人工智能导论在股票预测领域的应用

周五上午班第16组

李子南，邬静芙，钟日涵，丁泓勋，徐春骥

研究背景

“股票市场指股票发行、买卖、交易的市场，是证券市场的一部分。”由于其与经济活动的密切联系,它又被称为经济和金融活动的"晴雨表"。中国股票市场经过三十余年的快速发展,对我国经济社会建设和社会经济发展均产生了深远的社会影响。然而, 由于中国股市的频繁波动, 加上投资者对股价走势的错误判断, 往往会导致市场危机, 严重的还会演变成"股市灾难"。因此, 能否有效预测股票市场以及如何更精准预测股票市场对于学术界和金融界都是广泛关注的课题。

传统的股票预测方法使用统计原理和计量经济学模型来描述金融时间序列数据，比如在计量经济学领域里，自回归条件异方差模型能准确地模拟波动性的变化的时间序列变量，它在金融工程学的各种实证研究中都有着广泛的应用，它能使人们能更加准确地把握时间变量序列（如证券）的波动，许多金融公司也将其应用于预测时间序列波动中。在统计中，自回归模型和移动平均线(即ARIMA)，即box-Jenkins模型，是主要模式的时间序列。许多研究者基于股票历史价格趋势有关时间序列的数据来回归建模，并利用回归模型预测短期时间内股票价格的变化。研究者吴玉霞和温欣[1]选取"华泰证券"股票收盘价作为时间序列进行数据分析,在建立ARIMA模型的基础上，对创业板市场股票价格变动的规律和趋势进行了预测。实证结果表明, 该模型短期动态、静态预测效果较好, 可以为投资者和企业在进行相关决策时提供有益参考。但传统的回归方法对样本量和分布程度的要求较高，且难以保障预测的精度和稳定性，导致结果缺乏普遍性，因此，适用性十分有限，并不能满足大多数的应用场景。

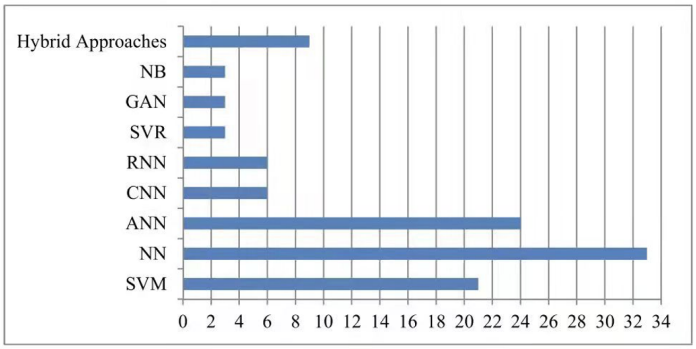


Fig. 2. Most frequently ML techniques.

资料来源 [5]

近年来, 人工智能应用技术已经逐渐开始迅速发展兴起。深度学习技术作为发展现代科学人工智能的一个重要技术分支, 在图像识别、语音学和视觉图像识别、自然语言处理等诸多应用领域已经成功取得了显著的科学研究应用成就。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。从上图中可以看到，常用的机器学习算法包括NN，SVM，ANN等，对于算法的详细内容我们将在后续综述中进行介绍。监督学习从训练数据集中学习函数，当新数据输入时，可以依据该函数进行结果预测。监督学习的训练集包括输入和输出，即特征和目标，这就使得该算法相较于传统的股票预测方法，有更大的优势，对数据没有过高高精度的限制。在金融大数据领域, 机器学习算法的应用主要集中在三个方面: 一是利用机器学习来建立或改善投资策略; 二是采用机器学习算法挖掘金融大数据，获得更多有效信息; 三是利用机器学习预测金融市场的波动和价格，也是关注度相对最集中的方面，我们综述将更加关注这个方面。

金融大数据的特征与内涵

大数据不只是简单的"大量数据", 它除了历史数据, 还有无时不刻生产的数据, 甚至包括未来将产生还未产生的数据。它具有规模大、功能多、更新快的特点。大数据存在于工业、农业、消费、金融、媒体等各个行业。其中,金融服务大数据以其庞大的应用规模和重要的学术研究应用价值已引起专家学者们的广泛高度关注。通过对这些数据进行分析和预测，互联网金融公司可以掌握用户的消费习惯，收入等级等信息，从而对用户的金融投资行为进行一定程度上的预测，这样就可以更有针对性的向目标用户群体投放金融产品的广告，从而使公司占据优势地位。许多学者认为, 大数据技术的使用将给金融领域带来深刻的变化, 企业若是能利用这些技术创造更多的价值，提高效率，就能形成经济产业， "金融科技（FinTech）"的概念也随之出现。由于金融科技的颠覆性创新，作为对于金融服务业的产品开发、业务进行、流程分析可谓是一种新兴而高效的解决方案。

