# 基于 SVM 的 IPO 首日投资策略分析①

施 剑

(复旦大学 计算机科学技术学院, 上海 201203)

**摘 要**: 将支持向量机方法应用于新股 IPO 首日价格变动的预测, 预测效果令人满意. 目前的股票价格预测研究都局限于通过已知的时间序列来预测将来的时间序列, 这类模型对于预测没有历史时间序列的新股 IPO 无能为力, 因此基于支持向量机的新股 IPO 价格预测模型对股票价格研究有着重要的参考价值.

**关键词**: 股票价格预测; 机器学习; 支持向量机

## IPO Stock Price Forecasting Using Support Vector Machines

SHI Jian

(School of Computer Science, Fudan University, Shanghai 201203, China)

**Abstract**: We use SVM to predict IPO stock price and achieve good result. There are different kinds of stock price forecasting models. But none of them can predict IPO stocks. So this paper has important reference value in stock price forecasting field.

**Key words**: stock price forecasting; machine learning; SVM

在美国、西欧等比较成熟的证券交易市场, 有着相对完善的监管体系和较理性的市场参与者, 新股 IPO 首日定价较为合理, 溢价和波动都比较小. 反观国内沪深股市, 因新股 IPO 首日没有涨跌幅限制, 经常遭遇暴涨暴跌等价格大幅波动的情况. 如能抓住价格大幅波动的机遇, 就可获得丰厚的收益.

以涪陵榨菜为例, 上市首日开盘价为 25.5 元, 仅上午半天的涨幅就达到近 20%, 下午开盘后仅 6 分钟就因为达到了深圳证券交易所规定的涨跌幅限制而遭到临时停牌. 恢复交易后, 又在短短 15 分钟的时间内达到了 50%的涨幅限制被第二次停牌, 第二次恢复交易后更是在 4 分钟内达到 80%的涨幅限制而被停牌到收盘集合竞价.

虽然涪陵榨菜因上市首日被爆炒而遭遇连续两天的跌停, 但即使以第 4 日的最低价卖出, 也实现了相对上市首日的开盘价 19.6%的收益. 而若在第 4 日以最高价卖出, 则实现了相对上市首日的开盘价 40.8%的收益.

再看另一个例子, 2010 年 8 月 31 日上市的金利科技, 以 24.5 元开盘以后, 便一路上冲至 36.25 元, 最后以 32.2 元报收. 而在之后的两个交易日, 金利科技都以涨停板收盘, 在第 4 个交易日一度上冲至 42.8 元, 也就是相对上市首日开盘价涨 74.7%.

新股波动幅度较大, 没有庄家控盘, 持仓成本透明等因素决定了通过爆炒新股套利的机会长期存在. 本文通过基于 SVM 的预测方法, 让计算机自动判断股票 IPO 首日是否有利可图, 若为正面判断, 则可通过首日买入, 次日卖出的交易获利.

## 1 相关研究

股票价格波动是否可预测、如何预测, 成为近年来研究的焦点问题之一. 国内外对股票价格波动进行预测的方法种类很多, 然而不论哪一种模型, 目前的研究都局限于通过已知的时间序列来预测将来的时间序列, 这类模型对于预测没有历史时间序列的新股IPO无能为力.

目前, 应用较为广泛的股票价格智能预测方法包括小波变换、遗传算法、人工神经网络以及支持向量机算法.

① 收稿时间:2013-03-16;收到修改稿时间:2013-05-17

人工神经网络是非线性、非参数预测技术的主要代表. 神经网络不需要建立所研究的问题本身的精确逻辑和数学模型, 而是模仿人类思维方式训练神经网络的算法. 1991 年, Matsuba[1]将神经网络引入股票市场的价格预测. 2006 年, 林春燕等应用人工神经网络建立股票市场预测模型, 并对东风汽车和邯郸钢铁等个股进行了研究, 结果说明用人工神经网络方法对国内股票市场进行预测是可行的[2]. 但人工神经网络方法存在过拟合问题, 可能产生局部而非全局最优解[3,4], 所以人工神经网络方法在对有强噪声和高维的数据的股票市场进行学习的时候往往受到限制.

