人工智能 数据挖掘 机器学习 算法工程师 时间序列分析

关注者 865 被浏览 25,572

机器学习门下，有哪些在工业界应用较多，前景较好的小方向？

举个例子：有人做推荐系统，他们可以到各种商城、新闻类App做推荐的工作；有人做语音识别，他们可以去科大讯飞之类的和语音相关的公司工作；有人做人脸识别，...显示全部

关注问题 写回答 添加评论 分享 邀请回答 ···

9 个回答 默认排序

阿萨姆
机器学习/集成学习/异常检测

390 人赞同了该回答

推荐一个尚未受到足够重视但潜力很大的方向：**异常检测(anomaly detection)**，也叫异常分析(outlier analysis)，相关的还有novelty detection。

异常检测在工业上有非常广泛的应用场景：

- 金融业：从海量数据中找到“欺诈案例”，如信用卡反诈骗，识别虚假信贷
- 网络安全：从流量数据中找到“侵入者”，识别新的网络入侵模式
- 在线零售：从交易数据中发现“恶意买家”，比如恶意刷评等
- 生物基因：从生物数据中检测“病变”或“突变”

换句话来说，异常检测**就是从茫茫数据中找到那些“长得不一样”的数据。**但检测异常过程一般都比较复杂，而且实际情况下数据一般都没有标签(label)，我们并不知道哪些数据是异常点，所以一般很难直接用简单的监督学习。异常值检测还有很多困难，如极端的类别不平衡、多样的异常表达形式、复杂的异常原因分析等。

从人才供给上来看，专门研究或者应用异常检测的人才是非常有限的。而且大部分人往往都更青睐于传统互联网科技公司，留给银行和零售业的可用之人并不多。因此，**已经身处某个行业的朋友们很适合了解学习异常检测，从而弥补所属领域对于异常检测人才的需求。**

1. 应用场景与前景

像文章开头提到的，异常检测的主要应用场景是风险控制(risk control)，常见于金融机构、保险机构、银行等。以我的个人体会为例，各大银行都在扩充自己的数据分析团队，尝试用机器学习手段来降低如银行卡盗刷的案例。而且值得关注的是，大部分银行的风控手段往往都还有很大的升级空间，十月份的时候我和加拿大最大的银行之一的机器智能(machine intelligence)主管交流时，他告诉我他们的部门总共才7个人，最大的困难就是找不到合适的人，即缺少懂得用机器学习来做风控的又愿意加入银行的人。

换个角度来看，**对于银行和普通金融机构来说，最大的挑战是很难吸引科技人才。**大部分科技人才都还是选择加入互联网公司，比如国内的BAT或者国外的FLAG。

我也曾给另一个跨国保险公司做过诈骗识别的项目。他们所使用的风控软件叫做NetReveal，花费数百万美元，但误差率高达百分之90。换句话说，100个识别出的欺诈中只有不到10个是真的诈骗，浪费了大量的人力物力。在引入了机器学习的异常检测后，我们大幅度降低了误差率。

拿银行和保险行业的例子是为了说明**这个方向缺口很大，但相关人才很少，有符合技能的人才又往往不愿意委身于此。**因此，**异常检测在风控中的前景非常光明，属于为数不多机器学习能够落地的方向。**

2. 相关技术

异常检测可以通过监督学习或者非监督学习来做，但往往最终还是需要非监督学习。以反欺诈为例，大部分时候我们根本不知道什么是欺诈，什么不是。诈骗的定义往往是很模糊。往小了说，反诈骗似乎是一个二分类问题(binary classification)，但细想后会发现如果把每种不同诈骗当做单独的类型的话，其实这是多分类问题(multi-class classification)。而单一类型的诈骗几乎是不存在的，且诈骗的手段日新月新总在变化。因此即使拥有历史数据，我们也很难分辨不同种类的诈骗。

