关于机器学习的调研

常傲
西安交通大学，计算机系，710049，西安

摘要（Abstract）：本文通过对西安交通大学信息库和国家知网等网站的有关机器学习的调研，分析现今中国乃至世界的机器学习发展状况并使用抽样调查法调查人工智能对社会的影响，讨论预测中国应该在当前的国际形势下对机器学习抱有怎样的态度。

关键词（Key Word）：机器学习，人工智能，数据库，云计算

当今社会，机器学习已经从过去的高高在上变成现今的家喻户晓，越来越多的人对机器学习感兴趣、认为它能改变人类的未来；同时又有一些对机器学习有所研究的人对此抱有消极态度。那么，究竟什么是机器学习，它又对我们的生活有着什么样的影响呢。

本文将从教育和网络安全两方面来阐述机器学习的影响。

1 机器学习对教育领域的影响

机器学习很明显是使教育智能化，改变传统的教育方式，提高学习效率和各级领导和师生的对应关系，使教育变得更加开阔、透明、效率。

传统的高校课程教学主要关注学生理论知识学习，虽然学生理论基础较为扎实，但是实践能力比较薄弱，不适应新的新形势下的产业对人才需求。新的形势下高校需要更加深入了解产业需求，突出教书育人与实践更加紧密结合，学校的学习不应该只停留在表面的书本而应该是围绕国家的产业进行补充扩展，为国家输送人才。

课程理论教学部分既强调理解经典算法，又注重引入最新理论成果。在工科的基础课程中，高数、线性代数、概率论等均是非常重要的基础课程，也是机器学习的基本功。如果学生在这方面得过且过、不求甚解只求不挂，那么对未来的机器学习领域的研究必然是相当的吃力。同时，更要注重创新、与时俱进，很多高校抱着数年如一的老教材和 PPT 反复授课，或者是换汤不换药地换个封皮，这使得课程相当地乏味无聊。如果能够加入一些新的元素，结合目前的社会发展状况进行合理的拓展，对于教学质量无疑是极大的提升。

课程实践教学部分既强调解决问题能力，又注重方法优化。在现有课程教学中，多数教学强调使用现有的程序解决面临的实际问题，往往忽视学生编程能力提升和方法的优化。而当面临新的问题，无法直接利用现有方法直接解决，或者利用现有方法解决问题时得不到理想结果时，常常无计可施。课程建设和改革需要积极探索提升学生编程能力和优化水平。如今网络上有如此多的开源代码，可以帮助学生学习，但很多人往往不知道该怎样检索和使用，再加上国内搜索软件的糟糕，无异于黄钟毁弃瓦釜雷鸣。

课程建设突出产学融合新特征，体现创新型人才培养新目标。课程的教学大纲、教学资源、教学过程和教学质量反馈构成了一套动态闭环更新的教学过程，产学融合则能够在此过程中帮助有效提升课程整体质量。其中教学资源建设与完善需要与企业密切合作，通过委托研究、合作开发等形式加强合作。面对快速发展的人工智能相关科技，产业需求也在不断提升，需要与企业互动交流，依据最新技术需求与技术革新，对课程的教学目标、教学大纲等进行修订，并通过完善教学资源和教学过程，实现产学融合的新型课程构建。

