股票机器人

Vincentyao

15-08-09

具体做什么？

- 基于历史与当前数据，运算机器学习算法，做出对股市的预测。
  - 预测股市大盘的涨跌
  - 分析各个板块的变化趋势
  - 推荐可能盈利的股票组合

背景

- 有效市场假说（The Efficient Market Hypothesis）
  - 有效市场：如果在一个证券市场中，价格完全反映了所有可以获得的信息
  - 在任何时候，单个股票的市场价格都反映了已经发生的和尚未发生、但市场预期会发生的事情。
- 随机游走假说（The Random Walk Hypothesis）
  - 股票市场价格的变动是随机的
- 但市场真的不可预测吗？

国外的参考方案

- 基于Twitter，分析公众情绪，预测DJIA涨跌[1]。
  - 准确率87.6%
- 基于其他市场的股票和商品价格变化，预测NASDAC趋势[2]。
  - 准确率74.4%，年化30%
- www.forecasts.org/stock-index/index.htm 预测stock index

基本方法

- Fundamental Analysis：
  - 关注 股票所属的公司，譬如公司营收情况与预期
  - 多采用 传统有监督分析方法
- Technical Analysis
  - 关注 股票的历史价格与历史交易情况
  - 多采用 时间系列分析方法
- Alternative methods
  - 机器学习模型，如Artificial neural network

输入数据源

- 公司相关
  - 公司经营状况，证券分析报告
- 股票相关
  - 之前该股或者同行业股的交易状况
- 其他
  - 最新财经新闻，民众情绪分析(from weibo, 朋友圈等)，其他股票市场行情，商品和外汇价格，相关概念的搜索记录和热度

输入数据源

- Stock
  - NASDAQ, DJIA, S&P 500, Nikkei 225, Hang Seng index, FTSE100, DAX, ASX
- Currency
  - EUR, AUD, JPY, USD
- Commodity
  - Silver, Platinum, Oil, Gold

Feature选择

- 绝对值feature
- 变化率feature

$$
\begin{aligned}
\mathcal{N}(\nabla_\delta x_i(t)) &= \frac{x_i(t) - x_i(t-\delta)}{x_i(t-\delta)} \\
\mathcal{N}(\nabla_\delta X(t)) &= (\mathcal{N}(\nabla_\delta x_1(t)), \cdots, \mathcal{N}(\nabla_\delta x_{16}(t)))^T \\
\mathcal{N}(\nabla_\delta(\mathcal{F})) &= (\mathcal{N}(\nabla_\delta X(\delta+1)), \cdots, \mathcal{N}(\nabla_\delta X(n)))^T
\end{aligned}
$$

