

面向任职教育的“数据挖掘与机器学习” 课程教学研究

李莎莎, 刘 丹, 唐晋韬

(国防科技大学 计算机学院, 湖南 长沙 410073)

摘 要: 随着军队信息化、智能化建设的深入, 人工智能技术逐渐渗透到军队武器装备保障、指挥作战、情报分析等各个方面。本文所要探讨的“数据挖掘与机器学习”课程属于军队初级指挥军官的任职培训课程。本课程针对参与培训人员专业多样, 学习基础参差不齐等问题, 设计了有针对性的教学内容和教学手段, 实践结果表明了教学设计的有效性。

关键词: 任职教育; 数据挖掘与机器学习; 课程

中图分类号: G642 **文献标识码:** A **文章编号:** 1672-8874 (2019) 01-0108-06

Exploration of Teaching Data Mining and Machine Learning for Military On-Post Training

LI Sha-sha, LIU Dan, TANG Jin-tao

(College of Computer Science, National University of Defense Technology, Changsha 410073, China)

Abstract: With the ongoing construction of informationized and intelligentized army, artificial intelligence technology has gradually penetrated into the military equipment support, command operations, intelligence analysis and other posts. The course of Data Mining and Machine Learning discussed in this paper is aimed at on-post training for military junior commanding officers in artificial intelligence positions. Due to the diversity of the major and the uneven learning foundation, the course designed teaching content and teaching methods accordingly. The practical results show the effectiveness of the teaching design.

Key words: On-post training; data mining and machine learning; course

一、引言

随着军队信息化智能化进程的不断推进, 基于信息平台的新型作战力量建设要求大批能够快速有效地分析海量信息, 从中提炼出最有价值的信息以便科学指挥与决策的人才以及支持这项工作的系统, 其核心技术就是人工智能与大数据技术。针对以上需求, 衍生出人工智能岗位培训任务, 该任务致力于在部队各个岗位大力培养掌握了人工智能与大数据技术, 并具有研制/使用智能

武器装备与系统能力的专业技术人才。

“数据挖掘与机器学习”课程是人工智能专业岗位培训的专业课程之一。数据挖掘与机器学习是人工智能领域的一个重要分支, 也是当前数据智能时代的热点领域之一。随着“大数据”时代的降临, 决策将日益基于数据和分析而作出, 而并非基于经验和直觉。数据挖掘技术致力于并通过统计、在线分析处理、情报检索、机器学习、专家系统和模式识别等诸多方法从大量的数据中获取用于决策的知识, 辅助决策者做出正确的决策。机器学习是通过模拟或实现人类的学习行为,

以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心问题之一，也是目前数据挖掘领域最为强有力的技术手段。

数据挖掘和机器学习是高度综合的前沿学科，也是实用性极强的应用学习。本文所要探讨的“数据挖掘与机器学习”课程主要是针对人工智能岗位初级指挥军官，旨在介绍现有技术及其在不同领域的应用案例，激发学员从自己专业对该领域进行探索、学习、和应用的兴趣，对我军创新型复合型军事人才培养具有重要意义。要求学员通过该课程了解人工智能方向的基础知识和基本方法，熟悉机器学习、数据挖掘相关的系统和工具的操作实践技能及组织运用方式，能够从事人工智能相关的军事应用系统的应用和管理，能够利用人工智能技术解决本职工作中所遇到的实际问题。

“数据挖掘与机器学习”课程通常是针对计算机专业学员，要求学员具备“数据结构与算法”、“数据库原理”、“人工智能导论”等课程基础，且要求学员能熟练应用数据挖掘的常用工具也需要学员有一定的编程基础。而接受任职教育的学员专业背景比较复杂，学习能力及学习诉求多样，任职教育难度较大。本文首先深入分析了接受任职教育的学员的特点，并针对性地进行了教学内容和教学方法的设计，提出了针对不同问题的解决方案，从考核结果看，教学达到了预期目的。

