大数据时代的思考

3160103829 王宇琪

就像马云所说,"21世纪的竞争是数据的竞争,谁掌握数据,谁就掌握未来。" 在这个全球互联的时代,大数据为我们提供了一个前所未有的观测世界的角度。从蒸 汽时代、电气时代到现在的信息时代,随着互联网和计算机技术的发展,越来越多的 数据被产生、被记录、被挖掘,如果说唐太宗"以史为镜,可以明得失"聚焦的是现 在,那么大数据"以历史为经验,去预测未来"聚焦的就是未来。

第一个想探讨的是大数据与小数据。什么是大数据?这个大该如何理解?在 1981年,比尔盖茨认为 640KB 的内存就足够大了,而在当今,640GB 的硬盘可能都不能满足人们的需求。所以我认为这个大是一种相对的概念,它并没有明确的界限,而是有着深深的时代烙印。在当今,智能感知、智能交通、智能城市,TB、PB 级的数据,这是前人未曾想到过的计算与规模。那么数据真的是越多越好吗?有了更多的数据就能保证预测出完美的结果?从辩证的角度,这个大也是有利有弊,利在于丰富的数据提供了全方面的分析角度,具有更大的潜力得到更加完美的结果;弊在于信息的冗余、错误,可能会使你舍本逐末,偏离正确的道路。因此在数据的江湖之中,虽说大数据凭借巨大的潜力横空出世,大有一统江湖之风,但大数据与小数据的相辅相成才能相映生辉。因为数据本无大小之界限,它取决于运用数据的问题与立场,我认为以应用为导向,在适当的问题选取适当的方法才是正确之道。

不同于人文的文以载道,以文化人,科技的发展,则是一腔改变世界的热血。我 想这也是近来大数据、人工智能、区块链、物联网热门的根源,它们都赋予了未来一 个美好的愿景。大数据会带来哪些变革呢?第一是思维的变革,传统角度去解决一个 问题,更多的是从数理层面去简化、去假设、去构建抽象模型,而数据时代赋予了一 种新的数据驱动的角度,通过历史数据的训练、学习去优化模型。这并不意味着数据 驱动就一定优于数理模型,在很多工业场合,机理模型仍旧处于主导的地位,只是 说,大数据提供了一种新的思维角度,也有人尝试机理与数据的混合驱动,取己之 长,补其之短,达到更好的效果。第二是商业的变革,随着淘宝、头条、网易云音乐 等社交软件的信息挖掘,数据不再是陈旧的历史记录,而在很多用途上可以获取商业 价值。商家通过数据分析用户的喜好,提供更适合的商品推荐。公司领导在进行决策 时,除了使用他敏锐的个人直觉,也更多地倾向于数据的分析与预测。有名的例子就 是 Netflix 对大数据的采集挖掘,进而产生出《纸牌屋》这部让观众欲罢不能的美剧。 不仅仅是预测,航海、航天、矿业、农业等相关的数据本身,就具有巨大的经济价 值。最后是生活层面的变革,大数据结合医疗,优化了诊断与识别; 大数据结合工 业,完善故障预警与因果分析:大数据结合交通,构建无人驾驶与智能交通.....对于学 生的我们,更明显的就是原有课名加上大数据后,一下子变成了炙手可热的课程,可 见大数据在当下的火热程度,大数据工程师、分析师的岗位也一下子层出不穷,备受 热捧。热潮之下,更该做的是理性的探索与发展,技术结合应用方显其优势,而非一 味追逐技术的虚名。

大数据的发展绝非异军突起,而是与机器学习的算法相结合、与分布式存储、云计算相结合,共同引领一场数据革命。从机器学习的角度,可以分为有监督学习与无监督学习,在做有监督学习的分类、回归问题时,我们希望探究变量之间的因果关系,而在大数据的视野下,我们不再追求理性的因果关系,转而关注事物的相关关系。这也反映出真实世界的纷繁复杂性,很多问题都难以用严格的因果关系去约束,反而是许许多多的相关关系共同作用得到的。在无监督学习中,大数据的全面性则为特征的提取提供了更权威的选择。数据的全面性,也使得算法的泛化能力得到提高。

大数据推动的深度学习,近年来取得了许多不错的进展。随着数据量的爆炸,存储服务器成为系统性能的瓶颈,大数据也促进了分布式的存储与计算的发展。并行计算的发展也应运而生,GPU CUDA 加速、AI 芯片、类脑芯片等名词也频频出现,不断提高着人们处理信息、分析这个大数据的能力。可以说,大数据促进了许多研究方向的共同发展与进步。

按尤瓦尔·赫拉利在《今日简史》对大数据的观点,他认为大数据已经彻底颠覆了人们的生活。搜索、导航、知乎、谷歌,短短不到 20 年,人们极大地丰富了信息,也极大地提高了收集、寻找信息的效率。我觉得其中一个很有趣的观点是,当人们的信息、用户画像不断完善,它可以学习你掌握的知识、你的兴趣与爱好、风格与习惯,是否可以打造一个虚拟的智能去替代你的存在。同时数据的爆炸,也使人更加麻木、茫然,反而使教育者反思人的哪些能力是尤为重要的?批判性思考、沟通、合作、创意。最后想以《今日简史》的开篇总结:在一个信息爆炸却多半无用的世界,清晰的见解就成了一种力量。