

AhnLab 安全月刊

2017.01 Vol. 50
Machine Learning

利用“机器学习”的安全威胁检测

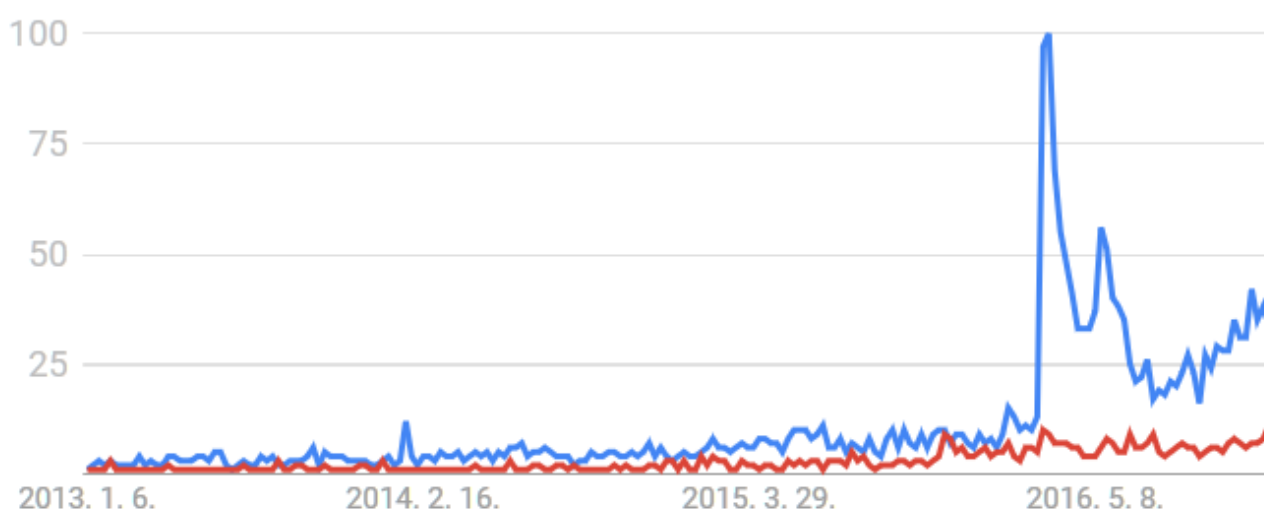
什么是机器学习(Machine Learning) ?

去年Google的AlphaGo和李世石9段的围棋人机大战告诉人们人工智能不再是遥远的未来。此外，IT行业的巨头企业正在将“人工智能(Artificial Intelligence)”作为未来的发展动力，并在各个领域展示了应用人工智能的事例。

本刊将分为两次介绍实现人工智能的方法中的一个“机器学习(Machine Learning)”和安全领域为什么需要机器学习。本期将介绍什么是机器学习。

从几年前开始，不仅是IT专门新闻，就从一般的电视新闻也很容易接触到“人工智能”相关的各种新闻报道。去年，AlphaGo和李世石的围棋对决，引发了全世界的关注，号称世纪大战。今年1月2日~4日，网名为“Master”的神秘围棋手连续战胜全世界顶尖高手，连胜60场没有败绩。此后，Google承认“Master”就是人工智能AlphaGo。当人工智能AlphaGo打败李世石，然后又战胜全球诸多围棋泰斗，连胜60场没有败绩之后，“终结者(Terminator)”或“矩阵(Matrix)”之类的遥远的科幻似乎变近了，不少人有了一种“细思极恐”的感觉，既期待着人工智能为我们的文明带来跨越式发展，又担心这些比人类还聪明的家伙真有一天会把我们玩弄于股掌之间，甚至把我们灭绝¹。

【图 1】显示的是从2013年1月到2016年12月的“人工智能”和“机器学习”关键词在Google的搜索次数及变化趋势。



【图 1】Google Trends分析的2013年1月~2016年12月的“人工智能”和“机器学习”关键词的趋势

¹ <http://tech.sina.com.cn/i/2017-01-05/doc-ifxzkfuh5472600.shtml?qq-pf-to=pcqq.c2c>