二、培养对象特点分析

为了更加深入地了解接受任职教育学员的学习背景，我们利用问卷网向接受本次任职教育的15名学员发出调查问卷，结果如下：

（一）学员本科专业分布情况

从图1的学员本科专业分布情况可以看出，所有学员中，计算机专业类的学生只占30%，其中包括计算机科学与技术专业18%，信息与计算科学6%，信息安全6%。这也就意味着大部分学生缺乏数据结构、算法等计算机基础，也缺乏计算思维的系统培养。

（二）学员基础课程学习情况

考虑到一方面即使对同一专业，不同院校的课程设置也有所不同；另一方面，学生可以根据自己的喜好，选择一些自己感兴趣的课程。因此，

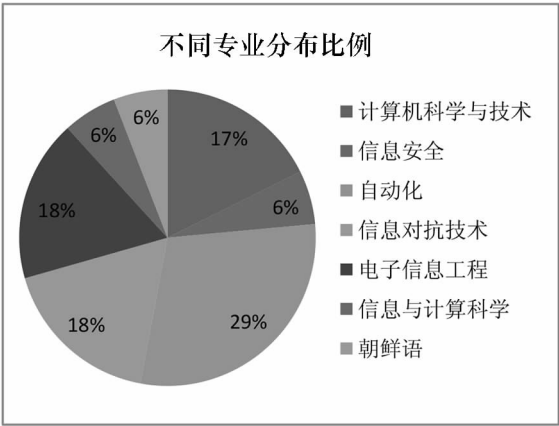


图1 学员本科专业分布图

本课程对学生的课程学习经历进行了进一步的调查，结果如图2所示，图中每条柱状线代表一门课程的学习情况，柱状线右侧的数字表示学习了该课程的学生数量。图2表明，一些数学基础课，如高等数学和概率论，学习率达到93.3%（仅一人未学习）；其次是大学计算机基础，学习率为80%；再次是程序设计和数据结构基础，学习率为60%~70%。其余课程，如离散数学、算法设计与分析、数据库原理、人工智能等课程的学习率都不足50%。从以上调查结果来看，学生学习“数据挖掘与机器学习”的数学基础还是有的，但是计算机方面的基础就显得差距较大，一半以上的同学都不具备数据库原理、人工智能等“数据挖掘与机器学习”课程先导课程的学习经历。

（三）学员编程能力

在学员的编程能力这部分，本课程调查了图3所示的四个问题。调查结果表明，学生主要的编程语言是C和C++，使用率约为67%，了解Java的同学不超过20%，完全没有接触过Python。另外，学生写过的最复杂的程序不超过1000行。事实上，60%的同学大学四年的代码总量不超过1000行，基本没有软件开发经验。以上调查结果说明，学生的编程能力停留在入门阶段。

（四）学员期望

考虑到学员的兴趣才是学习最强的动力，本课程还对学员自身的学习期望进行了了解。学员期望总结起来表现在三个方面：（1）与部队实践相结合；（2）深入浅出，照顾基础；（3）提供好的平台、资源、学习方法，以保证学习的可延续性。具体调查情况如表1所示，表中每一行内容对应一个学员期望。