金融大数据按行业可分为证券大数据、银行大数据和保险大数据。证券大数据包括股票市场、债券市场和衍生品市场的历史市场数据和交易数据。银行大数据包括存款额、货币量和信用卡用户的海量数据等指标。保险大数据包括各种理赔数据、保险数据等。这些金融大数据都来自于现实生活，具有地域跨度大、形式多样、结构不同的特点。通过互联网的辅助，大多数金融产品及服务都能得以实现。同时，降低了金融产品消费者和服务提供者之间的信息不对称。所以，大数据金融无疑是高效的。许多业务能够实现在线进行，还有自动化行动。产品得以在极低误差的情况下交予消费者。得益于金融大数据的强大的数据挖掘能力，许多与此相关的金融业务变得十分高效，同时其成本也可以大幅降低。研究的难点主要在于数据处理和挖掘，而非数据收集和整理。我们需要使用各种工具和技术来分类、处理和挖掘收集的大量数据。然而，市场环境是不可预测的，很难像处理传统数据一样找到适用性强的传统策略，这也使得新兴的大数据技术显得十分重要。

大数据技术在经济金融领域的重要价值包括统计和预测。通过大量数据的一系列计算方法,对一段时间内的趋势进行预测,预测结果更加准确。科技的高速发展使得金融大数据的内容开始从不同的角度得到填充。学者们利用爬虫技术在模型中引入文本特征，构造投资者情绪和分析方法等指标进行研究，得到了些许新的结果。如顾文涛等构建了适用于金融投资领域的财经新闻情感词典来对财经新闻进行文本分析,同时构造了新的预测模型:将财经新闻文本中所含的情感量化为情绪指数并与时变密度函数相结合,得到时变加权密度模型，并以此为基础，将多个预测模型结合，以模型得分为权重，构建得分加权模型，对股票收益率进行预测。根据试验结果, 添加情感指标可以有效提高模型的预测能力，得分加权模型的预测能力进一步提高，基本实现了精度和得分规则的双最优。赵明清、吴胜强[4]结合百度指数,利用时差相关系数和随机森林选取微博搜索初始关键词,通过爬虫技术获取微博文本,利用文本挖掘技术对微博文本作分词处理,判断分词后的微博情感倾向,分析影响微博影响力的相关因素,以信息增益确定微博权重，最终得到了两个具有良好预测功能的模型。在金融大数据领域，机器学习技术成为越来越多的学者选择的研究视角。

决策树与随机森林

决策树

在众多机器学习的方法当中，决策树（decision tree）是最贴近我们日常生活的方法之一，我们平时在生活中经常会用到决策树的思想。比如我们会通过西瓜的瓜蒂形状来判断这个瓜甜不甜，也会通过观察螃蟹的肚脐是尖的还是圆的来判断螃蟹的性别；或是像生物学博士“无穷小亮的科普日常”那样来鉴定网络热门生物时，通过界门纲目科属种的顺序对其进行大致分类，最后通过各式各样的更加细致的生物特征来确定该网络热门生物到底是什么生物。如果把上述的决策过程归纳成一颗树的形状，每个上层节点通过一些特定规则分裂成下一层的节点，最终的叶子节点就是分类结果。

决策树就是这样形式进行的一种思想，基于不同规则与一个或是多个不同的特征来进行分类决策。而决策树算法的学习过程的重点就在于如何选择最优最好的划分属性，即随着决策树划分层数的增加，该决策树分支的所含有的样本应逐渐倾向于一个类似的类别。也就是说节点分裂时要使得节点分裂后的信息增益最大化，以达到最大信息增益为决策树构建的特征选择依据。

决策树学习的三个步骤分别是：特征选择、决策树生成和决策树剪枝。首先的特征选择决定了所要构建与学习的决策树模型要使用哪些特征来进行判断。我们知道在训练数据集当中，每个测试集的样本属性可能会有很多个，不同属性其所对应的作用也有大小之分，这一点很贴切于股市预测。因而特征选择的作用就是筛选出和分类结果所相关性较高的特征，换句话说也是分类能力较为强大的特征。而信息增益是在股市预测中决策树搭建的特征选择经常使用的标准准则。决策树模型是指选好特征之后，就从根节点开始对每个节点进行所有特征的信息增益，选定在信息增益当中增益最大的特征作为节点

特征，根据该特征的不同取值再建立子节点以此类推，直至信息增益很小到一定程度或至没有特征值可以选择为止。决策树剪枝的目的是为了减少“过拟合”，通过减去部分分支来降低可能过拟合的风险，同时提高决策树的效率。

