美赛经验分享

刘懿

参赛经历

- 2020年参加美赛，获得S奖
- 选题：A题：海洋温度变化对捕鱼业的影响
- 2021年参加美赛，获得M奖
- 选题：D题：音乐相似度、音乐家影响力、流派的演化等

- 下面会结合2021年D题的参赛经历来分享经验。

2021年参赛概况

集成集体音乐（ICM）协会确定了您的团队，以开发一种衡量音乐影响力的模型。这个问题要求您检查艺术家和流派的进化和革命趋势。为此，ICM为您的团队提供了一些数据集：

1)“influence_data”1代表艺术家自己报告的音乐影响者和追随者，以及行业专家的意见。这些数据包含过去90年中5,854位艺术家的影响者和关注者。

2)“full_music_data”2提供了16个变量项，包括音乐特征（如舞蹈性，速度，响度和调子），以及98,340首歌曲中的每一个的artist_name和artist_id。这些数据用于创建两个摘要数据集，包括：a.艺术家“data_by_artist”的平均值，b.表示跨年“data_by_year”

注意：这些文件中提供的数据是较大数据集的子集。这些文件包含您应为该问题使用的唯一数据。

为了执行这个具有挑战性的项目，ICM协会要求您的团队通过以下措施，通过音乐艺术家随时间的影响来探索音乐的发展：

1. 使用Impact_data数据集或其中的一部分来创建音乐影响力的（多个）定向网络，将影响者连接到追随者。开发可捕获此网络中“音乐影响力”的参数。通过创建定向影响者网络的子网来探索音乐影响力的子集。描述此子网。您的“音乐影响力”措施在此子网络中体现了什么？
2. 使用音乐特征的full_music_data和/或两个摘要数据集（包括艺术家和年份）来制定音乐相似度的度量。使用您的度量，流派的艺术家是否比流派的艺术家更相似？
3. 比较流派之间和流派之间的相似性和影响。什么是流派的区别，流派如何随时间变化？有些类型与其他类型有关吗？
4. 指示data_influence数据集中报告的相似性数据是否表明所标识的影响者实际上在影响相应的艺术家。“影响者”实际上会影响追随者创作的音乐吗？是某些音乐特征比其他音乐特征更具“感染力”，或者它们在影响特定艺术家的音乐方面起着相似的作用？
5. 从这些数据中确定是否存在可能标志着音乐发展中的革命（重大飞跃）的特征？在您的网络中，哪些艺术家代表着革命者（重大变革的影响者）？
6. 分析一种类型音乐随时间变化的影响过程。您的团队能否确定能揭示动态影响者的指标，并解释流派或艺术家随时间的变化？
7. 您的作品如何表达有关音乐在时间或环境方面的文化影响的信息？或者，如何在网络中识别社会，政治或技术变化（例如互联网）的影响？

向ICM协会写一份一页纸的文件，说明使用您的方法通过网络理解音乐影响的价值。考虑到这两个问题数据集仅限于某些类型，然后又针对这两个数据集共有的艺术家，您的作品或解决方案将如何随着更多或更丰富的数据而发生变化？建议进一步研究音乐及其对文化的影

2021年参赛概况

Contents

赛前准备

- Github上有一些比较丰富的参考资料库：

1. https://github.com/zhanwen/MathModel
2. https://github.com/HuangCongQing/Algorithms_MathModels

- 参考书：

1. 《数学模型》姜启源
2. 《MATLAB 数学建模方法与实践》卓金武等

- 我认为对于算法只要大概了解它们的原理即可，网上的参考实现很多，而且可能也用不到太高深的算法。
- 主要要弄清楚算法能做什么、不能做什么。

赛前准备

- 熟悉往年题，大概了解一下每道题涉及什么方面。
- 阅读往年的优秀论文，网上很容易找到。
- 也可以通过图书馆的EBSCO平台：https://lib.tsinghua.edu.cn/info/1184/3739.htm
- 查找关键词：UMAP Journal，可以搜到历年的O奖论文。
- 通过阅读论文，主要学习作者是怎样把题目和建模方法关联起来的，以及写作思路。

比赛过程

- 第一天（2月18日）早上6点发布题目，共有6题。
- 建议尽快讨论选题的问题，最好下午之前就能确定选题。
- 后期换题好比“临阵换将”，影响心情。
- 选题确定之后，第一步就是建模，同时也可能需要搜集数据（针对部分题目）。
- 虽然有些人可能提倡一人负责建模、一人负责编程、一人负责写作的分工方式，但是实际上我们当时一人负责一个或者两个小问，其他两位同学也会参与其中，提一些意见，并不是严格的流水线作业。最后写作的时候各写各的部分，其中一位同学在前期的工作量较少，所以在写作时就多写一些，特别是Summary sheet。

建模部分

- 我认为最难的部分就在于建模，建模部分涉及到的算法可能有很多类，比如优化、分类、预测、评价等。
- 尽快分析题目，确定建模的方法。
- 建议讨论时集思广益，队友们都尽量参与到讨论之中。
- 以2021年D题为例，我们很快就能捕捉到两个关键词：影响力、相似度。

建模部分

- 影响力：给定了艺术家之间的“**追随**”关系。可以用社交网络的方法建模。

1. 用python的networkx库可以很方便地实现一些网络方面的基本操作。
2. 以艺术家为节点，B追随A就画一条A->B的有向边。
3. 图构建完毕后，就大有可为了。

