南京航空航天大学
毕业设计（论文）开题报告

学　　院：	理学院
专　　业：	信息与计算科学
题　　目：	支持向量机在金融数据挖掘中的一种应用研究
学生信息：	郭富士　　091603430
毕设地点：	南京航空航天大学
指导教师：	袁泉　　副教授

2020 年 2 月 18 日

1. 研究背景及意义

1.1 研究背景

1602 年，荷兰建立了世界上第一个以金融股票为主的证券交易所，即阿姆斯特丹证券交易所。在此之后，不断壮大的资本市场便在人类的各类生产经营活动中，发挥着不可替代的重要作用。其中，中国的股票市场是近几十年来变化最快的。自 1990 年的上海证券交易所（Shanghai Stock Exchange，SHSE）和深圳证券交易所（Shenzhen Stock Exchange，SZSE）成立的 30 年以来，中国股市的规模已经达到了 59 万亿人民币，境内上市公司已有近 3800 家。中国股市有力地推动了中国经济的发展，也是众多个人投资者认为的可以扩大财富的最佳捷径。

众所周知，股市是一个高度复杂的非线性系统，投资者在其中同时面临着高收益与高风险。在这种背景下，投资者希望能够获得一些投资建议，来尽可能地增加收益，减少损失。这便催生了关于股票市场的各种各样的预测问题，过去几十年，传统投资分析方法，包括基本面分析 (张秀云 & 郭树, 2011) 和技术面分析 (戴洁 & 武康平, 2002)，随着证券市场的成熟和壮大已经逐渐失效。随后，又发展了如回归分析、时间序列分析 (李民 et al., 2000)、马尔可夫法 (夏莉 & 黄正洪, 2003) 等利用统计学来建立预测模型的方法，但它们无法描述股票交易数据的非线性性，预测效果并不理想。而最近几年，量化投资作为一种逐渐发展起来的新兴投资方法，基于机器学习技术挖掘股市数据的变化规律而建立的投资策略越来越受到投资者的青睐。投资者利用这些通过大量的数据分析得到的模型对新数据进行预测，很大概率能够获得预期收益，而且基于数据，量化投资是绝对理性的，不受投资者个人情绪的干扰。因此，利用机器学习方法对股票进行预测的前景是十分乐观的。

1.2 研究意义

与国外相比，我国的量化投资历史较短，发展还不够完善，但随着我国证券市场的成熟与壮大，量化投资的实践机会也越来越多。借此机会，我们将应用新的机器学习方法。支持向量机（Support Vector Machine，SVM）是 Vapnik 等人提出的，一种基于 VC 维理论和结构风险最小化（Structural Risk Minimization，SRM）的新兴学习机 (Cortes & Vapnik, 1995)。相比于其他机器学习方法，支持向量机具有难以比拟的优越性：首先支持向量机的求解过程是一个凸优化问题，保证得到的解是全局最优的；其次它能够很好地解决高维数据、小样本和非线性问题，且具有良好的泛化能力。基于此，支持向量机在模式识别、文本分类和目标检测等数据挖掘问题中都获得了较好的应用效果。

2. 国内外研究现状

2.1 股市预测方法的发展

自从股票诞生以来，投资者们都梦想着能够准确预测股票的涨跌，以此来获得超额的收益，但结果总是不尽人意。主要原因有二：其一是不确定因素，如国家政策 (许均华 & 李启亚, 2001)、自然灾害等，对股市影响深远但却很难量化；其二是数据的总量，受限于计算机技术，在此前无法对超大规模的数据进行挖掘与分析。但是随着学者的进一步研究、统计理论的完善与计算机技术的进步，现在已经能够克服上述困难对股市预测进行研究。解保华 (解保华 et al., 2002) 使用单位根、方程比

