公共安全事件全球数据在数据分析实训课程中的应用

王丽萍[1] 马静恒* 张洪萍[1] 李娜[1] 罗凯文[1]
（陆军勤务学院 军事物流系，重庆 401311）

摘 要：公共安全事件全球数据具有受关注度高、变化快、特征易理解的特点，将这些数据集应用于数据分析的实训课程中，能够激发学生的学习兴趣，提高实训课程的授课质量。实训课按数据理解、数据预处理、数据分析、数据可视化和报告提交的五步流程，以任务驱动和兴趣驱动相结合的方式展开。通过描述性统计分析和机器学习建模的编程实现，既让学生深入了解到公共安全事件的影响，也能熟悉基于python的数据分析方法，积累数据分析的实践经验，拓宽理论知识的应用范围，提高解决实际问题的能力。

[关键词]公共安全事件；数据分析；任务驱动；兴趣驱动；实训
【DOI】10.12252/j.issn.2096-6288.2021.07.1414

近年来，多起公共安全事件席卷全球，给全世界人类的正常生活造成了严重的影响。包括WHO在内的许多组织和机构实时的分享了全球公共安全事件数据[1]，为其在数据科学领域展开研究提供了数据支撑。数据分析在公共安全事件的监控、传播规律、预防与诊治等方面发挥了重要的作用，也增强了相关专业从业人员的责任感和使命感。可以说，公共安全事件的数据同数据分析早已息息相关。

数据分析是一门理论教学和实践紧密结合的课程[2]，而其中的实训环节更是大学生在步入工作岗位之前积累实践经验的有效手段。然而，高校实验室所采用的数据大多比较陈旧，且实验形式单一，学生既不关心数据来源，也不去深入理解数据涉及的业务领域。更重要的是，数据存在着行业跨度大、与现实脱节、分析结果难以验证等问题，这让许多学生产生了数据分析是否能够真正解决实际问题的疑虑。伴随着数据时代数据价值的提升[3]，大量真实数据的获取已变得十分困难，尽管公共安全事件数据有些尚未完善，但相比于实验室数据，其最大的特点就是受关注度高。因此，公共安全事件的基础数据集对于大多数学生来说也不会觉得陌生，加之数据随着课程的进行也在发生着动态变化，可以弥补以往数据分析实训过程中数据采集和结果验证的两个薄弱环节，所以非常适合作为数据分析课程的实训样本来使用。本文结合数据分析的课程目标，将公共安全事件全球数据运用到实训内容之中，以任务驱动和兴趣驱动相结合的方式[4]，尝试改进传统实训课程依靠多组示例数据集的流水线式教学方法，仅用一套数据集贯穿整个实训过程，以达到提高学生实践能力的目的。

一、基于公共安全事件全球数据分析实训流程的设计

作为全球关注的热点问题，公共安全事件数据具有受关注度高、变化快、特征易理解的特点。从2020年1月到2020年11月，全球已经有超过200个国家和地区报告发生公共安全事件累计人数超过5000万人，样本总量超过100 000组。在公开的数据中主要统计了国家和地区每日确诊人数、死亡人数和治愈人数。由于数据分析涉及的专业面广，且大多数学生对于医学流行病学方面的知识还缺乏了解，参考数据科学的流程和方法[5-6]，基于公共安全事件全球数据的分析实训教学仍然面向应用型教学，即以数据理解、数据预处理、数据分析和数据可视化作为实训主线，辅以数据采集和机器学习算法编程作为拓展实训项目，实训流程如图1所示。