退一步说，即使我们真的有诈骗的历史数据，即在有标签的情况下用监督学习，也存在很大的风险。**用这样的历史数据学出的模型只能检测曾经出现过与历史诈骗相似的诈骗，而对于变种的诈骗和从未见过的诈骗，我们的模型将会无能为力。因此，在实际情况中，一般不建议直接用任何监督学习，至少不能单纯依靠一个监督学习模型来奢求检测到所有的诈骗。**除此之外，欺诈检测一般还面临以下问题：

> 1. 九成九的情况数据是没有标签(label)的，各种成熟的监督学习(supervised learning)没有用武之地。
> 2. 区分噪音(noise)和异常点(anomaly)时难度很大，甚至需要发挥一点点想象力和直觉。
> 3. 紧接着上一点，当多种诈骗数据混合在一起，区分不同的诈骗类型更难。根本原因还是因为我们并不了解每一种诈骗定义。

一般来看，我们把异常检测的技术包括：

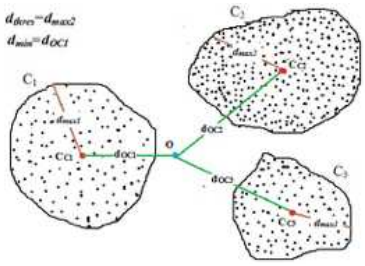
1. 建立在统计学意义上的检测方法：

- 极值分析(extreme value analysis)。这样的方法往往仅对单独维度进行研究，使用上有很大的局限性。
- 对数据分布进行假设，如对异常数据和正常数据进行不同的分布假设，并用EM算法拟合数据。这样的方法局限性在于假设往往和实际有较大出入，效果一般。

2. 基于线性分析的检测方法，特指在低维度上分析数据间相关性的方法。这样的方法包括维度压缩如PCA，Factor Analysis等。这类方法的问题在于把数据压缩后或者找到低维嵌入后，数据的可解释性下降，我们很难解释为什么异常是异常。

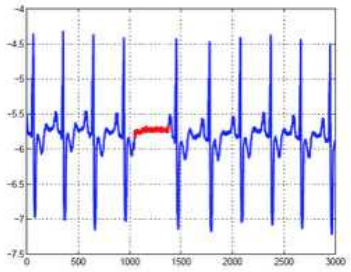
3. 基于时空上的异常检测，特指异常和其所处的环境有关：

- 空间关系造成的异常：

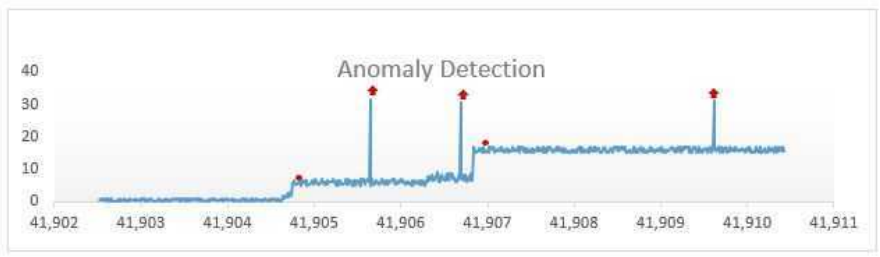


O点在单独来看的情况下是正常点，但考虑到临近点后就是异常点 [1]

- 时间序列上的异常：



红色部分单独来看不是异常，但考虑到临近点后就是异常点 [2]



时间序列上的突然上升或者下降都可能是异常点 [3]

4. 建立在相似性分析

- 建立在距离度量的上的异常检测(distance based)，如K-近邻为原型的也可归为此类
- 建立在密度分析上的异常检测，如经典的 Local outlier factor(LOF)

5. 其他各种异常检测方法，包括：

- 集成异常检测(outlier ensemble)：代表性的算法有isolation forest，feature bagging
- 监督异常检测，半监督异常检测，主动学习(active learning)
- 图中的异常检测，也包括网络中的异常检测