决策树在近四十年的发展中主要发展出了三种算法：ID3算法、C4.5算法和CART算法。研究者们利用决策树直接构建来进行股市预测模型的文献较少，大部分股市预测于决策树的模型均加入了其他算法对模型以及决策树的优化。接下来本文将对上述三种决策树算法在股市预测方面的应用进行简要综述。

ID3算法

研究者刘利与何先平[1]尝试利用决策树ID3算法与模糊理论结合，并加入遗传算法对ID3决策树进行优化，最终易形成可以进行股票交易决策的一种改良后的模糊决策树模型。在上述算法思路结合过程后，两位研究者提出了GAFDT模型。并选取了某股票交易公司在2006年6月10日至6月30日的股票交易数据进行模拟构建决策树。测试显示，虽预测股票上升或下降趋势的准确度上相对传统FDT方法预测来说准确度要更高，但是这一模型由于选取的时间跨度较短，同时未考虑模型可能存在的过拟合问题，所以该模型的实际使用效果会比测试要差。

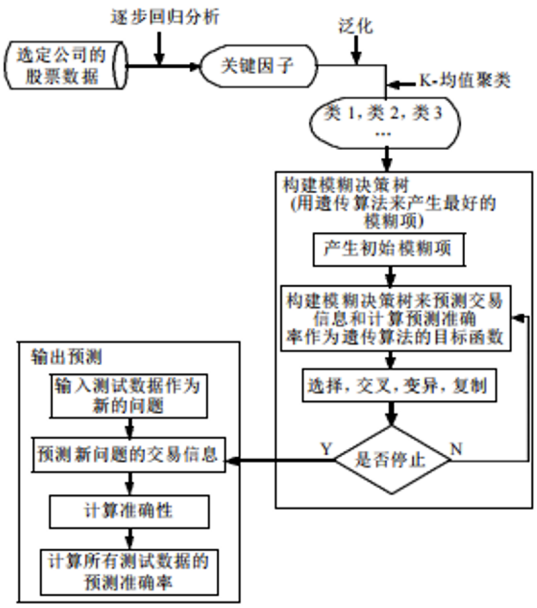


图 1　GAFDT 详细模型

表 4　预测准备率 (FDT 对 GAFDT)

	FDT		GAFDT		
	平均值	最好值	平均值	最好值	最好模糊项
EPISTAR	0.710 0	0.772 7	0.829 0	0.895 1	7，9，9，8
SiS	0.719 4	0.811 1	0.828 4	0.905 5	9，9，9，9
UMC	0.714 8	0.809 1	0.800 8	0.895 1	8，8，8，8

C4.5算法

研究者胡扬与王领[2]通过优化C4.5决策树对50支股票进行模拟，并加入了多个额外的相关技术指标进行模型优化，结果得出，在利用决策树进行股票预测时，应组合多个技术指标能使预测准确度增加不少，同时也说明了该模型可以有效知道投资者在股票市场中规避风险以及进行合理投资。

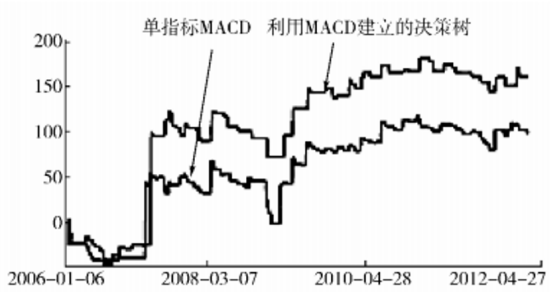


图 3　单指标(MACD)与决策树的收益对比图

CART算法

研究者王禹、陈德运和唐远新[3]选取了Cart算法与boosting算法相结合，还加入了两种纵向性特征预测来提高预测的准确度。经测试，通过利用boosting方法来优化CART决策树算法来建立股市预测模型，相比于传统的股市预测算法要准确度精确度都更高，还能更高效的寻找到不同特征数据间的可能的隐藏存在关联，防止股市中噪声数据等的过度干扰（因为股市是一个具有许多噪声的环境）。尽管研究者们加入了两个输入变量提高了预测的精确度，但这一模型并没有对该决策树模型可能存在的过干扰问题进行研究探讨。

表 1　实验列表

Tab. 1　Experiments

实验	输入	方法	说明
实验一	F_1	Cart 树	采用 Cart 决策树进行预测
实验二	F_1	Cart 树结合 boosting 算法	Cart 树与 boosting 算法进行预测
实验三	F_2	Cart 树结合 boosting 算法	Cart 树与 boosting 算法结合，并增加新特征