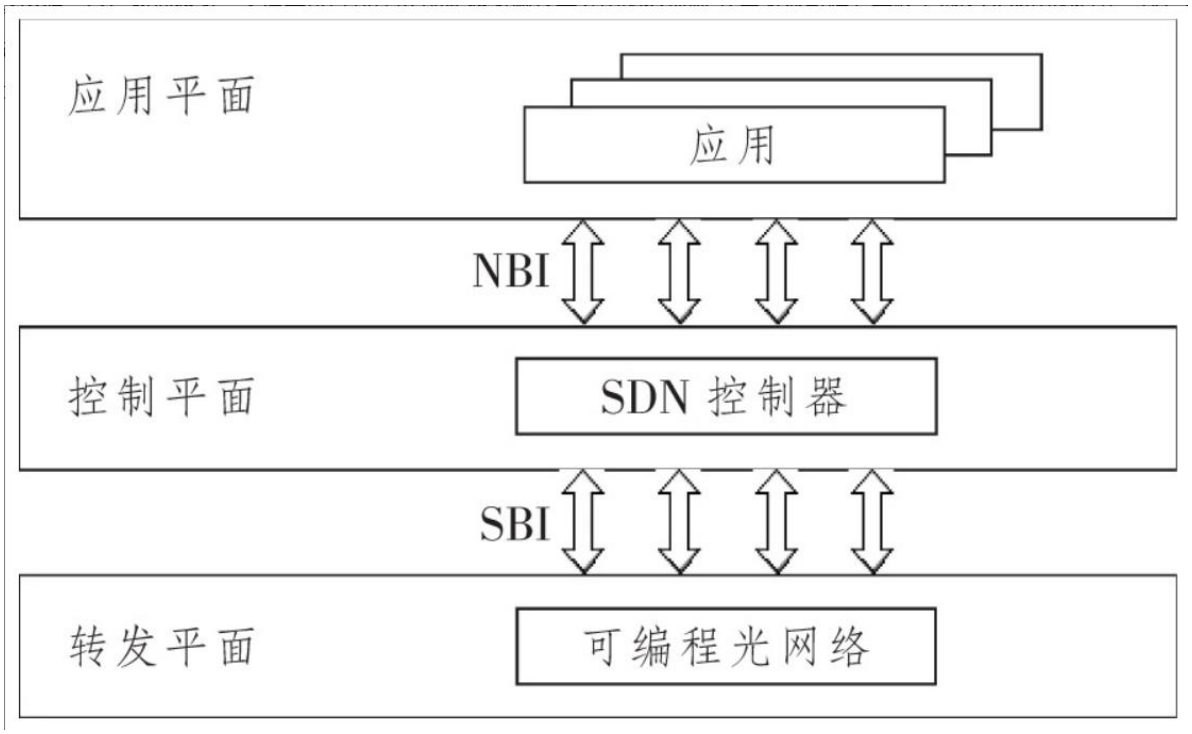
2 机器学习在安全检测的应用

软件定义光网络(Software Defined Optical Networks,SDON)是将软件定义网络的概念和技术应

用于光网络的一种新型网络结构。SDON 控制平面构建灵活、开放和智能的光网络体系结构,并通过南向接口及北向接口实现灵活的服务访问和硬件控制。但集中式的 SDON 控制平面在日常应用中可能会受到恶意的攻击或入侵,特别是当受到攻击者劫持时,整个光网络可能会陷入瘫痪。如何防止 SDON 收到敌人的攻击就成了非常重要的问题。

Chandola 等对基于统计学、信息论和机器学习等手段的异常检测分析方法进行了比较,认为机器学习是一种具有发展前途的入侵检测方法;Sequeira 等提出了一个基于主机数据收集和处理的实时入侵检测,可以有效识别计算机终端侧伪装者与真实用户之间的区别。可见机器学习对于网络安全方面也有极大的贡献。

SDON 模型



SODN 的潜在威胁包括: 未经授权的访问、数据泄露、数据修改、Dos、安全策略误用。

基于时间序列的异常检测思路是,通过参考更长时间内数据的总体走势进行长期环比,在短期范围出现频繁操作则被认定为非正常现象,存在入侵行为,对其进行异常检测符合网络安全的预期。入侵者短时间内的创建、修改和删除光路操作会对走势造成干扰。大部分情况下都是使用曲线对序列进行拟合,这样可以清晰呈现变化趋势。若新出现的数据破坏了这种序列的走势,曲线因此不再平滑,即说明该部分出现了异常。

对曲线进行拟合有很多方式,比如滑动平均和回归等。本文使用指数权重移动平均算法来拟合曲线。在 EWMA 算法中,下一点的平均值是由上一点的平均值和目前点的实际值修正而来。而对任意一个 EWMA 值来说,数据的权重是不相同的,更靠近当前值的数据有着更高的权重。得到了平均值之后,就可以基于公式判定之后出现的新数据是否在设定的阈值范围之内,即通过与实际值进行比较,判定新数据是否超出了该范围以决定是否判断为入侵。

EWMA 算法的表达式为

$$v_t = \beta v_{t-1} + (1-\beta)\theta_t \ ,$$

式中:vt 为 t 时刻 的 EWMA 值;系数 β 为加权下降的速率,该系数变小意味着下降的速度变快,在数学中一般会以作为一个临界值,小于该值的加权系数值不作考虑;θ t 为时刻 t 的流量实际数值，此时表达式应为

$$v_t = (1-\beta)(\theta_t + \beta\theta_{t-1} + \cdots + \beta^{t-1}\theta_1)$$

3 遗传算法在机器学习中的应用

众所周知，算法+数据结构=程序设计。而如今大火的人工智能是不可能离开数据结构和算法这种基础底层的东西，而当机器学习发展到一定程度时又可以反过来推进算法的证明与实现。

遗传算法（Genetic Algorithm）是机器学习算法中的基础，也是机器学习的核心。它是一种通过模拟自然界种群进化的随机搜索算法。遗传算法具有良好的全局搜索能力，而且有较强的内在并行性，可以方便地进行分布式计算，加快求解速度。但是遗传算法也存在不足，其局部搜索能力较差，导致单纯的遗传算法比较费时，在进化后期搜索效率较低。

爬山算法是一种局部择优的方法，采用启发式方法，是对深度优先搜索的一种改进，它利用反馈信息帮助生成解的决策，属于人工智能算法的一种。该方法的局部搜索能力很强，但外卖配送路径优化问题是一种多约束问题，由于爬山算法缺乏对问题的可行空间的全局性采样，所以较容易陷入局部最优解。根据遗传算法和爬山算法各自的不足，把两个算法结合起来，借鉴郎茂祥等[10]学者的改进方法，在传统的遗传算法操作步骤中加入爬山操作，来提高算法的局部搜索能力和收敛速度。算法策略的如下。

