基于公众关注度的股票市场的分析与预测

综合论文训练答辩

李雨田

清华大学

2015 年 6 月 18 日

指导老师　余宏亮

报告人　李雨田

学号　2010012193

数据获取及预处理

目的

- 统一数据结构
- 高效率
- 易用

数据获取及预处理

目的

- 统一数据结构
- 高效率
- 易用

数据获取及预处理

目的

- 统一数据结构
- 高效率
- 易用

数据获取及预处理

框架

- 基于 JSON 和 Redis 数据库
- 基于事件循环机制
- 提供 JavaScript 和 Python 库，与 Numpy，Pandas 和 StatsModels 对接

数据获取及预处理

框架

- 基于 JSON 和 Redis 数据库
- 基于事件循环机制
- 提供 JavaScript 和 Python 库，与 Numpy，Pandas 和 StatsModels 对接

数据获取及预处理

框架

- 基于 JSON 和 Redis 数据库
- 基于事件循环机制
- 提供 JavaScript 和 Python 库，与 Numpy，Pandas 和 StatsModels 对接

数据来源

- 挑选上证 50 指数成分股
- 从东方财富网股吧获取用户关注度数据
  - 个股所有讨论帖点击量
  - 对应的时间和内容
- 从国信证券客户端获取历史股票价格和交易量

数据来源

- 挑选上证 50 指数成分股
- 从东方财富网股吧获取用户关注度数据
  - 个股所有讨论帖点击量
  - 对应的时间和内容
- 从国信证券客户端获取历史股票价格和交易量

数据来源

- 挑选上证 50 指数成分股
- 从东方财富网股吧获取用户关注度数据
  - 个股所有讨论帖点击量
  - 对应的时间和内容
- 从国信证券客户端获取历史股票价格和交易量