仅在2005年，人工智能在韩国也不是一个受欢迎的领域。这或许是受到1980年末人工智能面临的第二次寒冬的影响，但可能是受到了韩国的以硬件中心的工业环境的影响。但不同于韩国，国外的几家新兴IT企业则不断扩大人工智能方面的研究。其中一个就是Google，其中心有PageRank人工智能(或算法)。

Google之后的IT公司经过了Web2.0范例，人工智能自然被认为是技术优势及与竞争者的区别，并认为是与企业生存直接相关的一种技术。所有的数据被处理和共享，从这些数据中获取尽可能多的知识(knowledge)的企业可以获得优胜奖杯。即，企业必须努力获得尽可能多的数据，不仅是“处理和共享”容易，最终能够将其转换为知识。在过去，不管未来会发生什么事情，先是将数据积累，以后再说。但是现在，这些数据让人们的头开始痛起来。

如前所述，西欧的IT企业对人工智能的应用和研究已经进行了几十年，如“买尿布的人同时购买啤酒”、“将搜索到结果以客户最满意的顺序排序”的应用，还有“昨天已观看电影‘终结者’，那么今天将提供15%折扣的‘矩阵’”等。而韩国IT企业，最近才意识到人工智能的重要性并开始重视该领域，但力不从心。硬件工业只要有任性和努力就可以做到，但是人工智能没有那么容易。

机器学习

人工智能领域有多个，其中，最近较为有名的是AlphaGo背后的团队“Deepmind”。此外还有许多领域，如启发式、优化、信息检索和模式识别等。最近的算法趋势很难明确地区分其领域。本文着重介绍一般的“机器学习”。

气温	雨	风	出去玩？
5		●	否
10			是
8			否
6	●		否
24		●	是
30			是
33			否
32			否
30	●		否
15	●		否
16		●	是

【表 1】根据气温的户外活动情况

假设我们的生活具有一定的模式，而笔者根据气温的生活模式如【表 1】所示。下雨天、非常冷的天气、非常热的天气，笔者都不会出去玩。根据上述模式，笔者自己建立了下面如【表 2】的规则。

```

(气温 < 10) 或 (温度 >= 31)
  ↳ 不出去玩
(气温 < 31)
  ↳ (下雨)
    ↳ 不出去玩
  ↳ (不下雨)
    ↳ 出去玩
  
```

[表 2] 根据气温的户外活动规则定义

如果提前建立这种规则，仅根据明天的天气预报就可以决定明天是否出去玩，而不用苦恼。假设上述【表 1】的数据是一般人口的生活模式，那么当具有IT知识的炸鸡店老板获得【表 2】，可将该规则利用在促销活动。例如，可以预测户外公园的订购量，在公园分发传单来引导订单，继而增加销售。

但是，如果【表 1】的数据量从几十到几百万件或更多，人照样可以建立模式呢？可能不仅非常困难，而且很难验证。基于如【表 1】的任意数据，通过机器即计算机建立如【表 2】的规则的过程可以说明为“机器学习”。请参考【图 2】。



【图 2】机器学习的主要过程

【图2】中，橙色箭头表示利用数据执行机器学习的过程。即，如要执行机器学习，就需要数据，并且通过机器(计算机)来处理，其结果生成规则(模型)。该过程需要经过无数的反复和实验，是一个消耗很长时间的这个过程。规则图标的两个小箭头则表示应用学习的过程，即通过生成的规则(模型)导出相应的信息。与先前执行机器学习的过程相比，该过程相对简单。

AlphaGo在与李世石对决之前已根据许多围棋棋谱，经过长时间的学习结果，获得了规则(模型)。在与李世石对战时，简单输入当前的棋盘布局，可在限制时间内计算出获胜概率最大的走法。

有关机器学习的介绍到此结束，下面来了解以下为执行机器学习时所需要检查项目有哪些？

机器学习，什么是重要？

求“出去玩”答“是”的个数

求“出去玩”答“否”的个数

如果答“是”多的话，始终判断为“是”