2003 年 Kim 讨论了运用 SVM 进行预测的可行性, 并用此方法对韩国的股指进行了预测, 结果表明 SVM 的预测效果明显优于人工神经网络的预测效果[5]. 2005 年, W.Huang 等人用 SVM 对日经 225 指数进行预测[6]. 在国内, 2005 年杨一文等人利用 SVM 对上证综指趋势做了较准确的多步预测; 2006 年, 张晨希等人利用 SVM 对股票进行短期预测[7]. 这些预测实验结果都表明, 支持向量机方法比神经网络方法优越.

然而, 现有研究大多通过学习某一标的物的历史交易数据, 进行其后续时间序列上的单步或多步价格预测, 对于没有历史交易信息的新股 IPO 价格走势, 目前的预测方法则无能为力.

## 2 SVM基本原理

SVM 通过引入核函数来控制分类器容量. SVM 的基本原理是: 将原始特征向量空间中较难实现的分类器, 通过定义适当的函数进行非线性变换, 这一函数称为核函数. 通过这样的变换将原始特征向量空间映射到新的高维空间, 然后对个新的高维空间中求解最优线性分类面, 从而降低分类器复杂度. 当选用特定的核函数时, 可通过核函数在低维特征向量空间中对应特征向量上定义的计算推导出由此核函数导出的高维特征向量空间中的两个特征向量间的点积. 这样高维特征向量空间中的数据处理便能够方便的在低维特征向量空间中进行. 由于通过向量间的点积运算便可求解特征向量机, 故不必担心计算量上的维数灾难, 而着重解决怎样选择适当的核函数的问题, 以改善特征向量在高维特征向量空间中的分布, 从而获得结构简单的分类器. 所以, 求解 SVM 的过程就是在高维特征向量空间中求解样本数据之间最优分类面的过程.

常用的 SVM 核函数有如下三类:

① 线性核函数(linear)

线性核是对线性分类器在线性可分的情况下使用. LibSVM 的作者认为, 当输入特征维数较大时, 宜采用线性核函数, 不需把数据映射到高维空间, 因维数较大时在高维空间中性能没有多大的改善[8].

② 多项式核函数(poly)

Keerthi S 等认为, 多项式核函数在阶数较高时学习能力较好, 在阶数较低时泛化性能较好, 当阶数增加时所训练的时间会急剧增加[9].

③ 高斯径向基核函数(rbf)

Scholkopf B 等认为, 在没有先验知识的情况下, 一般采用径向基核比较好, 空间复杂度小且易于实现[10]. 因此, rbf 核在很多领域得到广泛的应用.

## 3 模型的构建

### 3.1 特征向量选取

新股没有历史交易信息, 在其上市交易前所能获取的数据非常有限, 本文选取其中对新股炒作有一定影响的特征构建特征向量.

① 上市数量.

上市数量直接决定了该股盘子的大小, 也决定了掌控这只股票所需的资金量. 盘子较小的股票如中小板、创业板个股易于被操控, 而控制大盘股则相对困难.

② 发行价与发行市盈率.

市盈率是价值投资者衡量个股估值的标杆, 市盈率的高低决定了一个股票的价格是否偏离它的实际价值. 然而投机者或许并不过于关注市盈率, 而是仅仅通过股票的票面价格来判断该股是否有炒作空间, 如人们往往认为 80 元的股票很贵而 8 元的股票很便宜.

③ 开盘涨幅.

这里所说的开盘涨幅是指 IPO 当天 9: 25 集合竞价结束产生的开盘价相对发行价的涨幅. 该涨幅能反映股民对这只股票的热情程度, 涨幅过低说明该股吸引力不够, 而涨幅过高则可能过于偏离实际价值致使价值回归现象的产生.

④ 集合竞价换手率.

集合竞价换手率反应了庄家对该股的控制愿望, 若集合竞价换手率很低, 则说明没有庄家介入. 若换手率过高, 则说明散户抢盘明显, 或许会影响庄家收集筹码.

