数据库系统概论新技术篇

大数据与机器学习

Part II：大数据机器学习

卢志武

中国人民大学信息学院

2017年4月

目录

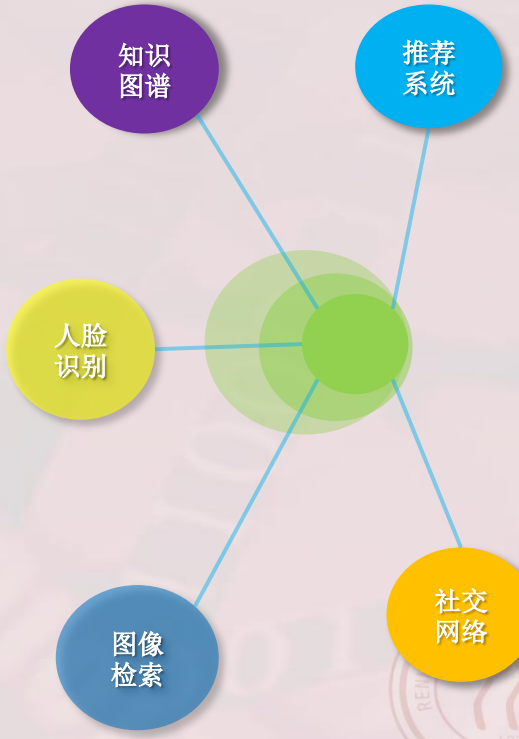
- 大数据机器学习的基本概念
- 大数据机器学习的实现平台
- 大数据机器学习的总结与反思

大数据给机器学习的挑战

- 数据源多样化，能否自动地从每种数据源中提取特征？
- 数据量非常大，算法的运行效率能否满足实际需求？
- 数据分布会发生变化，学习模型的假设是否还成立？

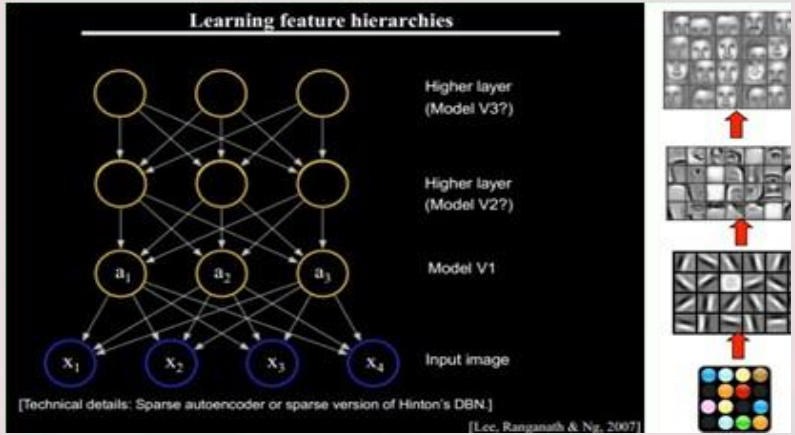
大数据机器学习的特点

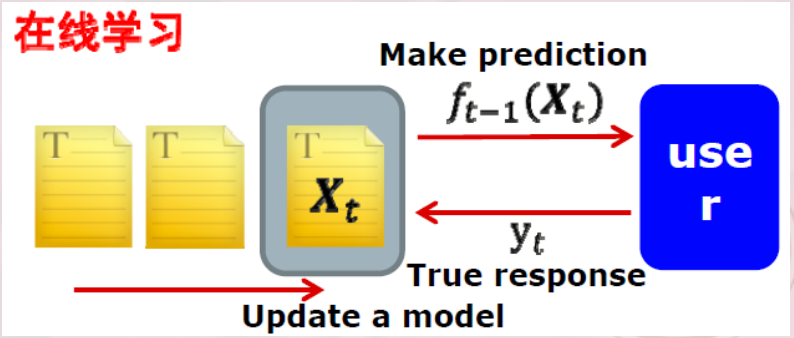
- 各种技术的融合
  - 单一技术难以满足大数据分析多样性需求。
  - 多种技术融合，有助于提高系统稳定性。
- 数据理解是难点
  - 海量数据高度非结构化，数据的快速准确理解成为关键。
  - 深度学习和在线学习在大数据理解领域显得尤为重要。
- 分类会逐渐弱化
  - 快速检索是有效利用海量数据的前提。
  - 大数据时代，分类问题会逐渐弱化，而检索则会变得更重要。