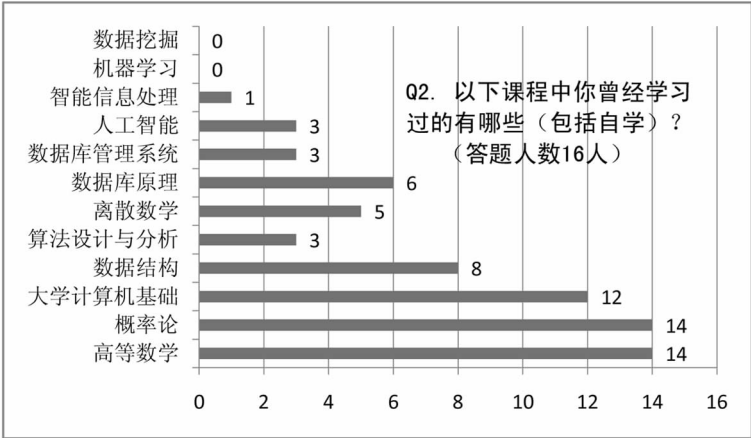


图 2 课程学习情况分布

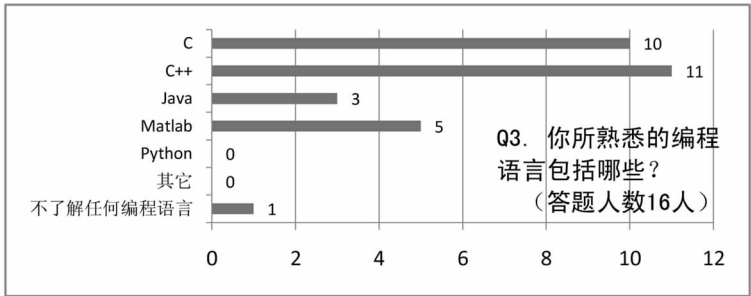


图 3 学员编程能力调研

表 1 学员期望调研

学员序号	学员 期 望
学员 1	期望培训可以兼顾基础，增加拓展，开发思维，适应部队需要。
学员 2	我们基础比较薄弱，希望能够理解。
学员 3	倾向于应用。
学员 4	希望老师上课时，能够结合案例进行分析讲解，不要纯理论的讲解。
学员 5	希望可以提高学习能力，利用好学校的资源，能把好的思想和技术带回单位，为以后的工作打下基础。
学员 6	希望可以老师在授课的同时多跟我们实际工作方面有贴合。
学员 7	希望介绍更多学习方法，学习资源，学习平台。
学员 8	理论结合实践，尽量多的参考资料下鼓励自学。
学员 9	无
学员 10	希望能教一些面向部队的知识。
学员 11	我希望在讲理论的时候可以告知理论明确的实践指向，这样我们学习可以有方向性。
学员 12	希望能对该专业课程有更多的了解，并在此基础上能尽可能的深入学习，希望能多掌握部队实用的技术与知识。
学员 13	多多指教。
学员 14	希望老师的课程能够深入浅出，照顾一下我这种一知半解的学生。
学员 15	无

根据以上的对学员学习基础和学习期望的调查,我们发现:(1)虽然学科基础参差不齐,但由于学员95%以上都是来自工科专业,基本的数学基础都是具备的。(2)学员编程基础弱,实践教学困难。(3)学员更倾向于与未来工作需求相贴合的实践导向型学习,更倾向于掌握能够持续学习的学习方法。

三、教学内容及教学方法

针对调研结果所发现的学科基础差、编程基础弱、理论联系实际需求强烈等问题,本课程对具体的教学内容和教学方法作了以下设计。

(一) 选择 Python 作为实践语言,基于实训平台进行基本语法练习

Python 是人工智能时代最为热门的编程语言之一,其特点是易于学习、易于理解,对非计算机专业的人员学习起来极为友好,目前已成为国内外众多高校非计算机类编程基础课的首选语言。另外,Python 语言还兼具易于扩展、有丰富的标准库的特点,有利于任职教育结束后的自我学习。综上,我们有理由相信 Python 是最适合人工智能岗位任职教育的编程语言^[1]。