表 2　结果比较

Tab. 2　Comparison of experiments

误差	方　法		
	Cart 树五特征	基于 boosting 的 Cart 树五特征	基于 boosting 的 Cart 树七特征
mse	0. 790 283	0. 701 280	0. 568 172
rmse	0. 888 979	0. 837 424	0. 753 772

随机森林

随机森林（Random Forest）作为一种相对较新的机器学习方法，近年来在股市预测方面的应用与关注度在逐渐提升。2001年Breiman在决策树算法的基础上提出了此随机森林算法，随机森林室友多个随机子集生产决策树的实例所构成，且不同决策树之间没有关联，联系离散数学中树与森林的概念，故形象的称之为“随机森林”。

在维基百科中随机森林的定义为：“随机森林或随机决策森林是用于分类，回归和其他任务的集成学习方法，其通过在训练时构建多个决策树并输出作为类的模式（分类）或平均预测（回归）的类来操作。个别树木。随机决策森林纠正决策树过度拟合其训练集的习惯。”其最大的特点就是可以解决决策树算法中的过拟合问题。

随机森林的构建包括特征与标签提取、特征预处理、样本内训练、交叉验证和样本外测试等步骤[4]。随机森林是一种由许多决策树结合通过Bagging的方式组成的分类器，是一种集成式学习器。我们从原始数据集生成 N 个 Bootstrap 数据集，对于每个Bootstrap 数据集分别训练一个弱分类器，最终用投票、取平均值等方法组合成强分类器。

研究者邓晶与李璐[5]通过改进随机森林在参数优化阶段的参数组合，通过网络搜索算法构造了一个最优参数指标的股票预测随机森林，有着较高的预测能力。

该模型的具体构造步骤如下：

1. 收集并获取股票的价格、股票历史涨幅等相关数据，计算得出相关应用的技术类指标并设置重点来组成数据集，并对数据进行分析演算和处理。
2. 将网格搜索算法作用到该随机森林模型中，并使用交叉验证的方法对各组参数的表现（即准确率等相关参数）进行评估，从而得出最优参数组合。
3. 构建随机森林预测模型，然后使用准备好的测试集对参数优化后的随机森林股票模型进行测试，对比决策树和支持向量分类器以及参数优化前的决策树，并进行模型评价。
4. 最后根据股票的涨跌情况等信息的预测分析为投资者提出投资建议。

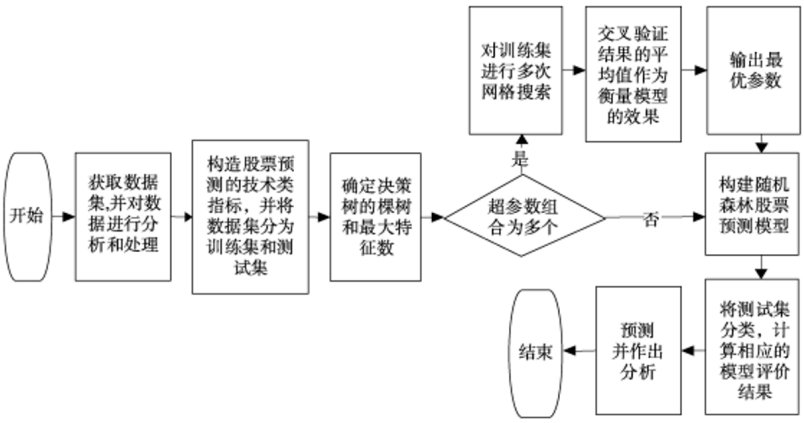


图 1　参数优化的随机森林股票预测流程

Fig.1　Random forest stock forecasting process with Hyper-parametric optimization

表 3　不同股票预测实验结果

Tab.3　Experimental results of different stock prediction

证券简称	支持向量机	决策树	原始随机森林	参数优化后的随机森林
平安银行	0.5977	0.6568	0.7751	0.8343
万科	0.5965	0.6316	0.7485	0.7661
深振业 A	0.5630	0.6222	0.7333	0.7481
神州高铁	0.6456	0.6392	0.7342	0.7405
美丽生态	0.6024	0.6625	0.7108	0.7590

随后研究者利用该模型堆5个月内多支不同股票进行了预测，并且进行了对比试验，证明了随机森林模型在股市预测方面的有效性，且在参数优化后能达到更高的准确性。同时这也说明中国股票市场的可预测性，但是由于股票的不断发展与变化，在未来的随机森林与股市预测中，堆参数指标的选择与构造都还要进一步优化与完善。随机森林算法在股市预测仍然具有较大潜力。

支持向量机(SVM)

支持向量机(SVM)是间距最大化、定义在特征空间的线形分类器, 是一个二分类模型,其中的核技巧(核函数)将数据向更高维空间映射,因此能够高效地处理非线性问题,使其能对无法直接线性分类的数据较好分类。它的学习策略是将间隔最大化,能够用解凸二次规划问题的方法解决,但其易出现局部最优解问题[1]。