本文以 IPO 次日开盘价减去 IPO 当日开盘价结果数值是否大于零作为分类的目标值, 若该结果大于 0, 则分为赢利类, 以数值 1 标记. 若该结果小于或等于 0, 则分为亏损类, 以数值-1 标记. 对于赢利类的股票, 可以采取在 IPO 首日以开盘价买入, IPO 次日以开盘价卖出的套利手段.

### 3.2 核函数及参数 C

依据 Scholkopf B 的相关结论, 结合实际情况, 本文选用了高斯径向基核函数.

参数 C 为平衡系数, 它的作用是在确定的数据子空间中调节置信范围和经验风险的比例, 使其推广性能最好, 不同数据子空间中最优的 C 不同. 在确定的数据子空间中, C 的取值小则表示对经验误差的惩罚小, SVM 有较大的经验风险值和较小的复杂度, 称为“欠学习”现象. 若 C 的取值大, 则 SVM 有较小的经验风险值和较大的复杂度, 称为“过学习”现象. 每个空间存在一个或多个合适的 C 使得 SVM 的推广性能最好. 当 C 大到一定程度时, 支持向量机的复杂度达到最大值, 经验风险和推广性能几乎不再变化.

## 4 实验与分析

为了避免股票在单日内的价格波动影响实验结果, 本文仅比较 IPO 第二日的开盘价是否高于 IPO 首日开盘价, 如遇 IPO 第二日全天跌停的, 则取之后首个非全天跌停的交易日的开盘价. 这就意味着如分类器作出正面判断, 则在 IPO 首日开盘买入, 次日不论涨跌均卖出, 有效规避了主观影响因素.

### 4.1 LIBSVM 简介

本文运用 LIBSVM 软件包进行实验. LIBSVM 软件包是台湾大学林智仁博士等设计开发的一个易于使用、快速有效、操作简单的 SVM 通用软件包, 可以解决 C-SVC、n-SVC 等分类问题、e-SVR、n-SVR 等回归问题以及 one-class-SVM 等分布估计问题, 提供了线性、S 形函数、径向基和多项式四种常用的核函数供选择, 可以有效地解决对不平衡样本加权、交叉验证选择参数、多类问题、多类问题的概率估计等.

### 4.2 实验流程

应用支持向量机进行 IPO 价格预测步骤基本如下:

① 试验样本的选取, 确定训练集数量、检测集数量.

② 标准化.

③ 参数选取.

④ 训练模型.

⑤ 预测.

⑥ 结果分析.

### 4.3 样本选取

本文选取了 07 年 1 月至 09 年 11 月的 246 只 IPO 股票为样本, 其中 07、08 年的 200 只为训练集, 09 年的 46 只为测试集. 数据组织形式如表 1 所示.

表 1 测试数据组织形式

| 股票简称 | 股票代码 | 上市数量 (万股) | 发行价 | 开盘涨幅 | 竞价换手率 | 次日开盘价高于首日 |
|---|---|---|---|---|---|---|
| 中国人寿 | 601628 | 60000 | 18.88 | 95.80% | 6.59% | 1 |
| 广博股份 | 002103 | 3840 | 6.6 | 75.76% | 3% | -1 |
| 恒宝股份 | 002104 | 2304 | 8.43 | 75.56% | 3.47% | -1 |
| 信隆实业 | 002105 | 5440 | 3.4 | 110.30% | 7.00% | 1 |
| 莱宝股份 | 002106 | 3904 | 20 | 72.50% | 0.97% | -1 |
| 沃华医药 | 002107 | 1440 | 10.85 | 72.44% | 2.55% | 1 |
| 沧州明珠 | 002108 | 1440 | 10.16 | 82.58% | 2.65% | -1 |
| 兴化股份 | 002109 | 3200 | 10.8 | 57.40% | 1.79% | -1 |
| 三钢闽光 | 002110 | 8000 | 6 | 58.30% | 3.10% | 1 |
| 威海广泰 | 002111 | 1696 | 8.7 | 256.30% | 6.08% | 1 |

### 4.4 参数选择

我们把数据集分成两部分, 以其中的一部分作为训练集, 以剩下的部分作为测试集. 用不同的参数针对训练集训练分类器, 再用测试集测试分类器分类的准确率. 最后用准确率选用准确率高的参数.