学习一门语言,前期需要大量的基础语法的实训练习。本课程借助国防科技大学开发的 Trustie 实训平台^[2],利用学生课余时间,以周为作业单位,展开进阶式的基础语法练习。借由 Trustie 实训平台,老师可以根据要练习的主题发布通关式练习,每个题目由通关任务、背景知识、参考答案三部分组成,参与实训的学员可以在平台提供的程序编辑区编写程序,结果输出区会给出编译结果。最后,平台自动将程序的输出和参考答案相比较,并给出是否通关的提示。学员可以查看参考答案,但是查看参考答案的题目将不计分。利用实训平台进行编程学习有以下优点:一方面,该平台的通关式答题提升了学习的乐趣;另一方面,实训平台提供每个题目的知识背景,应用示例,并能自动判定程序对错,学生能够边学习边练习,并自己找出错误,给了学员很大的自我学习空间,与学员提出的“希望提供更多的学习方法和资源,不谋而合”。最后,本课程以最终实训结果计入考核成绩为激励,确保实训完成度和完成质量。

(二) 采用多样教学方式,增强知识点的理解

与记忆

学员在刚接触到新的领域时,由于要转换到不同的思维方式,难免感觉新知识难以记忆、理解、掌握。一方面,有的学员会产生畏难情绪,放弃学习;另一方面,有些学员会强行记忆,没办法把知识转换成思维的一部分加以理解、掌握,记忆不深,难以灵活运用。课程针对这一问题,设计多种不同的课堂教学方式,旨在以趣味性的方式调动学生的学习主动性,促使其形成独立思考,质疑,探索,讨论,再思考的良性循环,最终将知识点完全消化吸收,转换为自己思维习惯的一部分。例如,在“决策树分类”这一章中,课程首先讲解决策树的算法原理;然后给出一个分类案例,并详细讲解了如何用 Python 实现用决策树解决该案例中的分类问题;最后,通过竞赛的方式,要求学员根据课程内容,通过调节参数、转换模型、改变训练方式等提高算法性能,并解释为什么所用方法能改进算法性能,竞赛排名靠前的同学将获得奖品。通过这次竞赛,有效提升学员对参数、性能评估的理解。又如,在“神经网络分类器”这一节中,通过合作学习^[3]的方法进行扩展阅读。具体来说,扩展阅读分为三个主题,“循环神经网络”、“卷积神经网络”、“编写 python 程序实现感知机分类”,本课程将学生分为 5 组,每组 3 人,分别阅读三份材料,然后互相交流材料内容,最后每组随机抽取一位同学讲述其未阅读的那份材料的内容,其余人员点评。再如,在“数据挖掘的基本流程”这一节,则选取恰当案例,采用分组讨论的形式,用数据挖掘的思维方法分析案例中问题及其解决方案,并将形成的结论做成海报,向所有学员讲解,由于所有学员讨论的都是同一案例,因此也可以促进各组之间的交流和更全面的思考。

(三) 以学生的专业背景为课题分组作业

本课程包含一次贯穿整个课程的分组作业。分组作业的目的有三个:一是串联课程所学的知识点,锻炼学员运用数据挖掘与机器学习的思维解决实际问题的能力;二是扩展学员的知识面,了解更多数据挖掘、机器学习与自己任职相关的应用;三是锻炼学生自己查资料解决问题的能力。由于参与培训的学员本身工作需求各异,课程选择了针对不同工作需求的课题,要求学员通过查资料和阅读论文,了解国内外在相关领域数据挖掘与机器学习技术的应用情况,并选择其中一个

具体的问题,谈谈如何收集数据、收集哪些数据、数据中可能存在哪些问题、如何清理数据、如何利用课程所学的模型对问题建模、如何评价模型的好坏等等。具体要求是:3~5 人一组,根据自己的专业和兴趣,在以下课题中选择一个,通过查阅相关资料,完成任务理解、目标设定、数据需求分析、模型选择等任务,每周形成周报告,最后提交项目总结报告,并完成口头报告。分组作业一是学员个性化管理。例如,基于学员信息管理系统实现学员个性化管理,并能支撑教学体系调整的决策。分组作业二是情报分析。例如,基于网络信息采集、情报监听、无人机侦查等获取的信息,实现特定情报预警等。分组作业三是智慧医疗。例如,依赖医院的信息系统构建智慧医院,实现智能分诊、预后跟踪、推荐用药、甚至辅助诊断。分组作业四是装备保障。例如,基于武器装备管理系统实现故障预警、装备技术分析、保障人员和物资智能调配等。其它分组作业还包括训练管理、后勤保障等。