股票为非线性、高噪声、波动性较强且需要对多特征数据进行分析的数据。据此特点,在股票预测中,SVM的独特优点就是可以使用核函数,利用非线性映射把股票数据映射到更高维空间,进而可以使用线性函数对股价的升跌做出分析。

支持向量机中,在大量已具有的核函数中选取的问题目前还无法很好地解决。同时,当该方法运用于大规模训练样本时消耗了巨大的计算机存储器空间和运算时间,其在大数据分析时代股票预测中的发展空间也因此受到了相当程度的制约。

近年来随着人工智能在股市预测中应用的拓展与深化,研究者为了发挥各预测方法的优点建立了许多组合模型,其中支持向量机的运用尤为重点。支持向量机与包括遗传算法(GA)、时间序列算法(Arima)、近邻传播(AP)等等算法结合后产生了各种各样的组合模型,其中一些展现出从预测效率到预测准确度的优势,甚至使SVM被称为"金融市场中最强大且最具预测力的工具",在人工智能运用于股市预测的权威论文数的运用频次占比接近1/3[2]。接下来本章将针对三类各具独特侧重点的支持向量机结合模型展开综述。

遗传算法-支持向量机(GA-SVM)

遗传算法是一个以生物进化理论（Darwin）为理论基础、进一步发展出来的自适应全局优化随机搜索算法,其本质上是一个不依靠具体问题的、自组合、自适应的直接检索方式。其底层思想主要来自于达尔文的生物进化论、魏茨曼的生物学选择理论及其孟德尔的族群基因遗传理论,具有坚实的生物学基础。它以作为经典高性能模型,通过统计、模型、寻优的方式而逐渐成熟[5]。

将遗传算法强大的全局优化搜索性能与支持向量机相结合,张伟等人建立了遗传算法-支持向量机模型。该模式通过使用遗传算法(GA)的全局自动寻优能力率先智能地找出了SVM的最佳参数,然后再去除了所有冗余特征,从而大大提高了估计的命中率,也大大减少了估计工作量和预计时间[6]。

其模型建立过程如下:

1. 选定原始特征向量并初始化种群

每一特点都被界定为一个具遗传效应的dna片段,用一条总长度为特征个数的二进制串描述每个特点,该二进制串为染色体(若染色体某位为“1”,则表明该位特点被选中,反之则表明该位特点被遮蔽,每一个染色体代表不一样的特征子集),并由此随机地形成初始群体。

2. 寻找核函数的最优尺度参数σ和惩罚因子C

对于SVM所采用的高斯核函数,必须先按经验规律确定尺度参数σ和惩罚因子C的范围,在此后进行离散化和二进制编码。对种群进行计算适应度、选择(符合条件的作为下一代)、交叉(交换基因)、突变(改变单个基因,防止局部极值)的操作。如σ(0,10),步长为10/1024,可得尺度参数σ的二进制串为X=X1×2×3×9×10惩罚因子C∈(0,100),量化步长为100/1024,得到Y=y1y2y3 y10,于是染色体为XY=x1×2 x10y1y2 y10。遗传算法在寻优完毕后,按照参数和二进制编码之间的映射关系对fσ=fσX,C=fCY进行编码,从而把染色体转换成实际的尺度参数σ和惩罚因子C。

3. 优化特征向量,获得最佳特征向量集
4. 通过使用更新后的SVM对测试集进行检测,可以得到高准确率

将所训练模型实际运用于股票预测中,并与统计建模、时间序列模拟方法、神经网络等进行了比较实验,后采用均方根误差(RMSE)作为模型的评估指标证明,GA-SVM相较其他几种方法除训练耗时较长外均具有优势。

表 6　SVM 与 GA 优化的 SVM 的预测效果比较

预测模型	准确率/%	训练耗时/ms	预测耗时/ms	特征个数
SVM	62.79	110	68	20
GA 优化参数	71.32	77 890	78	20
GA 优化特征	70.54	128 063	66	9
GA-SVM	77.52	189 406	62	6

表 7　GA-SVM 与其他主要非线性预测方法效果比较

预测模型	准确率/%	训练耗时/ms	预测耗时/ms	特征个数
GA-SVM	77.52	189 406	62	6
蚁群优化 SVM	70.96	66 570	66	20

时间序列算法-支持向量机(Arima-SVM)

时间序列算法(ARIMA)是针对平稳数据序列的、传统的线性时间序列预测模型,而支持向量机则是一种性能优异的非线性分析算法。该模式用时间序列计算捕获股价的线性规律,用支持向量机捕获股价的非线性规律,并开创性地使用了基于小波分解的时间序列计算-支持向量机组合模式,对股价收盘价格这一经典的非平稳时间序列进行预测[14]。