（1）编码方法：采用自然数的编码方式，用节点的编号进行编码。

（2）适应度函数：把目标函数作为适应度评价函数，本文的 52 目标函数是极小化问题，因此取其倒数用作适应度函数。

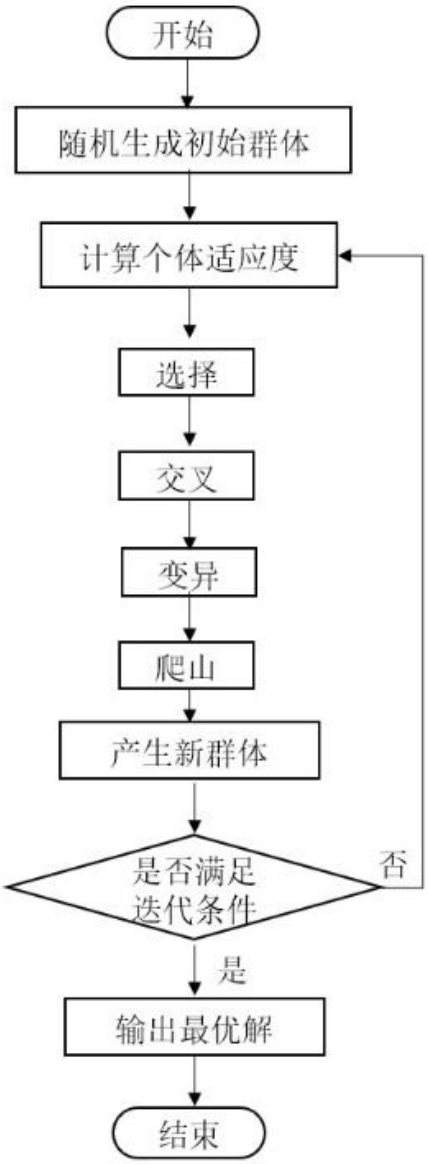
（3）选择策略：将个体的适应度从大到小排序，选择适应度大的个体和适应度虽然小，但是幸存下来的个体作为父代。

（4）交叉算子：采用 PMX 的交叉方法，操作方法为：首先，随机地在一对染色体中选择两个杂交点，如 P1=0854|7316|92，P2=0546|9132|87，再交换交叉片段，得到 P1'=0854|9132|92，P2'=0546|7316|87，最后按照映射关系（7-9,1-3,2-6）进行替换得到子代 P1''=085913276，P2''=0542713689。

（5）变异算子：采用交换基因的方法，随机产生变异的基因位置，对个体的基因进行交换。

（6）爬山算子：采用基因换位算子来实现爬山操作,具体操作为：随机选择染色体中的两个位置，交换该位置上的基因，判断基因交换后的个体适应值是否增加，若增加，则以交换后的个体取代原个体。

（7）终止条件：运算到指定迭代次数则终止。



4. 结论

本文通过教育、安全、算法这三个角度，多维度地阐述了机器学习的领域之广而内涵之深。这门科学也是“前不见古人，后不见来者”。但也正是如此，才吸引了古今中外这么多的学者为止绝倒，但对于机器学习真正领域的挖掘还远不止如此，或许未来的某一天我们能够解开机器学习的真正面纱，又或者只是这条路上一个默默无闻的脚印。但无论如何，毫无疑问的是它一定会造福整个世界，它的内涵更是值得每一个计算机领域的人切磋琢磨。

参考文献：

【1】张恒. 面向产学融合的机器学习新型课程建设探索.[EB/OL].2020.11.25.
https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDAUTO&filename=KJFT202033020&v=etkx4O6VD34mg3jOucsIlV5eg%25mmd2F5IT9seXNeMxWIkUEQOaJH1KxnmZ3L5NwbVP%25mmd2F1r.

【2】朱嘉豪 徐凯 王炎豪 陆煜斌 宣涵 沈建华. 基于机器学习的软件定义光网络入侵检测策略.[EB/OL].2020.11.25.
https://kns.cnki.net/KXReader/Detail?TIMESTAMP=637418919467435000&DBCODE=CJFD&TABLEName=CJFDAUTO&FileName=GTXY202006009&RESULT=1&SIGN=ONY8ml1bWsAxSOhidd%2fFt8g4DMI%3d

【1】朱桐, 江欢. 基于遗传算法的外卖配送路径优化研究.[EB/OL].2020.11.25.
https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDAUTODAY&filename=GXQG202012024&v=NQdT%25mmd2BMVqyvC%25mmd2BwOKLS3so%25mmd2BGw31f4yt%25mmd2BDhueScYCpcaxQHVFRhUtQMYzQX5jvOxLpi.