如果答“否”多的话，始终判断为“否”

【表 3】ZeroR算法

Google的DeepMind由大量的配置复杂的计算机构成，但相比之下【表 3】的ZeroR算法相当简单。也就是说，仅一台计算机也可以充分计算。

再举一个例子，如果我们具有100年的沙特拉伯天气数据，该数据很庞大，大概有几十太字节(TB)。假设根据该数据学习“明天是否下雪？”后建立了知识，那么回答总会是“否”。因为该地区100年没有下过雪。为了求“明天是否下雪”的答案，需要花费庞大的预算来构建最新且昂贵的机器学习算法的环境，然后想像一下未来五个星期天天等待的场景。然而，其结果可能与ZeroR算法的结果没有什么区别。因为回答始终是“否”。此外，ZeroR算法利用最近普通的计算机一台就可以快速回答。无论如何，沙特阿拉伯的“明天是否下雪”问题并不是那么难，而且常理来讲人们可能并不想知道。相反，预测沙特阿拉伯明天的最高气温，以应对酷暑更会有意义。

如此，为了执行有意义的机器学习，首先选择好“问题”是非常重要的。记住，“雪”在沙特阿拉伯是没有意义的。另外，机器学习的结果“沿袭”已知的知识也是没有意义的。假设通过机器学习的得到了“30多岁的男性在周末的晚上主要在家里享受啤酒和动作片”的结果，并且为了得到这个结果而花了很多费用。这不是很荒唐可笑吗？或许你会问，“谁会支付昂贵的费用来做这么无聊的事情？”然而，有相当多的没有意义的昂贵的机器学习项目已进行或正在进行中，这只是一个比喻罢了。通过前面介绍的ZeroR算法和稍微改进版本OneR算法来可以确认这种机器学习是否有意义。这是机器学习的基准算法。如果与当前正在研究的算法结果没有多大差异，则表示“问题”没有意义，或者不需要使用机器学习来完成。

不管怎样，如果“问题”已定义好，然后需要查看数据。需要检查的项目如下。

- 您是该数据的所有者吗？那么权限范围呢？
- 您对数据领域的知识有多深？
- 您对数据的噪声(Noise)知道多少？
- 该数据有多大？
- 该数据是否具有周期性，可以划分？如果不是，那么从头到脚处理为一个？
- 该数据的趋势(正态分布)是否频繁变化？
- 该数据的少数具有意义，而不是多数？

[表 4] 机器学习“问题”的检查项目

前面举例的沙特阿拉伯的情况，如何确保相应的天气数据是重要的。如果我们将温度计和湿度计等传感器发布到沙特阿拉伯后，亲自搜集天气数据的话，我们是该数据的所有者并拥有所有权限。但是如果从weather.com等服务提供商收集到RSS feed或JSON(JavaScript Object Notation)形式的数据，那么对该数据既没有所有权也没有权限。

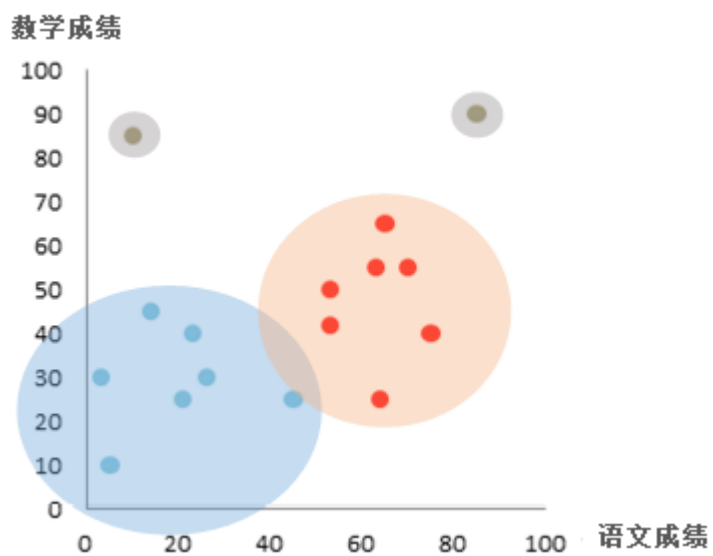
如果对数据没有主权，难以期望日后通过更改数据属性和形象的改进，因此对数据的拥有主权是重要的。例如，在沙特阿拉伯的天气数据中仅有风向信息也可以期待预测准确性的提高。但是，如果没有对数据的主权而无法添加风向数据，就是说要放弃提高预测准确性。