本文选择 Gauss 径向基核函数对新股数据分类: 并且采用 5-折交叉检验评价算法的好坏并且选择最优参数. 即将本文的 200 只股票分为 5 组, 每组样本数量为 40 只. 选择其中 1 组作为训练集, 其余 4 组作为测试集, 如此反复计算正确率. 实际操作中可采用 LibSVM 提供的参数选择工具 grid.py, 具体结果如表 2 所示:

表 2 参数确定

| 参数 C | 参数 g | 交叉检验率 |
|---|---|---|
| 1 | 0.01 | 57.7% |
| 1024 | 0.01 | 68.7% |
| 64 | 0.01 | 73.3% |
| 256 | 0.01 | 71.6% |
| 125 | 0.01 | 80.2% |
| 150 | 0.01 | 78.6% |
| 130 | 0.01 | 83.5% |

从表中可以看出选择 C 为 130, g 为 0.01 时交叉检验率最高, 为 83.5%, 所以在这里选择 C 为 130, g 为 0.01 的 Gauss 径向基核函数.

### 4.5 实验结果

使用参数 C 为 130, g 为 0.01 的 Gauss 径向基核函数支持向量机对 09 年的测试集进行预测, 结果如表 3 所示:

表 3 测试结果

| 测试集年份 | 2009 年 |
|---|---|
| 测试总数 | 46 |
| 正确数 | 37 |
| 正确率 | 80.4% |

虽然这样的正确率并不能算非常高, 但这是一个实用性非常强的结论, 具有很高的可操作性. 挑选预测结果为 IPO 次日开盘价高于 IPO 首日开盘价的股票, 在首日开盘时买入, 次日开盘时卖出, 即可实现套利交易. 如用 100 万元初始资金对 2009 年的实验结果进行模拟交易, 则年末账面资金为 241.2 万元, 即在一年内实现 141.2%的盈利.

## 5 实例验证

为了更好的验证预测效果, 作者于 2010 年 8 月至 2010 年 12 月参加了由财华社集团、深圳证券信息有限公司联合主办的全国性大型股票投资实盘竞技活动——“金太阳”中国股市实盘竞技大赛, 利用基于 SVM 的方法判断操作, 在 40000 多名参赛选手中脱颖而出, 成功挺进总决赛.

在全国准决赛阶段(2010 年 11 月), 依据本系统的判断, 取得了单月近 50%的收益率, 如图 1 所示.


图 1 炒股大赛结果

## 6 结语

本文指出了现有股票价格预测技术在新股 IPO 价格预测方面的局限性, 提出了基于支持向量机的新股首日投资策略方法模型, 并以相关实验证明了该方法的可操作性.

在新股价格预测领域还有一些值得继续深入研究的内容, 例如, 通过回归预测新股价格波动的幅度, 对新股波动不仅能定性判断, 更能定量预测其波动幅度. 另外, 也可结合投资者情绪, 设计一个累进的指标, 模拟新股炒作过程中的跟风效应以及过热后的价值回归, 以提升预测准确度.

## 参考文献

1 Matsuba I. Application of neural network sequential associator to long-term stock price prediction. IJCNN91, Singaporee, 1991: 1196–1202.

2 林春燕,朱东华.基于 Elman 神经网络的股票价格预测研究. 计算机应用,2006(2):476–484.

3 Tsaih R, Hsu Y, Lai CC. Forecasting S&P 500 stock index futures with a hybrid AI system. Decision Support System, 1998, 23: 161–174.

4 Grudnitske G, Osburn L. Forecasting S&P and gold future Prices: An application of neural networks. Futures Markets, 1993, 13(6): 631–643.