数据来源

示例

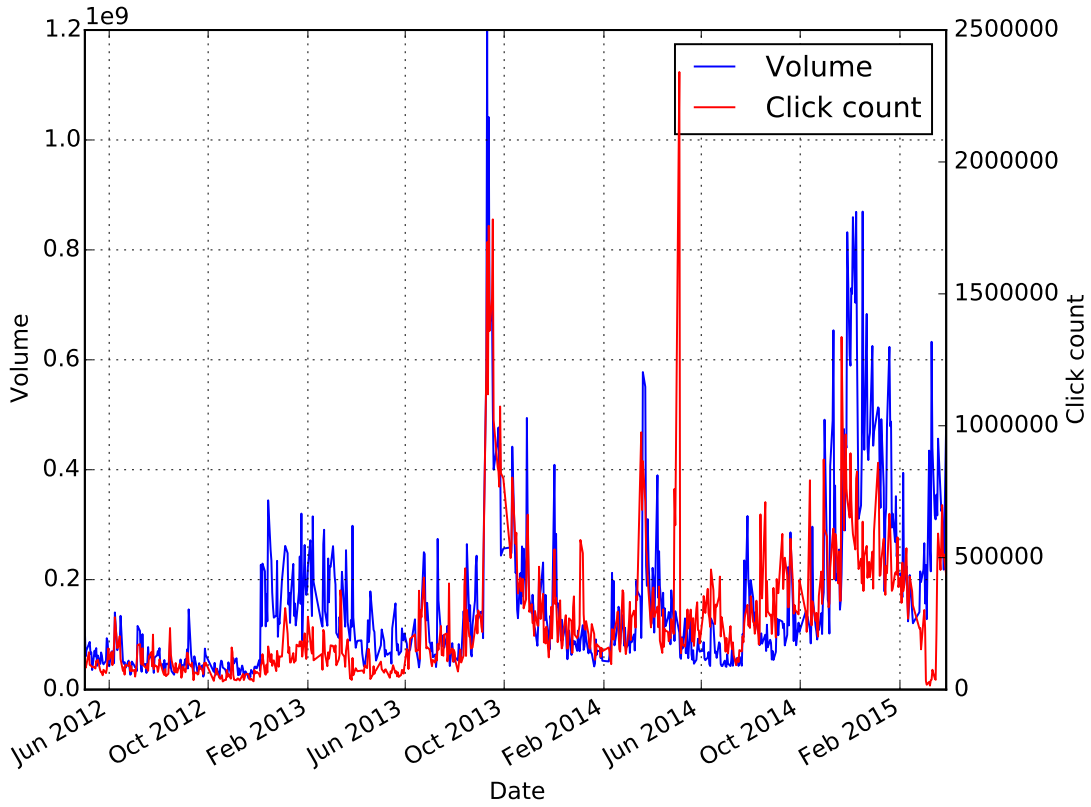

图: 浦发银行 (600000) 成交量与讨论帖点击量关系

数据来源

示例

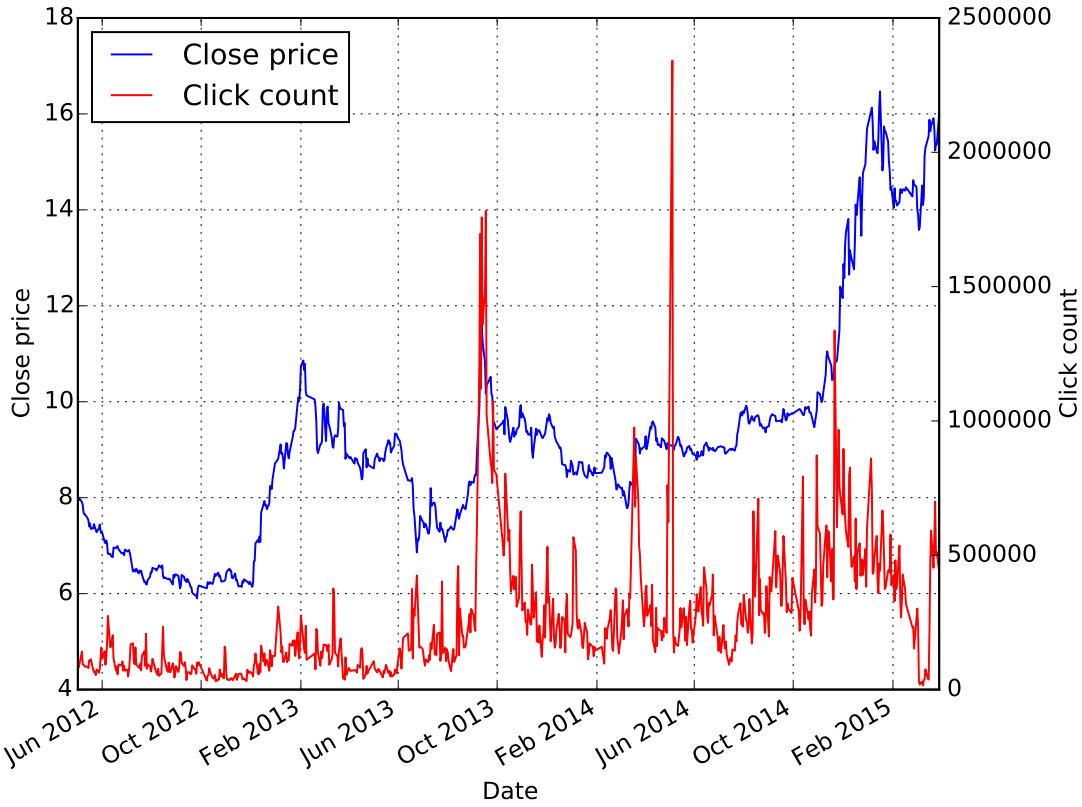

图: 浦发银行 (600000) 价格与讨论帖点击量关系

情感分析

- 拼接讨论帖标题和内容，去除非中文字符，使用翻译 API 翻译成英文
- 使用卷积神经网络分析情感，该模型基于 Twitter 数据训练得到

情感分析