- 相似度：给定了很多歌曲的特征（十几项），已经量化。

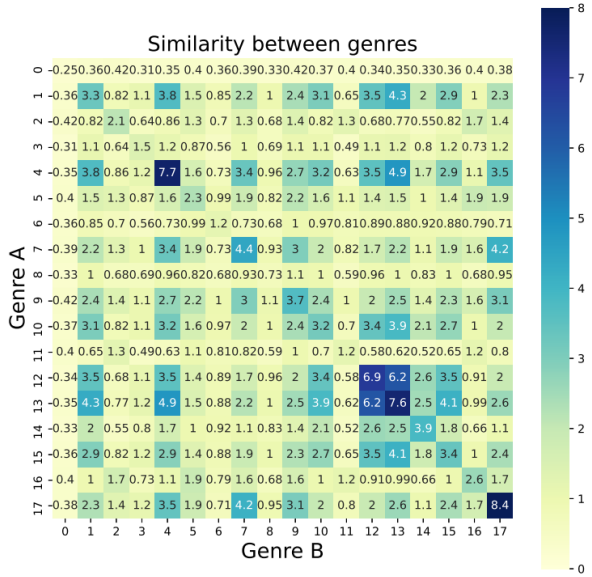
1. 可以归结成向量相似度（向量距离）问题。
2. 容易想到欧氏距离，简单粗暴；不过可以建模得更精细，比如给向量中的每一维赋予不同的权重（根据统计特征、层次分析法等）。

编程部分

- 常用工具：MATLAB、Python。
- 需要熟悉基本用法，MATLAB的函数、Python的库非常繁杂，不用面面俱到，但是要简单了解MATLAB（Python）能做什么、不能做什么！
- Excel也是好帮手，可以快速统计出数据的一些特征。
- 去年参赛时，核心的算法代码写得不多，主要是基于networkx的一些图算法（BFS、最短路等）以及聚类算法（KNN等）；并且用SPSS做了点数据处理（比如方差检验）。
- 不要忘了，数据前期处理以及后期展示也涉及编程！

编程部分

- 后期展示需要画图，论文里的图是否简洁美观，我认为会影响到读者的第一印象。
- 可以用MATLAB或者Python的matplotlib、seaborn等库画图，在网上很容易找到代码。
- 我们不仅画了常见的条形图，还画了热力图。



写作部分

- 常用工具：LaTeX。本地配环境可能有点麻烦，如果不想弄的话可以用下面提到的Overleaf。
- 基本语法不难，很容易速成，熟悉图、表、公式的基本用法即可。但是很容易写错（比如少个括号），所以最好写一段就编译一段，可以更快地定位到错误。
- 常用模板：mcmthesis。或访问：https://www.latexstudio.net/
- 写作平台：Overleaf（协作需要收费）
- 校内也有一个：http://overleaf.tsinghua.edu.cn（只有校园网能上，校外要用VPN）
- 当然用Word也可以，有些O奖论文就是用Word排版的。

写作部分

- 建议负责写作的同学赛前读一读O奖论文，看看这些文章的框架是什么样的，每个部分一般写什么内容。
- 英语水平要过关，可以求助Google翻译，但是汉译英总归不太好，而且效率低。
- 我一般是自己写，遇到不熟悉的词汇或者表述才会求助翻译。

写作部分

- 建议不要让一位同学写全部内容，这样任务太集中了，而且他还需要了解其他两位同学做了什么工作。
- 可以遵循“谁负责谁写”的规则，负责建模的同学写建模部分，负责编程的同学写编程部分，另一位同学负责整体把控、统稿，写问题背景等部分。
- 论文尽量完整，大段文字还是尽量少出现，可以用公式、图表来丰富内容，用斜体、黑体等提醒读者注意。

Definition 5.2. *Let A and B be two genres and S_A,S_B are sets of artists from corresponding genres. We define influence from genre A to B as*

$$I_g(A,B) = \frac{\sum_{p\in S_A}\sum_{q\in S_B} I_a(p,q)}{|S_A||S_B|} \tag{10}$$

- Case study看似是“语文建模”，实际上也能丰富文章的内容，建模结论更有实际意义。

时间安排

- 去年参赛时，我们在第一天上午就确定了要做D题，虽然之后曾一度感觉D题不太做得下去，但是还是坚持下来了。
- 之后埋头建模了几天，当时还没有返校，大家都在家，所以通过腾讯会议交流，效率比较高。
- 第三天开始写论文。由于三名同学都有编程基础，所以写论文时的代码问题（比如画图、排版）基本上自己就可以解决。
- 第四天晚上9点多，提交论文。
- 全程未熬夜，早上8点开工，晚上10点收工。

杂项

- 搜集数据、查找资料时，可以用Google，百度不靠谱。
- 如果上Google有困难可以用镜像站：https://www.library.ac.cn/
- 论文有页数限制（25页），注意官网的要求，今年的要求是整个PDF（包括summary sheet、附录等）一共只能有25页。我觉得不一定要附代码，也不一定要写满25页。
- 不要太赶ddl，虽然2月22日早上9点才是最终ddl，但是那时容易出现网络阻塞、交不上的情况，最好留点时间提前量。
- 提交时对邮箱地址没有要求，可以用QQ邮箱等，据说学校邮箱不太靠谱。

祝同学们取得好成绩！