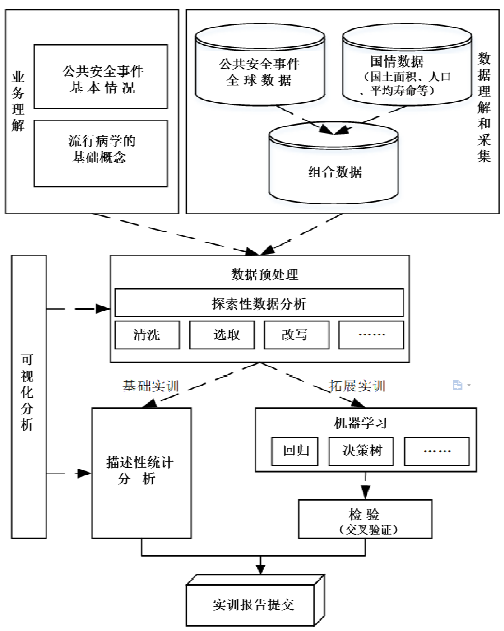

图1 基于公共安全事件全球数据分析的实训流程图

公共安全事件全球数据集主要来源于kaggle的开放竞赛项目以及GitHub网站[7]，采用基于Jupyter Notebook平台的python编程来逐级实现[8]。在实训的组织形式上，采用任务驱动，给学生明确实训即实战的基本要求，限定实训报告提交时间，实训成绩纳入期末总评成绩。从成立实训小组开始，所有工作就都交给学生自主完成。在实训的难度把握上，以观察、分析和理解为指导思想，以方法调用为基本手段。课程从以下五个具体方面对学生开展实训：

（1）数据理解。概要理解流行病学数据分析的基本概念，包括确诊率、死亡率、SIR模型。结合csv数据文件理解数据的存储形式，了解数据的来源和采集方式，认识数据集的特征变量。

（2）数据预处理。通过探索性数据分析（Exploratory Data Analysis，EDA）理解原始数据不完整、不一致、有噪声的性质特点，有针对性的对数据进行预处理。引导学生去采集和融合更多的特征（比如国家和地区的土地面积、人口密度、平均年龄、经济发展指标等）来丰富数据，为后续的分析阶段做好准备。

（3）数据分析。进一步明确数据分析的任务目标，由于是应用型数据分析，将本环节分成两部分，即描述性统计分析和机器学习建模，描述性统计分析阶段采用统计学的基本方法，机器学习建模采用回归和分类的相关算法。需要强调的是，最终的考核不应当以机器学习的预测结果作为依据，取而代之的是过程考核。另外，鼓励学生尝试用更多的模型进行拓展性实验。

（4）数据可视化。将可视化分析融入步骤（1）-（3），将各环节的分析过程以科学画板的方式迭代呈现。

（5）报告提交。引导学生从国家政策、地区环境、人口密度、人口比例、经济发达水平、流行病趋势等方面分析结果，形成个性化的实训报告。

二、基于公共安全事件全球数据的分析实训案例

（一）描述性统计分析

以特征工程结合EDA展开对数据的描述性统计分析，可以帮助学生更好的理解数据集以及了解公共安全事件是如何对全球人类造成影响的。

1.了解数据基本情况。

（1）实训内容

首先明确数据的基本情况，认识数据的不同类型，包括数据样本总数、统计的国家和地区数量、统计的起始时间和结束时间，区分定量数据和定性数据、结构化数据和非结构化数据。其次要构建新特征和修复旧特征，这些特征包括每日确诊病例数，每日死亡病例数，死亡率，生长因子（每日新增病例与前一日之比）。

（2）实训目标

学会使用python读取数据文件、处理数据集中的缺失值；理解矩阵数据集的概念；认识Dataframe[9]；熟练调用python方法处理Dataframe；掌握标准化和归一化的方法。

2.分析事件的影响和走势。

（1）实训内容

通过公共安全事件全球数据，理解病毒蔓延在国家和地区间呈现的指数增长趋势。由于给出数据在早期全球蔓延呈现出的明显的指数曲线，可以通过图形对比分析不同国家和地区间的公共安全事件走势情况。主要内容包括：

①在全球的传播趋势。绘制全球确诊人数和死亡人数走势的曲线图；绘制每日确诊人数和死亡人数的直方图；绘制每日病死率的曲线图。

②在中国的传播趋势。分别绘制武汉市、湖北省和湖北省之外省份的公共安全事件走势曲线图，可以直观的比较公共安全事件在中国的发展情况，分析中国的防控措施对病毒传播的影响。