- 拼接讨论帖标题和内容，去除非中文字符，使用翻译 API 翻译成英文
- 使用卷积神经网络分析情感，该模型基于 Twitter 数据训练得到

情感分析

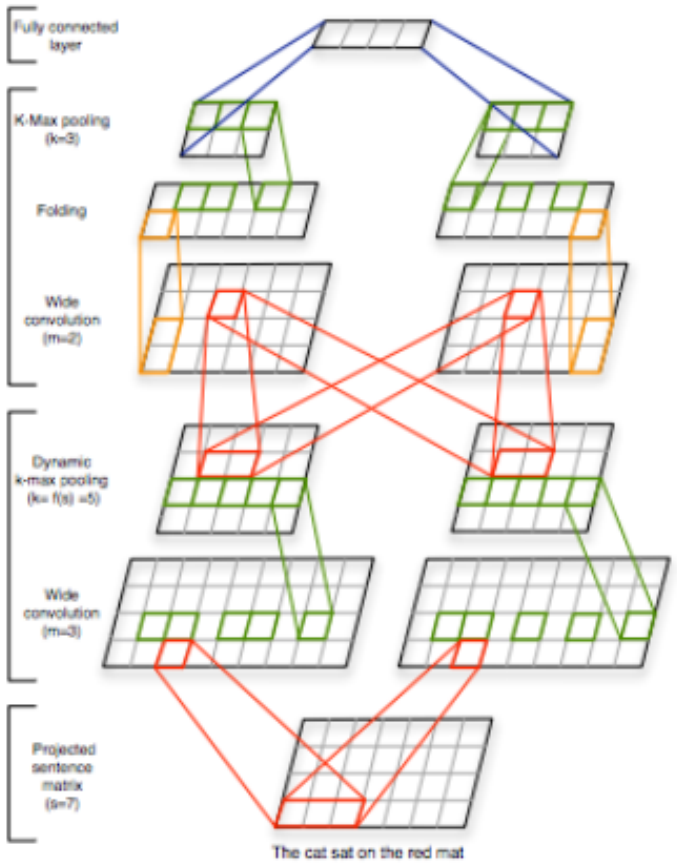

图: 情感分析卷积神经网络

因果关系分析

格兰杰因果关系

假设检验

$$y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + \sum_{i=1}^{m} b_i x_{t-i} + residual_t$$

