利20余项。这些学生大多进入头部互联网企业，成为算法研究员或算法工程师。

北京邮电大学2016级免试研究生胡同学，从本科第4年开始，先跟随高年级学长学习机器学习知识，并开始学习做研究；在研究生阶段，在教师带领下从事异质网络表示学习方面的研

究，先后在TKDE2018[7]和KDD2019[8]上发表论文；在研二暑假期间，被派遣到蚂蚁集团实习，历时3个多月，率先采用异质图神经网络解决互联网套现用户检测问题，在AAAI2019上发表论文[9]并申请专利。该技术在蚂蚁集团内部累计落地10余个业务场景，其中包括智能客服（在标签推荐场景CTR提升12.8%）、可疑账户识别（准确率提升11.4%，覆盖率提升21.4%）、支付（IOT广告场景累计CTR提升2.81%）等。该学生后来入职蚂蚁集团，工作表现优秀，两年内从P5晋升到P7。

南京邮电大学2017级免试研究生陆同学，在研究生阶段从事动态图表示学习方面的工作，先后在AAAI19[10]等会议上发表论文；在研二暑假期间，被派遣到腾讯公司实习，历时4个多月，采用图神经网络解决了社交关系对新闻推荐的影响，发表论文并申请专利；后到新加坡管理大学实习8个月，先后在KDD2020[11]、AAAI21[12]等会议上发表论文；获得2020年腾讯犀牛鸟精英人才优秀奖（排第二），并入职腾讯担任研究员。

北京化工大学2019级研究生王同学，在研究生阶段从事网络信息传播方面的工作，在TNNLS[13]等期刊上发表论文；在研二暑假期间，被派遣到高德实习，采用对比学习方法进行知识图谱推荐。

5 结 语

数据科学是国家面向科技、经济和社会发展的重大需求，属于“十四五”国家战略性新兴产业发展规划中“新一代信息技术”的重要组成部分。培养该领域的研究型人才对于国家发展战略有着至关重要的作用和意义。培养数据科学高水平人才应注意数学、计算机科学和领域知识的多学科交融，注重基础理论、学术研究与企业实践有机结合。未来，除了进一步探索数据科学高水平人才的培养规律外，还将探索数据科学领军人才的培养规律以及人工智能等相关方向人才的培养规律。

参考文献：

[1] 陈振冲, 贺田田. 数据科学人才的需求与培养[J]. 大数据, 2016, 2(5): 95-106.

[2] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[3] Coursera. Stanford, machine learning[EB/OL]. (2013-03-22)[2013-09-08]. https://www.coursera.org/learn/machine-learning.

[4] 石兵, 熊盛武, 饶文碧, 等. 数据科学与大数据技术专业建设研究与实践[J]. 计算机教育, 2021(4): 88-92.

[5] 谭红叶, 李茹, 吕国英. 数据科学与工程特色的计算机科学与技术人才培养模式构建[J]. 计算机教育, 2018(2): 14-17.

[6] 陆枫. 面向大数据时代的计算机系统能力培养改革与实践[J]. 计算机教育, 2017(3): 33-36.

[7] Shi C, Hu B, Zhao W X, et al. Heterogeneous information network embedding for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(2): 357-370.

[8] Hu B, Fang Y, Shi C. Adversarial learning on heterogeneous information networks[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery&Data Mining. NewYork: Association for Computing Machinery, 2019:120-129.

[9] Hu B, Zhang Z, Shi C, et al. Cash-out user detection based on attributed heterogeneous information network with a hierarchical attention mechanism[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii: AAAI Press, 2019: 946-953.

[10] Lu Y, Shi C, Hu L, et al. Relation structure-aware heterogeneous information network embedding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii: AAAI Press, 2019: 4456-4463.

文章编号：　　中图分类号：G642

[11] Lu Y, Fang Y, Shi C. Meta-learning on heterogeneous information networks for cold-start recommendation[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery&Data Mining. NewYork: Association for Computing Machinery, 2020: 1563-1573.

[12] Lu Y, Jiang X, Fang Y, et al. Learning to pre-train graph neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. California: AAAI Press, 2021: 4276-4284.

[13] Yang C, Wang H, Tang J, et al. Full-Scale information diffusion prediction with reinforced recurrent networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 1(9): 1-13.

基金项目：国家自然科学基金委员会，企业联合基金重点项目，“网络空间高隐蔽未知威胁智能检测与溯源研究”（U20B2045）。

第一作者简介：石川，男，教授，研究方向为数据挖掘，shichuan@bupt.edu.cn。

（编辑：赵 原）