其模型建立过程如下:

1. 将数据小波分析

首先使用Mallat算法(二抽取是该计算的基本原理,信号数据会在通过各层划分后比划分之前的总长度减半,而总输出数据长短和输入数据长短保持一致。经Mallat算法分析后的数据信息可实现二插值重构),对非平稳时间序列加以分析并重建获得的低频和高频信息,分别对应线性部分和非线性部分。

2. 线性部分平稳化处理后用于构建ARIMA模型

高频信号经重构后可类似地视为平稳的时钟序列,并构建了ARIMA模型估计的股票价值历史统计。

3. 非线性部分定阶后将其残差作为SVM的输入数据

由于低频信号因表现的长期趋势而具有高度非线性,因此可以利用SVM模式对其非线性规律加以建模分析与预测。该模式中SVM训练所使用的内核函数都是高斯函数,可以在MATLIB7.0平台下自编程软件,调出LIBSVM工具箱来进行SVM模拟,并通过减半的交叉检验后,经gridre-gression.py自动检索,得出模式的最优预测参数[14]。

表 1　三种模型预测结果对比表

日期	实际值	ARIMA	SVM	组合模型
2006－11－30	6.44	6.54	6.51	6.48
2006－12－1	6.47	6.58	6.54	6.52
2006－12－4	6.38	6.45	6.42	6.41
2006－12－5	6.67	6.83	6.74	6.71
2006－12－6	6.39	6.53	6.46	6.43
2006－12－7	6.27	6.39	6.33	6.32
2006－12－8	6.08	6.21	6.15	6.13
2006－12－11	6.14	6.27	6.21	6.18
2006－12－12	6.11	6.25	6.18	6.15

表 2　三种模型的 RMSE 对比表

模型	RMSE
ARIMA	51.3
SVM	21.43
组合模型	7.58

近邻传播-支持向量机(AP-SVM)

针对非线性数据，支持向量机是一种具有优秀性能的算法；相对应的，针对平稳数据序列，时间序列算法(ARIMA)是传统的线性时间序列预测模型。该部分介绍的模型开创性地将线性规律用ARIMA捕捉,对于非线性规律则用SVM映射到高维空间进行分析,并开创性地构建出了ARIMA-SVM组合模型（基于小波分解）,对股价收盘价格这一的非平稳时间序列进行预测[4]。

其模型建立过程如下:

1. 将数据小波分析

首先使用Mallat算法(二抽取是该计算的基本原理,信号数据会在通过各层划分后比划分之前的总长度减半,而总输出数据长短和输入数据长短保持一致。经Mallat算法分析后的数据信息可实现二插值重构),对输入的非平稳时间序列重构（基于小波分析的结果）获得分别对应非线性部分的高频信息和对应线性部分的低频信息。

2. 线性部分平稳化处理后用于构建ARIMA模型

高频信号经重构后可类似地视为平稳的时钟序列,并据此构建了ARIMA模型估计的股票价值历史统计。

3. 非线性部分定阶后将其残差作为SVM的输入数据

由于低频信号因表现的长期趋势而具有高度非线性,因此可以利用SVM模式对其非线性规律加以建模分析与预测。该模式中SVM训练所使用的内核函数都是高斯函数,可以在MATLIB7.0平台下自编程软件,调出LIBSVM工具箱来进行SVM模拟,并通过减半的交叉检验后,经gridre-gression.py自动检索,得出模式的最优预测参数[4]。

表 1　不同时间间隔下四种模型的预测准确率对比表

Tab. 1　Comparison of prediction accuracy of four models at different time intervals

时间间隔 / d	准确率 / %			
	SVM	BP	AP-SVM-9	AP-BP
5	55.4	58.1	58.3	56.8
10	57.3	56.9	59.8	57.9
15	54.8	54.9	56.4	55.8
20	56.7	54.2	51.3	55.6

神经网络

神经网络是数据拟合预测领域一直以来的研究热点，神经网络的基本结构源自于对人脑中神经元的链接和信息传递方式的一种简化模仿，其具有一定程度的学习，分类，记忆和模式识别等信息处理能力的人工系统。神经网络区别于其他模型的一个重要特点就是它的学习能力，它能够将训练的结果通过参数的形式分布的储存在网络节点的链接中。神经网络具有大规模并行处理，分布式储存，高度冗余和非线性处理等特点。这些特点使得神经网络十分适合股价序列的预测，也广泛受到学者们的青睐。