$$H_0 : b_1 = b_2 = \cdots = b_m = 0$$

因果关系分析

结果

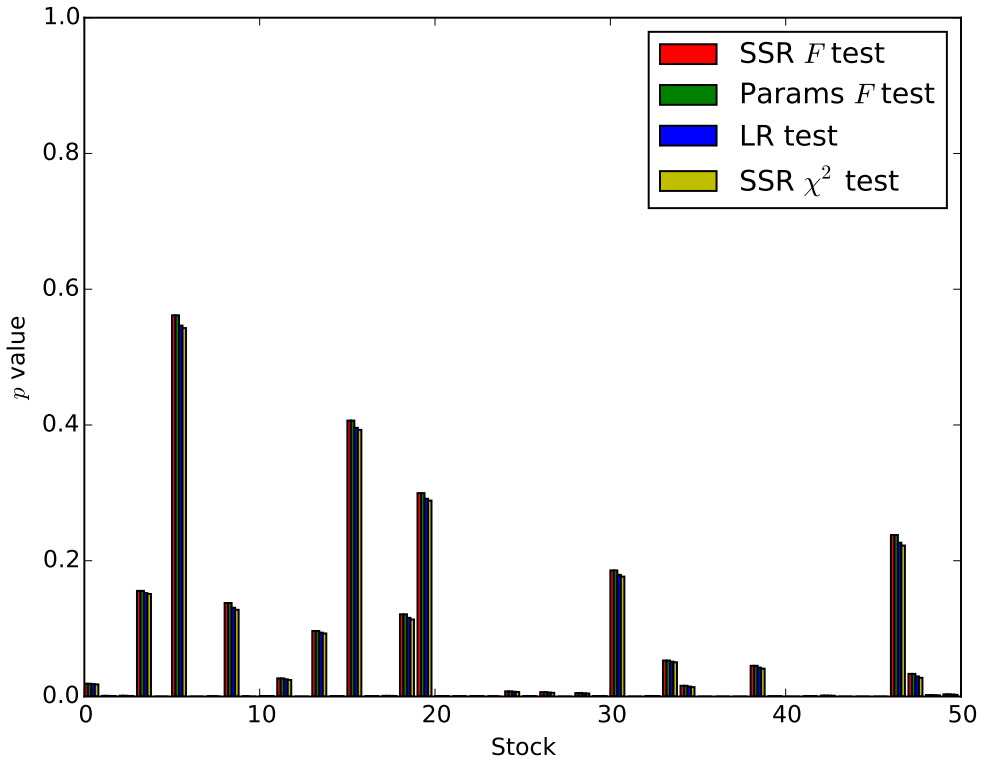

图: 上证 50 成交量与讨论帖点击量格兰杰因果关系检验

因果关系分析

结果

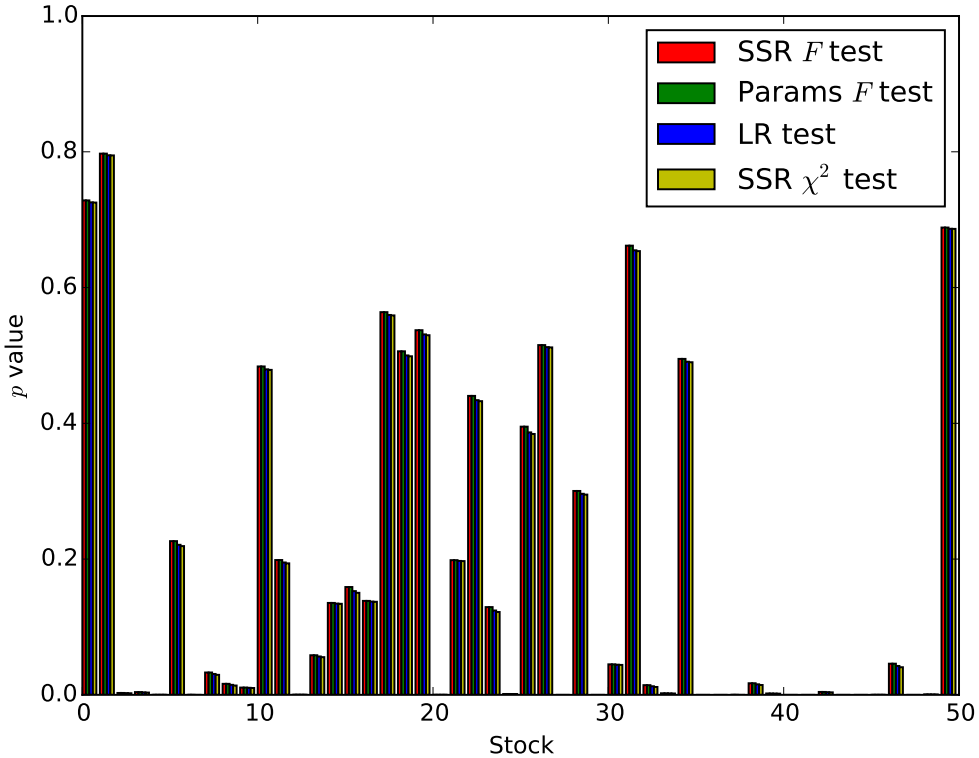

图: 上证 50 价格与讨论帖点击量格兰杰因果关系检验

因果关系分析

结果

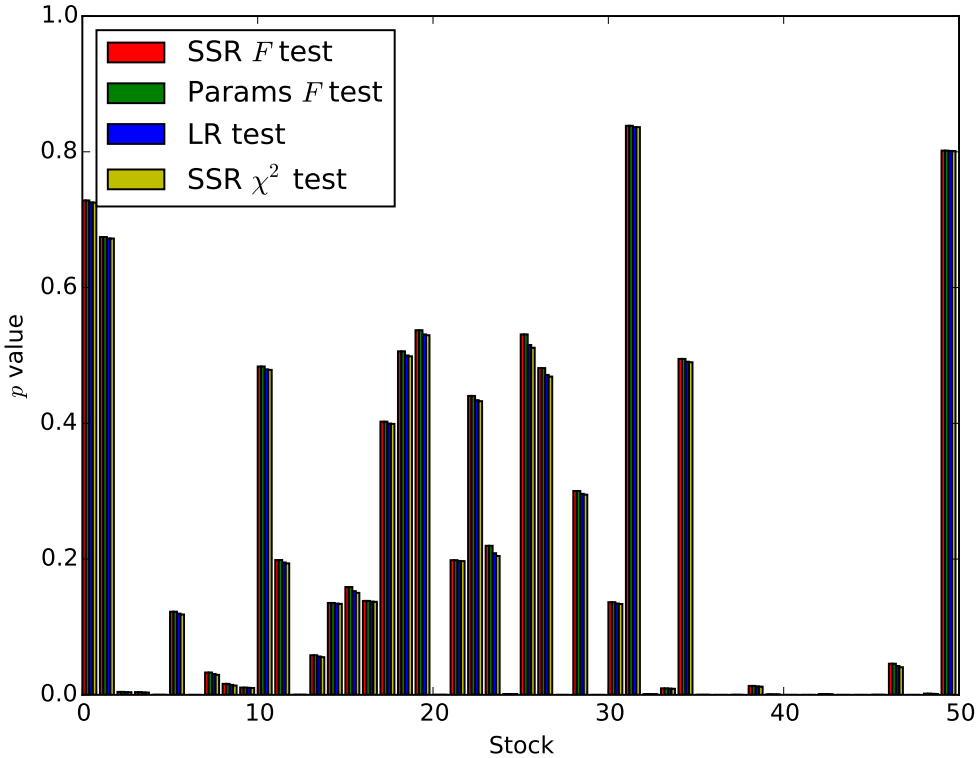

图: 上证 50 价格与积极讨论帖点击量格兰杰因果关系检验

预测模型

向量自回归模型

定义

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t$$

预测模型

框架

- 数据预处理
  - 计算量比
- 落后期选择
  - Akaike information criterion
  - Bayesian information criterion
  - Final prediction error
  - Hannan-Quinn information criterion
- 步长选择

预测模型

框架

- 数据预处理
  - 计算量比
- 落后期选择
  - Akaike information criterion
  - Bayesian information criterion
  - Final prediction error
  - Hannan-Quinn information criterion
- 步长选择

预测模型

框架

- 数据预处理
  - 计算量比
- 落后期选择
  - Akaike information criterion
  - Bayesian information criterion
  - Final prediction error
  - Hannan-Quinn information criterion
- 步长选择

预测模型

框架

- 数据预处理
  - 计算量比
- 落后期选择
  - Akaike information criterion
  - Bayesian information criterion
  - Final prediction error
  - Hannan-Quinn information criterion
- 步长选择

预测模型

框架

- 数据预处理
  - 计算量比
- 落后期选择
  - Akaike information criterion
  - Bayesian information criterion
  - Final prediction error
  - Hannan-Quinn information criterion
- 步长选择

预测模型

框架

- 数据预处理
  - 计算量比
- 落后期选择
  - Akaike information criterion
  - Bayesian information criterion
  - Final prediction error
  - Hannan-Quinn information criterion
- 步长选择

预测模型

框架

- 数据预处理
  - 计算量比
- 落后期选择
  - Akaike information criterion
  - Bayesian information criterion
  - Final prediction error
  - Hannan-Quinn information criterion
- 步长选择

预测模型

框架

- 数据预处理
  - 计算量比
- 落后期选择
  - Akaike information criterion
  - Bayesian information criterion
  - Final prediction error
  - Hannan-Quinn information criterion
- 步长选择