另外，即使拥有对数据的主权，如果研究机器学习项目的人缺乏对该数据领域知识，那将会是一个很不容易的工作。如果对医学信息数据执行机器学习的专家没有具备医学知识，结果将会是有限的。例如，关于肝指数的数据可以表现为实数(0.0以上)，或以适当的值为基准表示阳性(+)/阴性(-)。在机器学习算法中，对数据表示方式-整数和如+/-之类的二进制值(boolean)以不同的方式处理。因此，如果熟悉医学信息，可以将该数值变换为如‘+/-’的二进制来尝试学习。这个可以使机器学习更加多样化，并提供导出更好结果的机会。

此外，还有几个检查项目都是在执行机器学习中优先考虑的项目。尤其是在数据的趋势(正态分布)频繁发生变化的环境中，可能对机器学习的结果导致很大影响，因此需要提前调查。例如，假设国民收入数据由于国家发生了非常大的灾难，国民收入突然减少到三分之一，即数据的正态分布发生了很大变化。那么，如果将灾难前学习的结果应用于灾难后的数据，是否合适？即使世界上存在有多好的机器学习算法，都可能无法提前计算灾难。

作为例子提到的国民收入数据和灾难，这种趋势变化比我们想象的还要非常多样地发生，机器学习的准确性由于这种变化，随着时间的推移将越来越下降。因此，提前检查这些趋势变化的可能性很重要，并预先做好如何响应。

最后是“多数/少数”的问题。过去，将一般人分为几个分类后进行营销活动。而现在，需要将其分为更细分的类别来应对。在这里为什么会突然谈到“营销”一词，是因为在机器学习中与“Outstanding”一词有关。这一词意味着不是多数而是少数的一部分组，而机器学习算法试图忽视“Outstanding”来减少误差。如果我们所要的知识需要对总客户信息中前0.1%的VIP客户执行机器学习来寻求，那么就是说需要其他的方法。因为机器学习算法是减少Outstanding，并将数据一般化。这样的话，前0.1%的VIP客户数据将被忽视。这类似于下面【图 3】。机器学习算法将蓝色和红色两个分类区域分成“规则”或“公式”。然而，上面两个灰色区域是Outstanding区域，以阻碍分类性能而处于被忽视或删除的命运。即，数学成绩特别好的学生或者数学和语文成绩都很好的学生的数据将被处理为“例外”而不包含在学习。



【图 3】语文、数学成绩分布图

到目前为止本刊介绍了什么是机器学习，并了解到为执行具有意义的机器学习有哪些检查事项。在下一期中，我们将继续介绍为什么需要机器学习来防御安全威胁。

AhnLab 安全月刊

<http://cn.ahnlab.com>

<http://global.ahnlab.com>

<http://www.ahnlab.com>

关于AhnLab

Ahnlab的尖端技术和服务不断满足当今世界日新月异的安全需求，确保我们客户的业务连续性，并致力于为客户提供一个安全的计算环境。我们提供全方位的安全阵容，包括经过证实的适用于桌面和服务器的世界级防病毒产品、移动安全产品、网上交易安全产品、网络安全设备以及咨询服务。AhnLab已经牢固地确立了其市场地位，其销售伙伴遍布全球许多国家和地区。

AhnLab

北京市朝阳区望京阜通东大街1号望京SOHO塔2 B座 220502室 | 上海市闵行区万源路2158号18幢泓毅大厦1201室

电话：+86 10 8260 0932（北京） / +86 21 6095 6780（上海） | cn.sales@ahnlab.com

© 2017 AhnLab, Inc. All rights reserved.