(四) 单元练习加综合练习的渐进式实践方法

鉴于任职教育更重实际操作能力,因此在整个课程中,实践教学比例较高。虽然前期大量的实训练习帮助学员了解了 python 编程的基本语法,但是由于学员不了解数据结构、不了解面向对象编程,在使用 Numpy、Pandas、sklearn 等开源扩展包的过程中仍存在很大困难。因此,本课程采用每个章节一练习,最后利用单元练习的成果进行综合练习的方式开展实践教学。本课程设计了三个综合练习,分别针对结构化数据、图像数据、文本数据,三个综合练习覆盖了主要的数据挖掘问题,包括频繁项集挖掘、分类、聚类分析和时间序列分析,每个综合练习都贯穿了数据理解、数据清洗、结果分析等数据挖掘关键环节。每个练习分为若干问题,按数据挖掘的几个关键步骤组织,帮助学员进一步消化掌握数据挖掘解决问题的思路。

实验一是美国国会投票模式分析。实验目的:熟悉 Pandas、matplotlib、sklearn 等 Python 模块,掌握结构化数据分析的基本方法、流程。背景知识:美国国会 (Congress, United States) 的主要任务是立法,制定影响每一个美国人的法律,立法经常需要在一些有争议的问题上作出决定,例如医疗保险改革、战争、贸易汇率等,法案 (Bill) 经过一系列的审议程序后才能成为法律。法案一

般经过提出、委员会审议、全院大会审议等程序。一院通过后,送交另一院,依次经过同样的程序。法案经两院通过后交总统签署;若总统不否决,或虽否决但经两院 2/3 议员重新通过,即正式成为法律。在议案从提出到最终颁布的过程中包含了很多投票环节,由于这些议案是关系着每一个美国人的利益,因此投票中往往体现出了不同党派的利益纷争及其政治立场。通过分析同一利益集团的投票人的投票结果,可以发现法案之间的关联关系;根据投票人的投票情况,也可以判断投票人所属党派。实验任务:(1) 挖掘美国国会投票记录中的关联规则;(2) 根据投票结果判断投票人所属政党。实验方法与步骤:(1) 数据收集。本实验的数据来自加州大学埃文分校所提供的“1984 年美国国会投票记录”。自打开 data/house - votes - 84. data 文件,该文件中收集了共 435 条记录,每条记录对应一个投票人。每条记录有 17 个域,分别对应投票人的 17 个属性。属性的具体含义参见 data/house - votes - 84. names。(2) 数据质量分析与预处理。(3) 挖掘投票记录中的关联规则。a. 首先构造事务数据库,对数据重新编码。例如,民主党人编码为 1,共和党编码为 0,“赞成法案 1”编码为 3,“反对法案 1”编码为 4,“弃权法案 1”编码为 5,以此类推。b. 利用 fpgrowth 或 aprior 算法进行关联分析,挖掘频繁项集并生成关联规则。改变支持度和置信度的阈值。c. 观察得到的关联规则集的数目,并给出图表可视化。d. 分析关联规则是否都是合理的,对合理或不合理的规则试举一二例,分析其原因。(4) 根据投票结果判断投票人所属政党。a. 划分训练集和测试集。b. 在训练集上,以“民主党”“共和党”为类标,其余属性为分类属性,利用决策树、SVM 等算法,训练分类模型。c. 针对某个分类算法,改变参数,在测试集上观察分类精度的变化。d. 比较不同算法的分类精度。

四、考核评估

学生最后的成绩组成由卷面成绩,报告成绩和实训作业三部分组成。其中,卷面考核的知识点涵盖数据挖掘的流程、数据质量分析、关联规则挖掘、决策树分类、神经网络分类器、聚类分析、时间序列分析,知识点覆盖率达到授课内容的 90% 以上,难度中等;报告成绩由书面报告和