大数据机器学习的关键技术

- 深度学习
  - 很好地利用GPU等高性能计算设备。
  - 解决多源特征的自动提取问题。
- 在线学习
  - 解决传统算法更新模型代价过大的问题。
  - 有效地应对大数据分布发生变化的难题。
- 近似近邻搜索
  - 使得在大数据上检索成为可能。
  - 常用方法：哈希索引和基于树的索引。
- 并行计算
  - GPU并行计算：CUDA
  - 分布式系统：Spark





目录

- 大数据机器学习的基本概念
- 大数据机器学习的实现平台
- 大数据机器学习的总结与反思

常用的大数据机器学习平台

大数据机器学习平台简介

- **Hadoop**是一个由Apache基金会所开发的分布式并行计算框架。该框架最核心的设计是：MapReduce算法和分布式文件系统HDFS。
- **Skytree**是Skytree公司开发的机器学习平台，结合先进的机器学习算法，为企业提供大数据高级分析。已用于推荐系统、预测分析、市场细分等。
- **Spark**是UC Berkeley AMP lab所开源的通用并行计算框架，基于MapReduce算法实现，拥有Hadoop MapReduce所具有的优点，但比Hadoop更通用，迭代算法效率更高。
- **GraphLab**是CMU的Select实验室提出的一个面向大规模机器学习/图计算的分布式内存计算框架，构建了四种流行的机器学习算法的并行版本。

目录

- 大数据机器学习的基本概念
- 大数据机器学习的实现平台
- 大数据机器学习的总结与反思

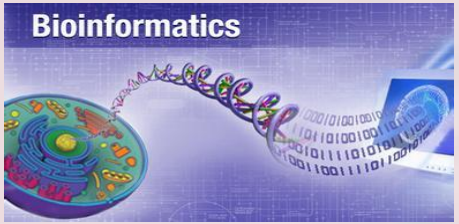
大数据机器学习无处不在

互联网搜索

美国大选预测

网络舆情监控

医疗辅助诊断



生物信息学

汽车自动驾驶

大数据机器学习的缺陷

- 大数据可能存在严重的偏差，机器学习的结果令人不太满意。
- 大数据虽很大，但是总存在例外的情况。
- 大数据机器学习的代价过大，而人从小样本中就能很好地学习。

搜狐 首页 用户名/邮箱/手机号 登录 注册 我的搜狐 邮件

美国大选是不寻常的选举年意外“倒戈”损特朗普

人人聚财金 2016-10-18 22:04:24 经济学家 阅读(374) 评论(0)

声明：本文由入驻搜狐公众平台的作者撰写，除搜狐官方账号外，观点仅代表作者本人，不代表搜狐立场。 举报

距离美国总统大选还有三周，在因秽语视频及性侵指控而选情告急之时，共和党总统提名人特朗普把对媒体的攻讦提升到了史无前例的高度。上周六，在佛罗里达西棕榈滩一次集会上，他说媒体正串通希拉里合谋操纵美国大选：

“希拉里动用的最强力武器是企业媒体、出版社。让我们说清楚一件事：在我们国家媒体不再关乎新闻。他们是跟任何说客或金融机构无异的特殊政治利益团体，有着整套的政治议程——其服务对象不是你们，而是他们自己。他们不计成本、不计代价，也不在乎多少人的生活会被毁坏，目标只有一个，把奸诈的希拉里选上台。对他们来说，这是场战争；对他们来说，根本没有出界一说。我们国家正在为生存而挣扎。相信我。”

凤凰 科技 凤凰网科技 > IT业界 > 正文

第一起自动驾驶死亡车祸！特斯拉事故复现分析

2016年07月04日 08:06

来源：36氪 作者：刘健茹 281人参与 53评论

小数据机器学习

- 更关注机器学习模型的可解释性。
- 更关注开放环境下的机器学习，即机器学习模型的稳定性。
- 更关注规则与知识融入到机器学习中。

- 主讲人：卢志武
- Email：luzhiwu@ruc.edu.cn
- 电话：010-62515670