③在欧美国家的传播趋势。分别绘制意大利、西班牙、英国、俄罗斯、巴西、印度、德国、美国的数据指数曲线，将多个国家的公共安全事件发展趋势进行对比分析。为了更直观的对各个国家病毒蔓延趋势进行比较，可以分别以每个国家出现第一例确诊患者的时间作为坐标起点绘制图形，再去分析各个国家病例的增长速度和增长原因。

④公共安全事件对国家和地区的影响。采用排序的方法筛选出病毒影响程度最大的前10个国家，分析病毒对这些国家的影响，分析不同国家不同死亡率的原因。采用高级的图形可视化方法，动态的观察病例在地域上的分布情况，如图2所示。

（2）实训目标

掌握基础图形的绘制方法；学会分析和区分指数曲线、对数曲线；理解平滑性的概念；熟练调用pandas、numpy、matplotlib、plotly、pyecharts库的各类方法；理解并绘制交互式图形。

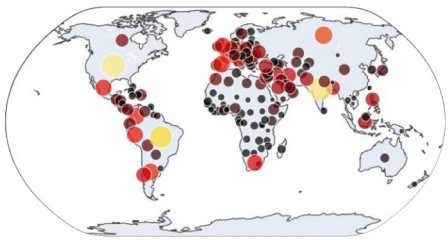

图2 通过数据可视化库绘制的公共安全事件全球数据地图

3.SIR模型验证。

（1）实训内容

流行病的基本数据模型用于研究疾病的传播速度、空间范围等。其中，SIR是流行病学最著名的模型之一，它不仅可以应用于传染病传播与扩散的研究，如对埃博拉病毒蔓延趋势的预测，还可以应用于其他众多领域。利用不同传染病群体的不同感染状态，可以构建公共安全事件的传播扩散模式。SIR将复杂网络群体中的所有个体划分为三种状态，即易感的（S）、受感染的（I）和康复的（R），可以使用这些状态的组合表示不同状态之间的转换顺序以及流行病所处的阶段。

实训中，需要学生依据公共安全事件的全球数据来求解微分方程，绘制出SIR模型的图形，并结合图形分析各个国家的公共安全事件事件所处的阶段。

（2）实训目标

了解SIR模型的原理；能够使用python定义函数及求解微分方程；能够通过对参数的优化来调整模型的拟合。

（二）机器学习建模

1.数据的补充和转换。

（1）实训内容

考虑到原始数据特征较少，同时国家与国家之间的国情差别较大，在对样本做进一步分析之前，需要补充新的特征变量，包括：国家人口密度、性别比例、65岁以上人口比例、平均年龄和人均GDP等。为了保证数据准确，可以仍然选择2.1.2节中提及的8个国家，这一步还需要学生完成一些必要的计算，比如人口密度是人口总数和国土面积的比值，人均国内生产总值是国内生产总值和人口总数的比值。

对原始数据需要进一步的改写，将确诊人数和死亡人数进行对数变换，对特征变量创建散点图矩阵并进行相关性分析。

（2）实训目标

学会处理问题数据的常用方法；理解数据分析的迭代原则；理解创建新特征的必要性和合理性；通过数据变换来提高系统的精度。

2.分类和预测建模。

（1）实训内容

采用交叉验证，区分训练样本和检验样本。采用多元线性回归模型对局部数据进行训练，掌握多重判定系数R^2的计算方法，绘制回归平面并观察拟合情况。

采用分类回归树（CART）对数据集进行训练，尝试使用不同的特征集生成决策树。结合决策树的生成过程理解模型过拟合问题，使用限定分支层数、限定分支样本最小数量的方法来实现简单的剪枝算法[10]。

