复旦大学管理学院本科生毕业论文开题报告

姓名	李悦凡	学号	12307100009
所在系	财务金融	专业	财务管理
指导教师	王小卒	职称	教授
校外指导教师 及其所属单位		职称	
论文题目	基于行业财务数据、宏观经济数据以及时间序列的沪深 300 指数机器学习预测研究		

选题意义、参考文献及论文大纲（可另附页）

请见附页

研究进度及时间安排

起止日期	主要工作内容
3/1-3/7	开题报告
3/8-3/21	数据收集及整理
3/22-4/15	构建模型以及撰写论文初稿
4/16-6/2	进一步修改论文，提交终稿

指导教师对开题报告的意见：

指导教师签名：　　年　　月　　日

《基于行业财务数据、宏观经济数据以及时间序列的沪深 300 指数机器学习预测研究》开题报告

选题意义

沪深 300 指数作为反映市场整体表现的重要指标具有重要的研究和投资价值。沪深 300 指数的表现不仅能够用于经济周期等问题的研究，而且，其衍生出的指数期货、期权产品可以用于对冲风险等投资决策。因此，准确判断未来沪深 300 指数走势具有很高的实用价值。

传统的沪深 300 指数预测方法包括基本面和技术分析方法，以及金融时间序列方法。尤其是在金融时间序列研究中，通过研究沪深 300 指数的序列自相关性，对沪深 300 指数的未来走势进行预测。这种方法具有一定的局限性。首先，该方法具有一定的前提假设，如沪深 300 指数的收益率服从高斯分布。而实际市场的收益率相比于高斯分布，往往具有高峰厚尾的特征，所以这个假设存在一定的局限性。其次，时间序列方法具有一定的滞后性，无论是自回归模型还是滑动平均模型，其自变量均为过去一定时间段内的历史数据。这些数据能够反映一段时期内的趋势，但是当市场出现重大的结构性变化时，这些模型往往表现滞后。

近年来，随着计算机技术的进一步发展，机器学习的概念越来越受到重视。相比于传统的统计方法，机器学习具备独特的优势。机器学习模型需要的前提假设较少，比如在均方误差（MSE）的传统统计学推导中，需要涉及到分布函数的假定和近似，而在机器学习方法中常用的 Bootstrapping 算法则可以通过随机重复采样来估计均方误差，两种方法得到的结果相差极小，而且 Bootstrapping 方法的前提假设较为宽松。其次，机器学习算法对于非线性的数据关系表现较好。在金融时间序列中，主要还是采用了线性回归模型，认为当前市场表现与之前的历史数据呈现线性关系，然而实际情况往往并不是如此直接的（见下图）。而机器学习算法并不要求数据之间一定存在这种线性关系，使得机器学习算法更加的灵活和准确。第三，机器学习算法可以便捷的添加新变量进行重新训练，比如目前研究较多的情感计算，可以通过情感计算来分析市场舆论的倾向，并作为变量输入到模型中进行新的训练。

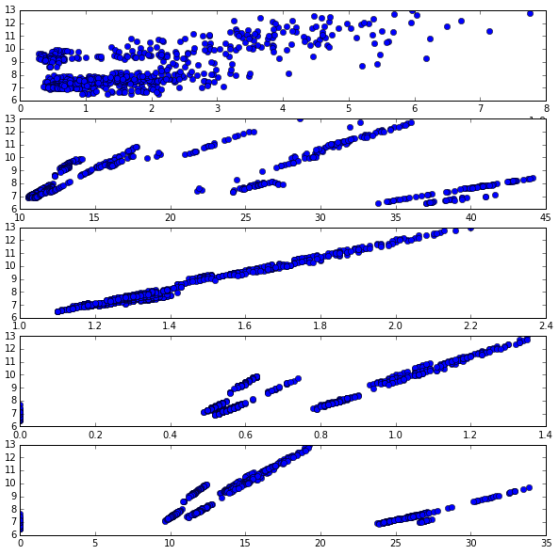

Figure 1 **市场指数与部分财务风险指标的关系**

目前国内关于机器学习在金融指数领域应用的研究主要集中在模型优劣的选择上，主要的侧重点是算法本身，而对于金融数据本身的特点研究较少。大多数研究倾向于使用支持向量机 (SVM) 和神经网络算法来对数据进行分析和预测。支持向量机和神经网络作为常用的非线性分类器在研究和实践中被广泛应用，确实获得了很好的效果，不过也有它们自身的一些局限性。

