【Café 速递】彭德华：聚类算法研究与科研经历分享

核心提示： 聚类算法可以应用在哪些场景？现有聚类算法的性能受到哪些因素的制约？对此又有哪些创新性的解决方案？科技论文的写作和投稿过程中，我们可能会遇到哪些障碍？彭德华同学结合自身科研经历，分享了他的心得体会。

主持：孟凡皓 录屏：沈婷 文字：沈婷

>>>人物名片

彭德华，武汉大学测绘遥感信息工程国家重点实验室 2020 级博士生，师从吴华意教授、桂志鹏副教授，主要研究方向为机器学习、聚类算法与理论。在 **Nature Communications, Future Generation Computer Systems, Neurocomputing** 等学术期刊发表 SCI/EI 论文 7 篇，申报发明专利 6 项，软件著作权 2 项。曾获得学业奖学金一等奖，中国研究生数学建模竞赛一等奖。

>>>嘉宾小语

- ✧ 从多尺度聚类到局部方向中心性聚类，再到流形学习、边界检测和图聚类，研究的过程中总会不断冒出新的问题指引我们的方向，蓦然回首才发现，已经走了很远的路了。

>>>报告现场

2022 年 11 月 6 日晚上 7 点，武汉大学测绘遥感信息工程国家重点实验室 2020 级博士生彭德华做客 GeoScience Café 第 344 期暨“研途指南”系列第二期讲座。彭德华根据自身科研经历，介绍了聚类分析中存在的问题和相关的创新性工作，分享了他投稿 **Nature Communications** 的科研历程与心得体会。讲座分为算法原理简介和科研经历分享两个部分。

图 1　彭德华作精彩线上报告

聚类算法创新性研究

大四保研后，彭德华在桂志鹏老师的带领下，开始了对多尺度聚类的研究。他所在的团队提出了一种融合分析尺度和视觉尺度的多尺度网格聚类（MSGC）算法[1]，并在中国大陆企业注册数据的实验中验证了算法的优越性。

在进行该项研究的过程中，彭德华系统总结了现有聚类算法在不同评价维度的性能。他认识到，现实数据中广泛存在的密度异质性和弱连接性是限制算法性能的重要原因之一，由此开启了他的第二项研究工作。

	基于划分	基于密度	基于层次	基于模型	基于网格	基于方向	基于图
任意形状簇	★	★★★	★★	★★★	★★★	★★★	★★
噪声识别能力	★	★★★	★	★★	★★★	★★	★
计算效率	★★★	★★	★★	★	★★★	★★	★
参数鲁棒性	★	★★	★	★★	★★	★★★	★
维度扩展性	★★★	★★	★★	★★★	★	★★	★★★
尺度适应性	★	★★	★★	★★	★★	★★	★
弱连接性	★★	★	★	★	★	★★	★★
密度异质性	★★	★	★	★★	★	★	★★
代表算法	K-means	DBSCAN	AGNES	GMM	STING	MeanShift	Ncut

图 2　现有聚类算法在不同评价维度的性能

为了有效应对聚类中面临的密度异质与弱连接数据分布，彭德华希望建立一种密度无关的边界识别度量指标来区分内部点和边界点，并使用边界点形成封闭的“笼子”来分离弱连接簇。局部方向中心性聚类算法（CDC）[2]是他的原创性成果，该算法通过度量每个点 KNN 方向分布的均匀性来搜索边界点，

度量指标在二维空间定义为角方差，在高维空间则转换为对单形体积的计算。

该项工作于今年 9 月被发表在 **Nature Communications** 期刊上，论文展示了 CDC 在单细胞 RNA 序列（scRNA-seq）、质谱流式细胞（CyTOF）、英语语音数据库（ELSDSR）等 47 个不同类型的数据集上与 38 种专业或通用基准算法的对比结果，显示出了 CDC 的巨大潜力。

