



机器学习框架

姜霄棠

January 12, 2015



1 背景与目的

- 量化投资调研
- 机器学习算法的性能问题
- 自动编程实践

2 机器学习相关算法

- 数据清洗
- 特征/成分提取
- 抽象层算法

3 开发任务与计划

- 基本时间序列预测功能实现
- 深度学习算法开发
- GPU 算法加速

股票价格预测与决策

- 对股票价格趋势进行时间序列分析
- 预测走势，进行决策

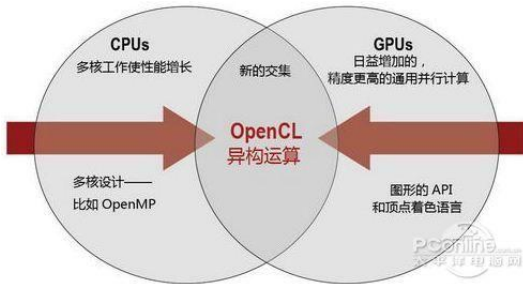


机器学习算法性能现状

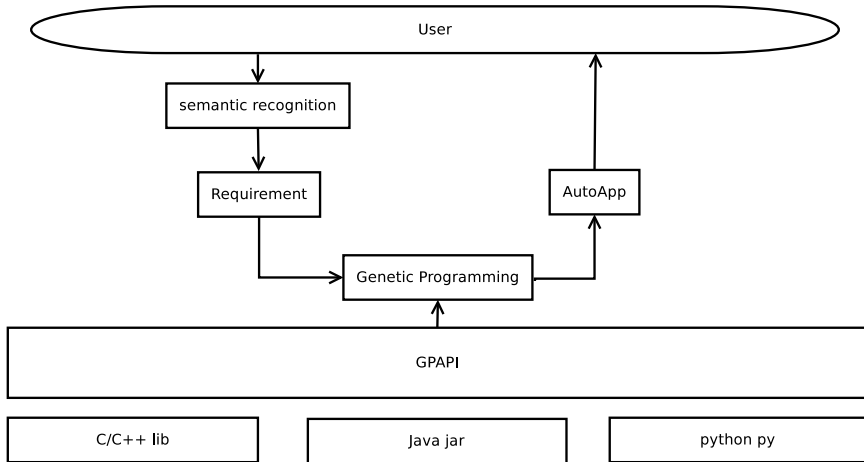
- 目前机器学习算法主要在服务器或学术研究中使用，其性能很少受到关注。
- 大数据下的机器学习算法目前一般依赖 **scale out**，如果有易于扩展且性能完美的设计，机器的运算效果可以提升 10 倍，进而减少功耗。
- 移动终端后面会越来越多地使用机器学习算法，目前普通使用的机器学习库不能满足性能功耗的要求。

异构计算标准化

- SIMD 技术和 GPU 加速技术可使程序性能大幅上升。
- 不同的机器上，由于硬件的不同，这些技术要么需要参数调优，要么完全不可用。



自动编程框架





目前 GP 框架的用法

- 编写根函数代码，规定关联结构，最终函数可由根函数按一个 xml 表拼装而成。
- 由 GP 框架学习 xml 结构，调整参数。
- 学习出来的 xml 即 Auto Program



时间序列

- AR 模型
- 相空间重构



图像相关

暂不用关注

- 压缩感知
- SIFT



流形学习

- 主成分分析（PCA）
- 多维尺度分析



数据转换

- 小波变换
- 标准化
- 核函数映射



预测：分类

- 决策树
- 随机森林（组合算法）
- AdaBoost
- KNN
- SVM（组合算法）



预测：回归

- 最小二乘法
- 逻辑回归
- LU 分解
- CART（组合算法）

定式优化算法

- 梯度下降/随机梯度下降
- 线性规划解法（单纯形法，内点法等）
- 二次规划解法（SMO 算法）

深度学习

深度学习是独立于以上算法的另一个分支。

- 自动编码器
- 稀疏编码
- 限制波尔兹曼机 RBM 与深信度网络
- 卷积神经网络 CNN

启发式优化算法

这部分由 GP 框架层实现，不属于机器学习库的内容。

- 遗传算法
- 粒子群算法
- 模拟退火算法



需要实现的算法

计划 2 月中旬完成。

- 决策树
- 随机森林
- SVM（已经完成）
- AR 模型回归（已经完成）
- PCA
- 小波变换（可选）



股票预测模型

计划 2 月底完成

- 数据提取
- 价格预测
- 执行器
- 执行器评估



深度学习算法

计划 3 月份完成

- 数据收集
- 算法开发
- 效果验证



GPU 算法加速

计划 4 月份完成

- PCA
- SVM
- 深度学习
- 线性回归