- Feature selection

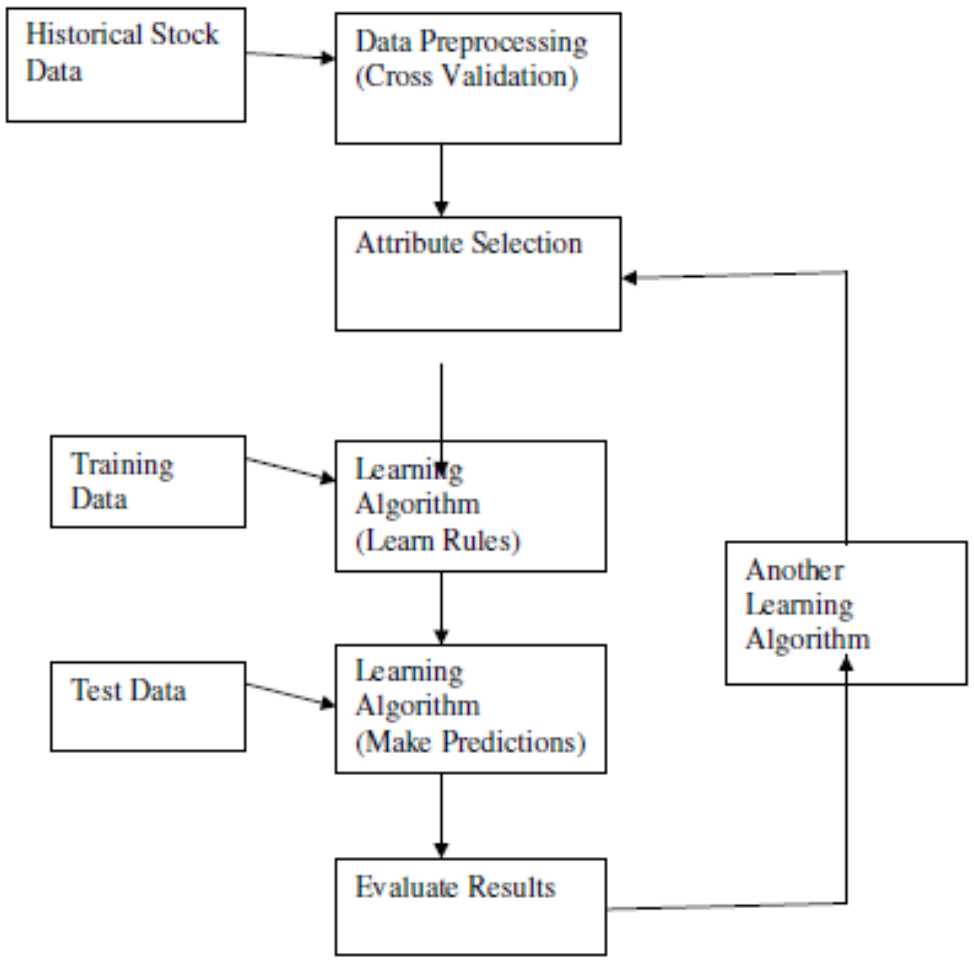
机器学习模型

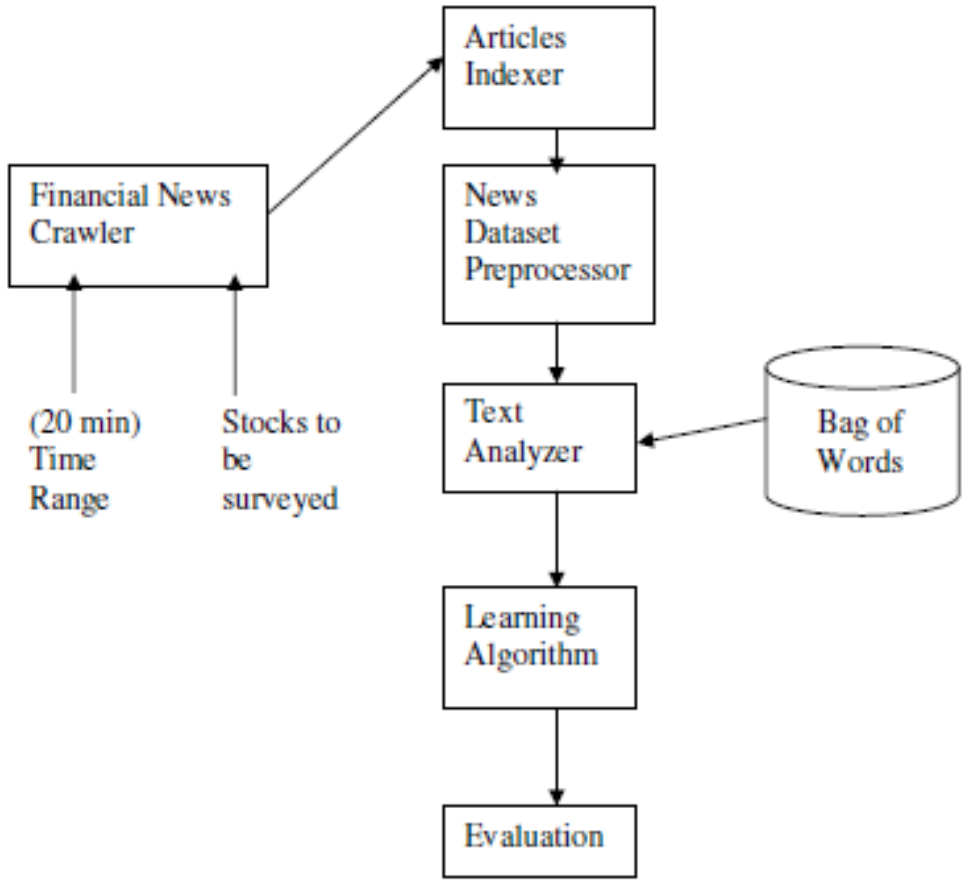
- 有监督模型，预测涨跌以及曲线
  - Decision Stump
  - SVM
  - Logistic/Linear regression
  - GBDT
  - Deep learning

- 时间序列模型
  - HWW
  - CRF
  - ARCH模型
  - RNN/ TDNN

- 文本信息处理模型
  - 概率语言模型
  - PLSA/LDA
  - Information extraction
  - Text categorization

系统框架


Historical Stock Data
Data Preprocessing (Cross Validation)
Attribute Selection
Training Data
Learning Algorithm (Learn Rules)
Test Data
Learning Algorithm (Make Predictions)
Evaluate Results
Another Learning Algorithm


Financial News Crawler
Articles Indexer
News Dataset Preprocessor
(20 min) Time Range
Stocks to be surveyed
Text Analyzer
Bag of Words
Learning Algorithm
Evaluation

技术难点

- 数据源的获取
  - 新闻数据，公众情绪数据，搜索记录数据。大规模爬虫，需要比较强的工程能力。
- 文本信息处理
  - 从繁芜复杂的互联网信息中获取真正有用的信息，需要自然语言处理能力，用于解析网页，提取内容，分析核心词，预测分类。
- 特征分析能力
  - 更多需要金融领域知识，人工经验。
- 机器学习模型
  - 股票市场是一个强非线性的变化过程，需要采用非线性模型，例如深度学习。
- 产品包装与推广
  - 可能的政策风险