（2）实训目标

进一步理解交叉验证的内涵；理解过拟合、泛化、相关性、降维、损失函数、剪枝等概念；通过误差度量的方法来评价学习算法的性能；熟练使用python的scikit-learn工具包。

三、实施效果分析

本实训案例设计了针对课程的五大理论知识，分别是数据理解、数据预处理、描述性统计分析、机器学习和数据可视化，指导学生使用python编程，熟悉了pandas、numpy、matplotlib、pyecharts、scikit-learn等核心工具包的使用技巧，另外，从实训课的实施效果中还获得以下启示：

（1）公共安全事件全球数据为数据分析提供了广阔的实训空间。这是一个任务驱动和兴趣驱动相结合的实训案例，当学生掌握了数据分析的基本方法之后，他们会有更丰富想法去尝试拓展性实验。比如尝试使用更多国家和地区的数据分析流行病学的一般规律，或是采用随机森林、XGBoost、LightGBM、深度学习等更高级的方法去预测实际事件走势。

（2）学生对公共安全事件有了更深入的了解。尽管他们都不是医学专业学生，但大多数人对　　SIR模型展现了极高的兴趣，由此可见拓宽知识领域对于提升数据分析能力的重要性。为此，教师在理论课程的讲授中，可以有的放矢的增加业务知识的讲解，能帮助学生在实训过程中更快的上手。

（3）这是以项目化的方式组织的专题实训，让学生以小组的方式完成项目，学生既是项目的参与人更是负责人，考核成绩会根据学生在项目中的实际贡献大小加以区分，在完成项目的同时也能够锻炼他们的合作和沟通能力。

（4）由于采用的是实际生活中的数据，分析结果又能从现实中得到验证，这对于教师的教学引导提出了较高的要求。但对于开放性话题的讨论，却可以拉近学生与老师之间的距离，从而进一步的激发学生的学习热情，有利于学生认清数据分析的定位与作用，因此应该善加利用。

四、结束语

针对数据分析实训课程中数据样本陈旧、行业跨度大、特征难以理解等特点，采用公共安全事件全球数据作为课程实训的数据集，设计仅依靠一套数据集且具有普遍适用性的数据分析实训流程，改变以往根据碎片数据来准备教学内容的被动形式。采用任务驱动和兴趣驱动相结合的方式，让学生从数据分析的项目实践中去获取经验，符合数据分析课程的教学目标，对提高学生分析解决实际问题的能力具有一定的帮助。

参考文献

[1]World Health Organization.Coronavirus Disease 2019 Situation[EB/OL].[2020-04-19].

[2]王鸣，薛燕.环境科学专业“数据分析与实验设计”课程教学改革探讨 [J].实验技术与管理，2020，37（2）：164-167.

[3]汤羽，林迪，范爱华，等.大数据分析与计算[M].北京：清华大学出版社，2018，3：6-13.

[4]陈雯柏，陈启丽.基于兴趣与任务驱动的人工神经网络课程改革理念[J].计算机教育，2015，18：29-31.

[5]阿尔贝托·博斯凯蒂，卢卡·马萨罗.数据科学导论-Python语言实现[M].于俊伟，靳小波，译.北京：机械工业出版社，2018，3：73-121.

[6]拉姆什·沙尔达，杜尔森·德伦，埃弗雷姆·特班.商务智能-数据分析的管理视角[M].赵卫东，译.北京：机械工业出版社，2018，5：176-182.

[7]GitHub.CSSEGISandData/[EB/OL].[2020].https://github.com/CSSEGISandData/.

[8]薛煜阳. Jupyter Notebook 在 Python 教学中的应用探索 [J].信息技术与信息化，2018（7）：168-169.

[9]丹尼尔·陈.Python数据分析-活用Pandas库[M].武传海，译.北京：人民邮电出版社，2020，1：2-97.

[10]李航.统计学习方法[M].北京：清华大学出版社，2019，5：80-87.

通讯作者：马静恒（1982－），男，四川泸州，硕士，讲师，研究方向为数据科学和智能计算。