我国目前金融市场样本数量不足。我国股市从 90 年代早期建立直到今天也不过短短 25 个年头，如果考察每日收益率作为样本，则总的样本点不超过 7000 天。或许在传统金融时间序列领域，7000 个样本点已经可以用于建模和预测，然而对于机器学习来说，这样的样本规模是偏小的，因为机器学习处理的数据维度较高。尤其是对于像支持向量机和神经网络这样参数较多、变量较多的算法，过少的样本点很容易导致过拟合问题，进而削弱模型的预测能力，根据部分文献的记录，神经网络算法的预测准确度只有 50%，也就意味着和随机猜测相比并没有显著的表现提高。所以，对于日收益率时间序列，可能相对比较强健而简单的机器学习模型是较好的选择，从这个角度来说，本人将会倾向于选择随机森林算法进行分类，同时训练支持向量机以及神经网络作为对比。

此外，现有文献中，并没有针对金融市场的风险进行调整。金融市场对于大幅下跌尤其敏感，而探测并且预报潜在的大幅下跌也是指数预测模型的一个重要功能。这就要在模型的训练阶段就对模型施加一个特定的惩罚系数矩阵，构造非均衡机器学习模型，这样训练出的模型就可以更准确的预报未来的巨大风险，尽管这样的做法付出的代价可能是对于未来上涨的预测准确度将会一定程度下降。

最后，我将会采用目前在机器学习领域较为流行的模型衡量指标 AUC (Area Under ROC Curve) 来评价各个模型。之前的文献比较关注于模型的准确度（Accuracy），但是这样的衡量方法是有局限性的。

		预测结果	
		正面	负面
实际数据	正面	A	B
	负面	C	D

上图是一个 Confusion Matrix，A、B、C、D 分别表示 “预计为正而实际也为正”、“预计为负可是实际为正”、“预计为正可是实际为负” 以及 “预计为负实际也为负” 这四种情况。我们平时常用的准确率（Accuracy）= (A+B) / (A+B+C+D)。可是，即使整体的准确率很高，也有可能是由于数据不平衡导致的。举一个最极端的例子，比如假设某国金融欣欣向荣，过去一年以 252 个交易日计算，共计上涨 251 天，而剩下一天下跌 10%。那么即使一个模型盲目全猜涨（即恒猜涨），也能获得超过 99%的预测准确度。然而在 AUC 的度量下，由于 AUC 同时需要考虑犯错的情况，所以全部盲目猜涨的情况下，AUC 仍旧等于 0.5。这样就可以一定程度上化解由于训练样本集不平衡导致的问题。

综上所述，本文将会试图从金融学的视角来考察基本的机器学习算法在中国金融市场指数预测中的应用和问题。

论文纲要

摘要

1. 前言
    1.1 课题研究意义
    1.2 课题研究方法
2. 金融市场指数预测方法
    2.1 金融时间序列方法
        2.1.1 ARIMA 模型
        2.1.2 GARCH 模型
    2.2 机器学习方法
        2.2.1 支持向量机
        2.2.2 神经网络
        2.2.3 随机森林
3. 针对沪深 300 指数收益率的实证分析
    3.1 基于行业财务数据的机器学习模型
    3.2 添加宏观经济因素后的机器学习模型
    3.3 加入时间序列概念后改进的机器学习模型
    3.4 模型表现分析
4. 结论

参考文献

[1] 裴双喜. 基于数据挖掘的金融时间序列预测分析与研究[D]. 大连海事大学, 2008.
[2] 孙吉红. 长时间序列聚类方法及其在股票价格中的应用研究[D]. 武汉大学, 2011.
[3] 卢钰. 基于参数优化的支持向量机股票市场趋势预测[D]. 浙江工商大学, 2013.
[4] 鲍漪澜. 基于支持向量机的金融时间序列分析预测算法研究[J]. 博士论文, 大连海事大学, 大连, 2013.
[5] 李巍. 基于宏观经济指标和人工智能方法的上证综合指数预测[D]. 西南财经大学, 2012.
[6] Lee C, Shleifer A, Thaler R H. Investor sentiment and the closed-end fund puzzle[J]. The Journal of Finance, 1991, 46(1): 75-109.