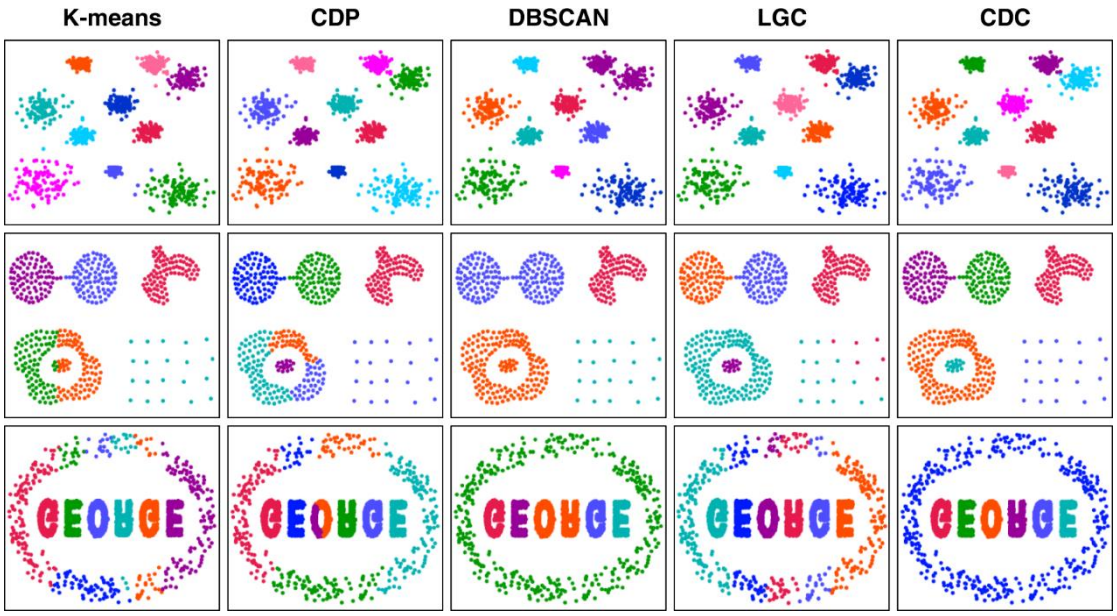

图 3　CDC 与其他经典聚类算法在合成数据集上的聚类效果对比

此外，彭德华还分享了他在流形学习、边界检测、图聚类等方向的研究工作。

科研经历分享

第二部分，彭德华从投稿经历、论文作图、回复审稿意见三个方面分享了自己的科研经历。

首先，彭德华分享了论文 *Clustering by measuring local direction centrality for data with heterogeneous density and weak connectivity* 的投稿经历。这篇文章被审稿人誉为“一项优雅、简洁且极具创新性的工作”，而彭德华则向我们揭示了论文发表背后的凄风苦雨。在初次被 **Nature Communications** 拒稿后，他根据修改意见补充了算法伸缩性实验并设计了高维可扩展方法；之后，论文中对生物数据集的处理方式受到质疑，于是他与武汉大学生命科学学院的老师和博士生合作完善了相关实验；面对审稿人不合理的验证操作，他开发了封装好的聚类算法工具箱以提升实验的可复现性，最终守得云开见月明。