传统神经网络

建立传统神经网络预测模型的基本步骤大致如下: 首先对样本数据集进行清理，对缺失值进行标准化，然后设计构建合适的网络结构，确定节点数、隐含层数和初始参数，通过不断学习优化找到最优的学习参数。其中根据数据集的拥有标注的多少，学习过程还可大致分为“无监督学习”，”有监督学习“和”半监督学习“。而股价序列的预测问题大多都是使用了监督学习。

与SVM、随机森林等算法相比，神经网络算法较早的被研究者应用于股票分析预测，早在上世纪 80 年代末，Halbert White[18]就尝试使用神经网络对IBM的股价进行估价预测，虽然陷入了过拟合和局部最优的问题中导致实验效果不好，但是在文章末尾，White分析提出了改进的方法并表示仍相信神经网络的潜力，为后续的研究者提供了方向和信心。为了提高传统神经网络模型的预测准确性，我国科学家蔡红和陈荣耀[19]将主成分分析与神经网络相结合，对变量进行分析和筛选，使得输入BP神经网络中的干扰变量大幅减少，成倍的减少了神经网络的训练时间，并且提升了预测精度。其在文中对在上证券所上市的首创股份的股价进行了模拟预测，其结果较为精确，但因为只对一只股票进行了分析预测，其偶然性还待考证。类似的，王文波[20]等人使用经验模态分解算法将原始信号分解成近似于独立的不同尺度的各种模态分量，再经过混沌分析和神经网络进行分析预测。结果表明，该方法具有良好的预测效果，但在长期预测中仍存在较大误差，且容易陷入局部最优。杨进和陈亮[21]将小波神经网络和ARIMA模型相结合。小波神经网络在应对非线性数据具有较好的表现，但在进行时序分析时表现较差。而ARIMA模型则是对时间序列数据进行分析和预测的模型。杨进和陈亮将小波神经网络和ARIMA的预测结果相结合，最终得到的预测结果精度较高，但其在实验中训练神经网络所使用的数据集的股价波动幅度较小，无法保证模型在股市普遍情况下的表现。郭怡然和王秀利[22]将BP神经网络和遗传算法相结合，遗传算法适于解决全局最优化问题，将BP神经网络中的参数作为遗传算法的初始种群，提取多种股市指标和数据作为输入参数，对比不同初始参数预测效果的效果，成功建立了我国A股市场风格轮换的分析预测模型，结果显示以股价为参数的BP神经网络对我国股票风格轮换的分析预测较为准确。綦方中[23]等人将主成分分析法和改进后的果蝇优化算法(FOA)引入BP神经网络，使用主成分分析法对原始数据进行降维处理，减少信息冗余，采用FOA优化BP神经网络的初始值和阈值。使用训练好的模型对A股股票价格数据进行仿真验证，结果显示在股价预测中,采用FOA优化的模型比原始神经网络的预测准确度更高。

深度学习

由于传统神经网络模型存在易陷入局部最优、收敛速度慢、易“过拟合”等一系列缺陷，随着近年来计算机GPU计算能力的大幅提升与相关理论的发展，相较于传统神经网络，网络结构更为复杂且拟合效果更好的深度学习算法得到了发展，相较于传统的神经网络，深度学习在理想状态下的表现要优于传统机器学习。但同时，深度学习所需要的数据量更多，训练时间更久，可解释性更差。深度学习主要包括卷积神经网络(CNN)、循环神经网络 (RNN) 、长短期记忆神经网络(LSTM) 三种网络。卷积神经网络的结构更类似于生物神经网络，与普通神经网络相比，CNN中加入了卷积层和池化层，减少了参数的训练次数，并更有利于提取高维的特征。RNN在处理时间序列数据方面有很大的优势，但RNN不具有长时记忆性，无法保留较长时间前的信息。LSTM对RNN的缺点进行了改进，其在RNN的基础上加入了门机制，提升了RNN对于长时间延迟信息的处理能力。

由于股票分析预测领域的时间序列性，非线性，大数据量等特征，使其十分适合深度学习的应用。近年来，越来越多的学者对如何将深度学习算法应用到股票预测领域展开了深入的研究。陈卫华[24]首次将深度学习运用于高频波动预测领域进行泛化预测，并将预测结果和19种传统模型在多种Loss函数下相比较。其结果显示，深度学习算法在每种Loss函数下的预测准确度都为最佳，相较于第二名，深度学习的预测精度提升了约9%-13%。研究还发现深度学习算法受关键参数的影响较小，在大多数情况下LSTM模型在测试模型中具有最好的预测效率，且随着训练数据量的增加，模型的预测效果逐渐趋于稳定。彭燕[25]等人先对股票数据进行了插值，小波降噪，归一化处理后输入不同层数相同神经元数和相同层数和不同神经元数的LSTM网络中进行预测效果的对比实验，找到最佳的层数和神经元数量，最终结果显示模型的预测准确率提高了约30%。