等方法对中国股票市场是否服从随机过程进行检验，结论是否定的。股票价格的未来走势与历史走势息息相关，找出这种关联性便成了研究的重点。一部分学者认为股票市场是线性发展的，提出了自回归模型 (李志林 & 王志刚, 2007) 和自回归移动平均模型 (翟志荣 & 白艳萍, 2010)。还有部分学者认为非线性模式能更好地描述股市，如 ARCH 模型 (唐齐鸣 & 陈健, 2001) 和 GARCH 模型 (魏巍贤 & 周晓明, 1999)。胡雪明和宋学峰 (胡雪明 & 宋学锋, 2003) 利用 MF-DFA 方法，来验证股市具有多重分形结构 (胡雪明 & 宋学锋, 2003)。周广旭 (周广旭, 2005) 建立了以径向基网络为核心的时间序列模型，来对股票走势进行预测。

近几年，人工智能的火爆使得以机器学习为核心的股市预测方法成为了主流 (林升 et al., 2019)。Patel (Patel et al., 2015) 用两种不同的输入数据方法，在印度股票市场上比较了四种机器学习预测模型的性能：人工神经网络、支持向量机、随机森林和朴素贝叶斯。 (林春燕 & 朱东华) 提出了一种 Elman 递归动态神经网络预测模型，在处理动态序列数据输入输出具有良好的性能。Nair (Nair et al., 2010) 建立了一种自适应决策树-神经模糊的混合系统，先利用技术分析进行特征提取，后使用决策树进行特征选择，再对特征进行降维，最后应用在神经模糊系统中，该模型有效地将决策树与模糊神经网络的优点结合了起来，起到了十分不错的预测效果。余锴 (余锴, 2018) 提出了一种基于贝叶斯推断的隐马尔可夫模型（Hidden Markov Model，HMM），相比于传统的 HMM 更加稳定且精确。

2.2 支持向量机在股市的应用

支持向量机是统计学习领域的新宠，相比于神经网络这类基于经验风险最小化原理（Empirical Risk Minimization，ERM）的机器学习方法，其优势使得它被广泛应用在众多领域中，如语音识别 (Ganapathiraju et al., 2004)、图像分类 (Foody & Mathur, 2004) 和回归分析 (陈永义 et al., 2004) 等。因此支持向量机能够很好地对股价进行预测，早在 2001 年，Fan 等 (Fan & Palaniswami, 2001) 就开始使用 SVM 对具有超常收益率的股票进行选择。Lin 等 (Lin et al., 2013) 将支持向量机应用在股票趋势预测系统的两个不同部分：在特征选择部分选用了 SVM 过滤器来对特征进行排名和选择；在预测模型部分运用了具有复合拟线性核函数的 SVM 来逼近非线性分离边界。辛治运与顾明 (辛治运 & 顾明, 2008) 利用了最小二乘支持向量机来进行股市预测，与时间序列分析相比，该模型有着更高的精度和更快的训练速度。彭丽芳等 (彭丽芳 et al., 2006) 将支持向量机和时间序列分析结合起来，较好地解决了传统时间序列预测模型仅局限于线性系统的情况。张玉川与张作泉 (张玉川 & 张作泉, 2007) 建立了一个简单的分类模型，对个股的价格涨跌进行预测分类，实证实验也获得了不错的结果。王芳 (王芳, 2015) 在对沪深 300 指数的预测研究中，不仅构建了使用遗传算法优化的支持向量机预测模型，还建立了基于将时间序列模糊信息粒化的模型，得到了较高的预测精度。

3. 研究内容和方法

3.1 研究内容

本文的主要内容是支持向量机在股票价格预测中的一种应用研究。将使用股票二级市场相关数据作为分析基础，结合股市蜡烛图，从中筛选出特定的数据形态组合，给出未来几个交易日最低价的统计分析结果。再针对这些特定的数据形态组合，提取其中的特征，利用支持向量机训练预测模

型，并给出相应的预测结果。

3.2 研究方法

1. 基于股票二级市场相关数据：历史开盘价、历史收盘价、历史最高价、历史最低价和历史成交量，利用 Python 进行数据清洗。再从中筛选出特定的数据形态组合，计算出连续可变交易日的最低价平均值。

2. 将部分数据作为训练集，由于样本特征数量小，样本数量大，分别使用 2 个不同的核函数：多项式核函数 $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j + b)^d$ 与高斯核函数 $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ 建立 SVM 回归预测模型。

3. 利用网格参数寻优选择预测效果最好，即在交叉验证中得分最高的模型，其中使用均方误差 MSE 作为评判标准，最后利用此模型在测试集上进行预测，得到最终结论。

参考文献

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273–297.