图 4　科研瓶颈期，彭德华与导师桂志鹏的聊天记录

谈到论文作图，彭德华以“信达雅”三字概况了个中要领。其中，“信”是图片传达的信息需准确无误；“达”是图片大小统一工整对齐；“雅”是图片的配色构图要有审美追求。

最后，基于自身论文写作投稿的经验，彭德华对如何回复审稿人的意见进行了总结分享。

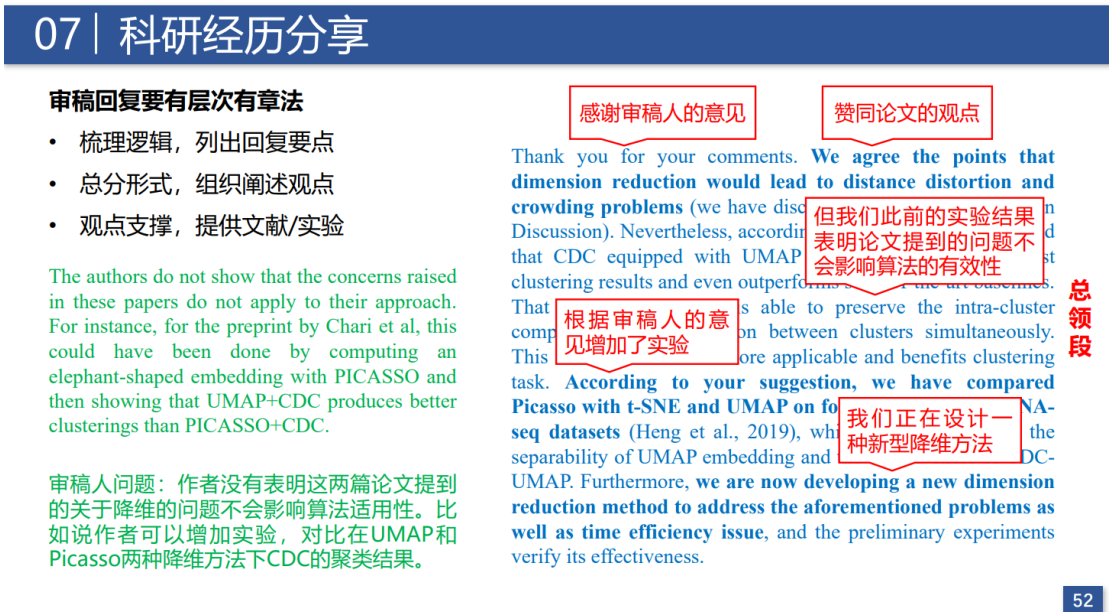

图 5　如何回复审稿意见

参考文献

[1] Gui, Z.*, Peng, D.*, Wu, H. & Long, X. MSGC: multi-scale grid clustering by fusing

analytical granularity and visual cognition for detecting hierarchical spatial patterns. Future Gener. Comput. Syst. 105, 96-118 (2020).

[2] Peng, D., Gui, Z.*, Wang, D., Ma, Y., Huang, Z., Zhou, Y. & Wu, H. Clustering by measuring local direction centrality for data with heterogeneous density and weak connectivity. Nat. Commun. 13, 5455 (2022).

>>>互动交流

提问人一：师兄在论文中和 PPT 中的图表都非常好看，想请教一下绘图使用的工具和方法。

彭德华：我作图使用的工具比较简单，常用的软件是 Origin 和 PPT，也有很多图片是用 Python 和 Matlab 代码生成的。我个人的做法是，先生成子图，再使用 Office Visio 合成一个高分辨率的图像。

提问人二：请问聚类算法常用的数学理论有哪些？掌握研究背景的理论部分需要多久呢？

彭德华：大部分机器学习算法中用到的数学没有那么复杂，所需要的数学基础主要是线性代数和微积分。另外，还需要具备应用知识的能力。如果想要做算法创新的话，很多理论的证明需要我们应用数学工具去证明一些命题，这个能力可能要慢慢地积累。

提问人三：复现论文的时候，数学公式如何“翻译”成代码？

彭德华：一方面，将数学公式“翻译”成代码是一个编程的问题，主要考验我们的编程能力。另一方面，有很多论文会提供开源代码，我们没有必要重复造轮子，可以在他人代码的基础上开发自己的工具。

提问人四：原始数据有多种属性，在聚类的时候如何确定对聚类结果更加重要的属性。

彭德华：在将算法应用于某一个场景时，我们确实会面对这样的问题。首先，我们可以使用统计学的方法进行特征分析，例如我之前提到的，通过变异系数等统计指标选择高变异基因。其次，可以借助主成分分析、降维等方法去分析每一个特征维度对数据的区分能力，然后考虑哪些属性需要弱化，哪些属性需要保留，以此提高数据分类或聚类的准确性，也可以提升计算效率。

提问人五：师兄你好，你的报告非常精彩，你的方向是机器学习的聚类方

面，在一个万物皆可深度学习的时代，你如何看待你的研究方向，是否有大环境的焦虑？

彭德华：这个问题其实你一说我就焦虑了，确实身边有很多人都在做深度学习，但是我认为，每个人还是要做出自己的特色。有相当大的一部分深度学习相关的研究论文，可能还是停留在应用层面，或者特征工程的一个层面，在理论的创新方面可能还是比较弱的。但是我认为自己的研究成果还是具备较高的理论创新性的，所以觉得这件事很有价值。当然，这并不代表我以后不会做深度学习，现在深度聚类也很火，我可能也会考虑将聚类和深度学习结合，去做一些深度聚类的工作。

总的来说，我们要对自己的研究工作有信心，要相信在自己的领域也能做出很多创新的东西，不一定要去和大家一起“卷”深度学习，每个人做出自己的工作的特色就可以了。

GeoScience Café 以“谈笑间成就梦想”为目标，于每周五晚 7:00 在实验室四楼休闲厅，邀请 1-4 位嘉宾，为大家带来学术报告或经验分享。报告内容包括摄影测量与遥感、地理信息系统、导航与定位服务等研究方向，听众可在报告结束后向嘉宾提问、与嘉宾交流探讨，同时每学期还会举办 2 期人文类讲座和 2 场导师信息分享会。每期报告会根据嘉宾意愿在 B 站开设直播，使不能来到现场的听众同步参与。报告 PPT 和视频会在征得嘉宾同意的情况下在 qq 群和 B 站上发布。

更多精彩内容（讲座预告、讲座回顾、报告 PPT、报告视频）敬请通过以下方式获取：