预测模型

结果

定义

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{y_{max} - y_{min}}$$

预测模型

结果

以下结果中有 $NRMSE = 5.857 \times 10^{-2}$ 。

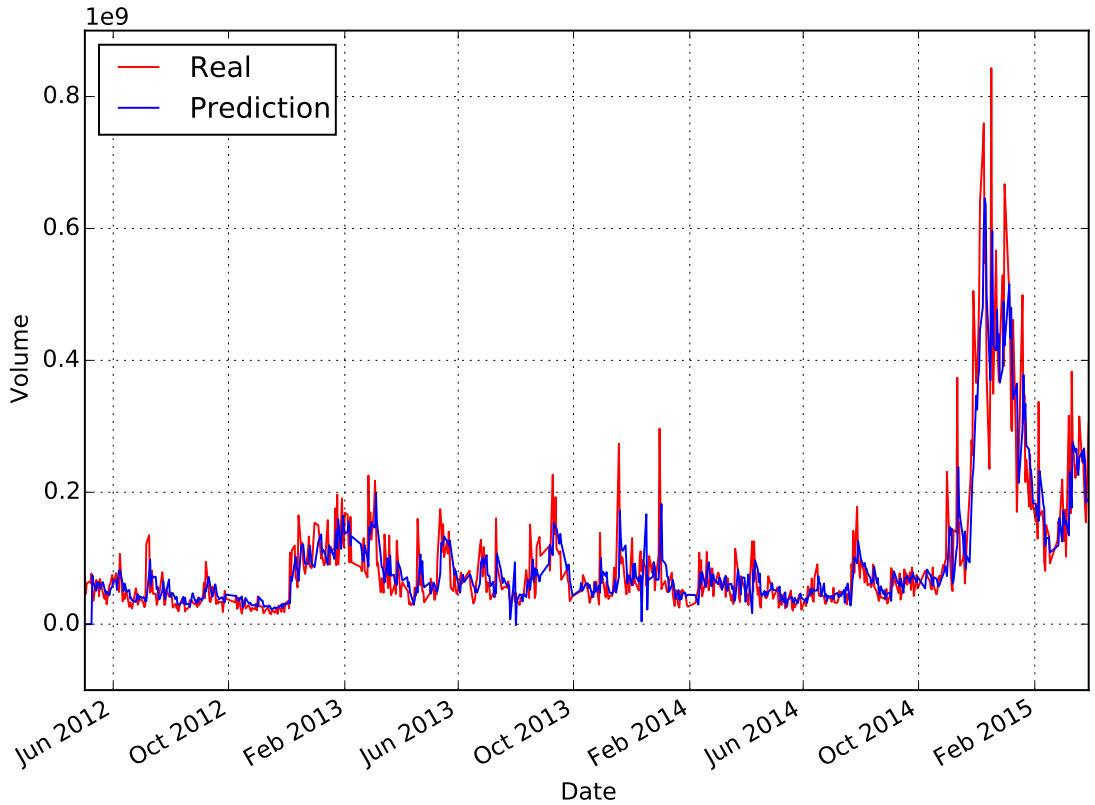

图: 招商银行 (600036) 成交量预测

预测模型

结果

以下结果中有 $NRMSE = 2.546 \times 10^{-2}$ 。

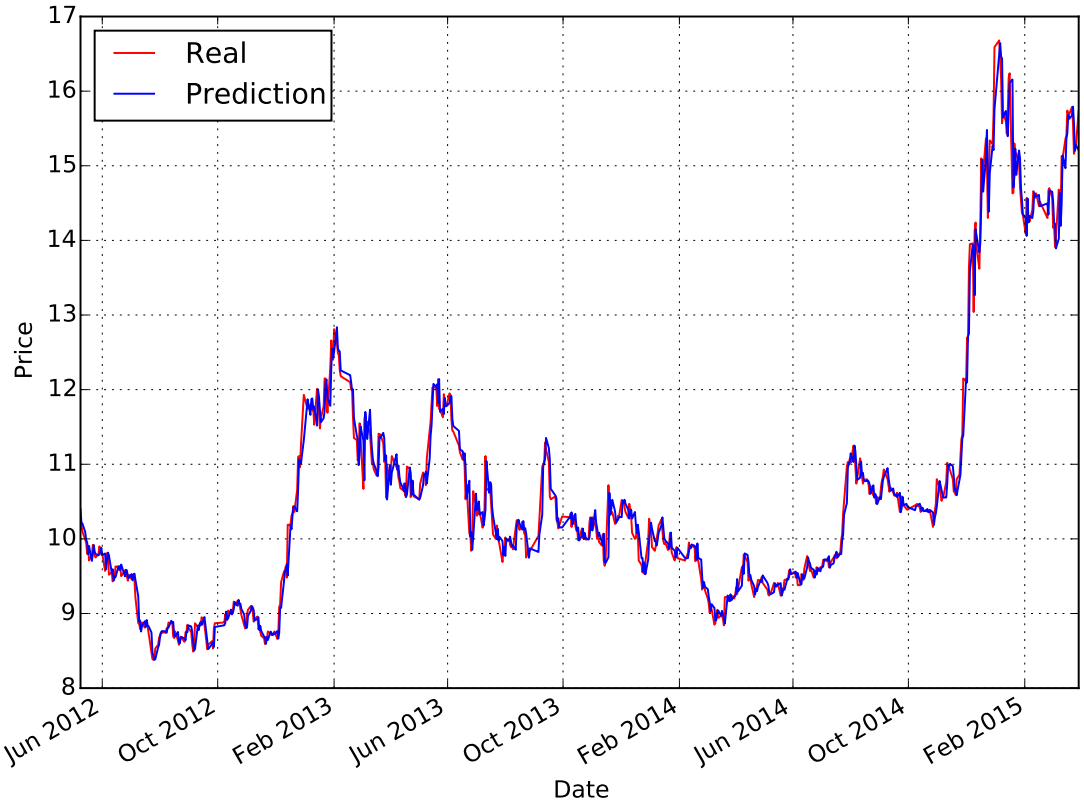

图: 招商银行 (600036) 价格预测