3. 学习路径推荐

虽然异常检测有非常广阔的应用场景，但据我所知还没有一门公开课或者中文书籍系统的讨论相关的问题。以英文材料为例，比较权威的是Charu Aggarwal的Outlier Analysis [4]，本文也多处参考了这本书的内容。

我自己觉得比较恰当的学习路径是：

- 掌握基础的、通用的机器学习知识，如周志华《机器学习》中前半部分的基础知识点
- 了解一些统计学的知识也有所帮助，因为最基本的异常检测是建立在统计学检验上的
- 学习时间序列分析也大有帮助，很多工业界模型都无法逃离“时间轴”
- 作为入门，可以阅读一下SIAM关于异常检测的教程(siam.org/meetings/sdm10...)
- 如果可能的话，建议系统学习上文提到Outlier Analysis这本教科书
- 进阶进行论文阅读的话，大部分研究都发表在数据挖掘会议上，主要包括KDD，ICDM，SIAM Data Mining，传统的机器学习会议不多

根据评论区朋友的补充，提供一些其他参考资料：

- @syzyzs 补充Udemy有一门相关课程：Outlier Detection Algorithms in Data Mining and Data Science
- @卡牌大师 补充“数据挖掘导论”中有一部分关于异常检测的综述，Sklearn的文档中也有讨论异常检测 (2.7. Novelty and Outlier Detection)
- @TyrionW 补充了一门相关的数据安全课程：CS259D: Data Mining for Cyber Security 以及 智能运维中涉及到的异常检测 (基于机器学习的智能运维_图文_百度文库)

从入门了解的角度，也欢迎大家参考我的知乎文章：

- 阿萨姆：反欺诈(Fraud Detection)中所用到的机器学习模型有哪些？
- 用Pyador进行『异常检测』：开发无监督机器学习工具库(一)

4. 总结

个人认为，**异常检测在工业应用上大有可为，是为数不多的有良好应用场景且人才缺口较大的领域。同时，因为大家对于互联网科技公司的向往，短时间内人才缺口很难被科班生补上，跨专业的朋友也有得天独厚的优势。**

但值得注意的是，作为一个小领域，甚至是一个没那么火的领域，相关的资料不多，且不成体系。而且资料往往是英文，需要很强的自学能力。**不难想象，自学难度以及学习曲线都非常陡峭。**

开玩笑的说，富贵险中求，对于技术发展要有我们自己的判断。在全民深度学习的时代，不妨了解一下这些“遗珠”，说不定它会成为你未来很多年的依身傍命之技。

---

[1] Mira, A., Bhattacharyya, D.K. and Saharia, S., 2012. RODHA: robust outlier detection using hybrid approach. *American Journal of Intelligent Systems*, 2(5), pp.129-140.

[2] researchmining.blogspot.ca...

[3] Anomaly Detection – Using Machine Learning to Detect Abnormalities in Time Series Data

[4] Aggarwal, C.C., 2016. *Outlier Analysis*. Springer.