Fan, A., & Palaniswami, M. (2001). Stock selection using support vector machines. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)* (pp. 1793–1798). IEEE volume 3.

Foody, G. M., & Mathur, A. (2004). Toward intelligent training of supervised image classifications: directing training data acquisition for svm classification. *Remote Sensing of Environment*, *93*, 107–117.

Ganapathiraju, A., Hamaker, J. E., & Picone, J. (2004). Applications of support vector machines to speech recognition. *IEEE transactions on signal processing*, *52*, 2348–2355.

Lin, Y., Guo, H., & Hu, J. (2013). An svm-based approach for stock market trend prediction. In *The 2013 international joint conference on neural networks (IJCNN)* (pp. 1–7). IEEE.

Nair, B. B., Dharini, N. M., & Mohandas, V. (2010). A stock market trend prediction system using a hybrid decision tree-neuro-fuzzy system. In *Proceedings of the 2010 International Conference on Advances in Recent Technologies in Communication and Computing* (pp. 381–385). IEEE Computer Society.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, *42*, 259–268.

余锴 (2018). 基于贝叶斯推断的 *HMM* 股价预测研究. Master's thesis 上海交通大学.

周广旭 (2005). 一种新的时间序列分析算法及其在股票预测中的应用. 计算机应用, .

唐齐鸣, & 陈健 (2001). 中国股市的 arch 效应分析. 世界经济, *3*, 29–36.

夏莉, & 黄正洪 (2003). 马尔可夫链在股票价格预测中的应用. 商业研究, *10*, 62.

张玉川, & 张作泉 (2007). 支持向量机在股票价格预测中的应用. 北京交通大学学报, *31*, 73–76.

张秀云, & 郭树 (2011). 股票基本面分析在实战中的应用. 中国证券期货, *5*, 28–29.

彭丽芳, 孟志青, 姜华, & 田密 (2006). 基于时间序列的支持向量机在股票预测中的应用. 計算技術與自動化, *25*, 88–91.

戴洁, & 武康平 (2002). 中国股票市场技术分析预测力的实证研究. 数量经济技术经济研究, *4*, 99–102.

李志林, & 王志刚 (2007). 股市预测的自回归方法. 统计与决策, *2*, 19–20.

李民, 邹捷中, 李俊平, & 梁建武 (2000). 用 *ARMA* 模型预测深沪股市. Ph.D. thesis.

林升, 綦科, 魏楷聪, & 张伟 (2019). 机器学习在股价预测中的研究综述. 经济师, (p. 29).

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, *26*, 476–0477.

王芳 (2015). 基于支持向量机的沪深 *300* 指数回归预测. Ph.D. thesis 济南: 山东大学.

翟志荣, & 白艳萍 (2010). 基于 matlab 的自回归移动平均模型 (arma) 在股票预测中的应用. 山西大同大学学报 *(*自然科学版*)*, *26*, 5–7.

胡雪明, & 宋学锋 (2003). 深沪股票市场的多重分形分析. 数量经济技术经济研究, *8*, 124–127.

解保华, 高荣兴, & 马征 (2002). 中国股票市场有效性实证检验. 数量经济技术经济研究, *5*, 100–103.

许均华, & 李启亚 (2001). 宏观政策对我国股市影响的实证研究. 经济研究, .

辛治运, & 顾明 (2008). 基于最小二乘支持向量机的复杂金融时间序列预测. 清华大学学报 *(*自然科学版*)*, (p. 20).

陈永义, 俞小鼎, 高学浩, & 冯汉中 (2004). 处理非线性分类和回归问题的一种新方法 (I)——支持向量机方法简介. 应用气象学报, *15*, 345–354.

魏巍贤, & 周晓明 (1999). 中国股票市场波动的非线性 garch 预测模型. 预测, .

指导教师意见（对课题的深度、广度及工作量的意见和对毕业设计（论文）结果的预测）：

指导教师签字：　　年　月　日

系审查意见：

系主任签字：　　年　月　日