(下转第 158 页)

编码类型的一个表单内部, 如:

<h:form enctype="multipart/form-data">

<s:fileUpload id="picture"data="#{register.picture}
"accept="images/png,images/jpg
"contentType="#{register.pictureContentType}"/>

</h:form>

⑨ 来自一个用户的任何输入, 再次显示在一个网站的页面上, 都可能成为一个漏洞的来源, 如果没有恰当地转义和过滤, 任何东西都可以被插入到页面的生成输出中, 例如, 插入引发跨站点脚本攻击的 HTML 限制子集. 但是, 对于要求人性化显示的页面, 如论坛帖子、wiki 页面等, 一般允许输入特殊格式化文本, 并正确回显.

Seam 的<s:formattedText/> JSF 组件, 利用 Seam 的文本解析器能够显示符合 Seam 文本语言的格式化文本, 包括对转义符(\)的支持、用反引号(`)引用代码块, 以及对 HTML 的某些限制子集的支持, 从根本上解决用户在页面输入/显示操作带来的潜在风险. 例如:

<h:inputTextarea id="text"
value="#{selectedBlog.text}"/>

<s:formattedText value="#{selectedBlog.text}"/>

⑩ Seam 组件在使用 EntityManager?API 与数据库进行交互时, 允许你在 EJB-QL 中使用 JSF EL 表达式优化 Java 代码, 例如:

User user=em.createQuery("from User
where username=#{user.username}").getSingleResult();

但是绝不允许:

User user=em.createQuery("from User
where username="+user.getUsername()).getSingleResult
();//易受到 SQL 注入攻击.

## 4 结语

在 Seam 框架开发 Web 应用程序时, 遵守上文提及的原则提升安全性, 能够使它在应用中免受各种常见 Web 应用程序风险的威胁, 从而为整个系统的安全提供更好的保障.

## 参考文献


1 OWASP.OWASP Top 10 for 2013. 2013-03-05 [2013-03-21]. https://www.owasp.org/ index.php/Top_Ten.

2 Samson Kittoli.Seam Reference Documentation. 2012-9-12 [2013-03-12]. http://docs.jboss.org/seam/2.3.0.Final/reference/en-US/html_single/.

3 Oracle and/or its affiliates.The Java EE 6 Tutorial. 2013-01-12 [2013-03-02]. http://docs.oracle.com/javaee/6/tutorial/doc/gijrp.html/.

4 mkyong.JSF 2.0 Tutorials. 2010-10-12[2013-03-02]. http://www.mkyong.com/tutorials/jsf-2-0-tutorials/.

5 Dan Allen.李鹏,韩智.译. Seam in Action.北京:人民邮电出版社,2010.

6 DocForge.Web application/Security. 2012-10-12[2013-02-10] .http://docforge.com/wiki/Web_application/Security.

7 Microsoft.Web 应用程序的基本安全实施策略. 2007-08-20[2013-01-15]. http://technet.microsoft.com/zh-cn/subscriptions/zdh19h94(v=vs.90).aspx.

8 John Conroy. How They Hack Your Website:Overview of Common Techniques. 2008-03-05[2013-01-15]. http://www.cmswire.com/cms/web-cms/how-they-hack-your-website-overview-of-common-techniques-002339.php.


(上接第 209 页)


5 Kim K. Financial time series forecasting using support vector machines. Neurocomputing, 2003, 55: 307–319.

6 Huang W, Nakamori Y, Wang S. Forecasting stock market movement direction with support vector machine. Computers and Operations Research, 2005, 32(10): 2513–2522.

7 张晨希,张燕平,张迎春,陈洁,万忠.基于支持向量机的股票预测.计算机技术与发展,2006,16(6):35–37.

8 Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. Taiwan: Taiwan University, 2008.

9 Keerthi S, Lin C. Asymptotic behaviors of support vector machines with Gaussian kernel. Neuraleomputation, 2003, 15 (7): 1667–2689.

10 Scholkopf B, Smola A. Leaning with kenels. Citeseer, 2002.