编辑于 2017-12-22

芯片（集成电路）话题的优秀回答者

2 人赞同了该回答

华尔街都是顶尖高手，除非你从小奥数都是全国名次，否则也就是打下手的份儿。

简单说：

小区物流，就是简单的人工智能，2-3年会全国铺开。

儿童教育定制，每个人类似闯关游戏，机器改卷子，机器排行榜。

工地安全监控、工作违规监控。图像识别，自动标记。

各种房屋、衣服、园艺、广告等设计工作，一旦机器自动生成，人挑选，简单易行，国际水平。

工厂生产流程监控和规划，很快效率提高N%。

高危人群标记：这个对社会影响更大

想象你有无穷个普通人的观察和思考能力，可以0成本帮手，每个行业都会大步提高效率。

发布于 2017-12-23

▲ 2 ▼ 添加评论 分享 收藏 感谢

C Li

我的妈我怎么啥都不会！

8 人赞同了该回答

Prognostics and Health Management

上面那个同学说的“”预警设备故障”应该就是指剩余寿命预测(Remaining Useful Life)，另外还有故障诊断等都会用到机器学习方法。

题主对时序信号有兴趣的话，剩余寿命预测就是一个时间序列的问题，可以用LSTM等模型。

建议题主看看去年KDD Workshop13:
Machine Learning for Prognostics and Health Management

其中有篇用LSTM Encoder Decoder预测剩余寿命的

编辑于 2017-03-22

▲ 8 ▼ 9 条评论 分享 收藏 感谢

许家瑞

卡内基梅隆 NLP

8 人赞同了该回答

model calibration

有许多机器学习模型输出的并不是概率，而是score。而这个score你可以直接转换成标签作为输出，但在业界很多时候是需要概率输出的，比如信贷预测、投注、保险。calibration就是试图将score转化成概率。

先占个坑，以后补。

发布于 2017-12-22

▲ 8 ▼ 7 条评论 分享 收藏 感谢

许豌豆

4 人赞同了该回答

我看到过的两个方向：

1 . 使用机器学习来提

2 . 使用机器学习来优化产品配方

发布于 2017-03-14

BigQuant
人工智能助力宽客玩转量化投资，使用AI开发量化策略。微信公众号:BigQuant

10 人赞同了该回答

关于AI的具体运用，Google：《Machine learning 101》

量化金融：链接

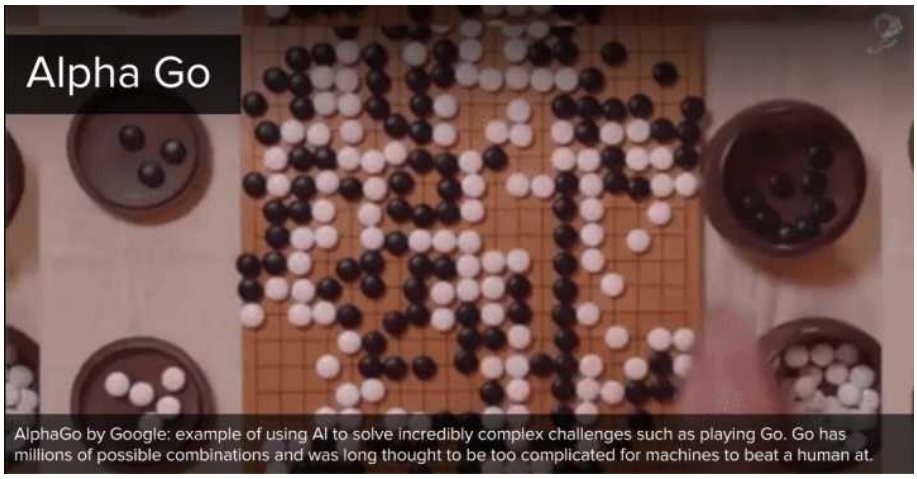
机器人：

图片风格转换

视频生成

实物翻译

棋类



发布于 2017-12-22

▲ 10 ▼ 添加评论 分享 收藏 感谢 收起

郑天一
数据分析 IT民工

1 人赞同了该回答

MIT Sloan Sports Analytics Conference

编辑于 2017-03-16

▲ 1 ▼ 1 条评论 分享 收藏 感谢

阿达

石油系统：

1、地质：地震勘探数据分析 ，比如断层、圈闭的自动识别

2、采油：油井产量预测，故障诊断

3、钻井：钻井过程中事故复杂预防。比如钻具事故、地层异常压力预报等

不过现在油价太低，石油行业投资萎缩，深度学习是否真正能降低生产成本还很难说。企业投入这方面可能有顾虑

编辑于 2017-12-23

▲ 0 ▼ 添加评论 分享 收藏 感谢

知乎用户
数据分析师、咨询师

挺难的，太窄了

发布于 2017-03-27

▲ 0 ▼ 添加评论 分享 收藏 感谢

写回答