口头报告两部分成绩组成，各占 50%。书面成绩考察调研资料是否详实可靠，分析是否合理，格式是否规范；口头报告由学员打分，取平均分为最终成绩，具体评分表如表 2 所示，A、B、C、D 代表四个分组报告，每个报告考察资料详实可靠性、分析合理性、逻辑是否清晰、PPT 是否图文并

茂以及表述是否生动。要求每位学员为每一组在每一项打分，每项成绩 0－10，0－6 分为合格，6－9 良好，9－10 优秀。每组总分 50。最终学员的成绩分布如表 2 所示，三项成绩相加后，所有学生的平均成绩 83 分，及格率达到 100%，达到预期授课目标。

表 2 考核成绩分布

分数段	90－100（优秀）	80－89（良好）	70－79（中等）	60－69（及格）	59 以下（不及格）
人数	3	8	5	0	0
比例	18.75%	50%	31.25%	0	0
平均成绩		83	及格率		100%

五、存在的问题与改进方向

从课程实施过程中的课堂氛围、随堂练习情况以及最终考核评估的结果来看，该课程的教学设计取得预期效果，但是依旧存在以下一些问题，本文针对以下困难也提出了相应的解决方案：

（一）不同专业之间存在知识鸿沟

一方面，根据课程前期的调查问卷，我们了解到学员在编程以及计算机基本理论方面知识较为缺乏，这就使得学员对“链表”、“队列”、“树”、“算法复杂度”、“数据库表格”等课程中需要的基本概念理解不足，仅靠课堂上临时的知识补充难以熟练掌握的要求；另一方面，学员来自其它专业，这些专业同样涉及本课程所提到的“熵”等概念，并带着对这些概念的固有理解，因此当与其头脑中的固有概念产生冲突或不一致时，导致学员理解困难。针对以上问题，在未来的课程教学中，本课程拟总结本期教学中的经验、标记教案中的学员应掌握的先导知识点，整理出前期阅读材料，提前发给学员进行预习，并结合课堂上的先导知识点讲解或预习小测，巩固学员对先导知识的掌握。

（二）教学进度参差不齐

学员的基础参差不齐，也造成了其学习进度参差不齐。这个问题在实践环节尤其严重，实践进度慢的学员会逐渐产生畏难和依赖情绪，降低其学习效果。针对这一问题，在未来的课程教学中，本课程拟采取分级分组、附加奖励的策略，根据单元实践的效果，将学员分成不同小组，每

个小组完成不同难度的作业，且每次作业的难度由简入难，难度最大的题目提供附加奖励。

（三）急需实验平台建设

本课程的实践数据量大、计算量大、实践环境安装复杂，对计算平台的要求较高。针对以上问题，在未来的课程教学中，本课程拟构建统一的基于云计算的实践平台。

六、结语

本文所探讨的“数据挖掘与机器学习”课程针对军队初级指挥技术军官人工智能岗位的岗前培训课程。在开课之初，课程先对参与课程的学员的专业背景、基础知识的掌握程度等信息展开了问卷调查。调查结果表明，参与课程的学员呈现出专业背景复杂，学习基础参差不齐，编程能力弱等特点。针对以上特点，课程设计了有针对性的教学内容和教学手段，实践结果表明了教学设计的有效性。

参考文献：

[1] 张良均,杨海宏,何子健,等. Python 与数据挖掘[M]. 北京:机械工业出版社,2016:3－5.

[2] 国防科技大学. Trustie 实训平台[CP/OL]. [2017－10－09]. <https://www.educoder.net/>.

[3] [美]安妮塔·伍尔福克. 伍尔福克教育心理学[M]. 12 版. 伍新春,译. 北京:中国人民大学出版社,2015: 478－479.

（责任编辑：邢云燕）