还有许多学者选择将深度学习与其他技术相结合，优化升级原始LSTM模型以得到更加准确的的预测拟合效果。欧阳红兵[26]等人等人提出了金融时间序列分析和波动分析与LSTM神经网络相结合的预测模型。以股票价格平均日指数为例，将构建的模型与支持向量机、MLP网络、K最近邻算法、广义自回归条件异方差模型四种模型的预测效果进行对比。结果表明，相比于其他四种模型，LSTM神经网络对时间序列数据中动态的长短期数据变化趋势能够进行更加精确的拟合和预测。若对股市的时间序列数据提前进行小波分解，LSTM神经网络对股价数据长短期趋势拟合精度将会进一步提升。张永安[27]等人提出了CEEMD-LSTM复合预测模型。CEEMD，即序列平稳化分解模块，可以逐级分解时间序列中不同尺度的波动或趋势,生成一系列不同特征尺度的本征模态函数(IMF)，将CEED输出的IMF用LSTM提取高维的抽象特征，最后综合各个IMF分量以及趋势项的预测值,得到最终的预测值。最后结果显示CEEMD-LSTM模型在多种误差函数下皆优于其他模型。任君[28]等人将SVM和LSTM与Lasso方法相结合，选取股票指标作为模型的输入变量，使用指数衰减法和网格搜索法优化SVM和LSTM网络中的参数，最后使用Lasso方法对变量进行筛选。最后实验结果显示该模型具有较其他模型更加优良的的抗风险能力和投资收益，且ELSTM-L模型对交易成本有着更高的宽容度。

集成学习

我们可以通过方差和偏差这两个参数来判断一个模型的好坏。偏差指模型的预测值与真实值的差距，方差指模型的预测值波动的范围。若一个模型的偏差很大，代表这个模型没有很好地捕捉到数据特征，没有很好地拟合数据，即欠拟合；若一个模型的方差很大，代表这个模型在不同训练集中表现差异较大，可能在某训练集中捕捉到了一些数据的特殊特征，即过拟合。当方差和偏差都比较小时，我们可以认为这个模型是比较强大的。

在机器学习中的有监督学习的算法中，我们期望得到一个稳定的，准确的，能适应所有情况的可靠模型，然而实际上出于种种原因，我们很难得到这样的模型。但是相比于得到一个强大准确的模型，获得若干个相比较而言并不是那么完美的，有偏好的（即只在某些情况下表现较佳）的模型要现实得多。而集成学习就是将这些弱模型通过某种方式组合起来，得到一个较为完美的强模型，它的思路是即便某些模型得出错误的预测，其它的模型也能纠正它们的错误。比较典型的算法有引导聚集发和增强学习法等。

引导聚集算法（Bagging）

“Bagging”一词是对“bootstrap aggregating”的简写，其中“bootstrap”又称自展法，是一种用小样本估计整体值的非参数方法。具体操作是从初始数据中有放回地抽取一定量的样本，通过对这些样本的计算得到统计量的置信区间。Bagging使用bootstrap方法从整体数据集中有放回地取出若干个数据集，针对每一个数据集训练模型，最后参考这些模型的预测结果得出答案。更具体地，对于分类问题，可以采用所有模型投票的方式得出答案；对于回归问题，对所有模型的结果求平均值。

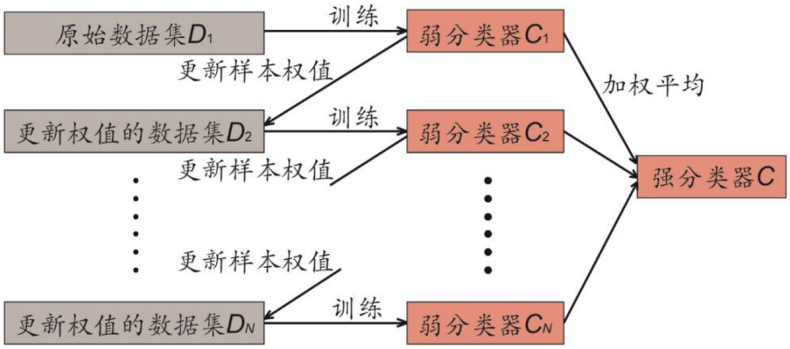
Bagging最大的优势是可以通过求平均的方式缩小基分类器的方差，从而减小过拟合。它的缺点是选取样本集的方法会带来一定量的偏差，影响最终结果。

上文中提及的随机森林算法就是用到了Bagging的思想。被应用在许多股市预测的模型当中。

增强学习算法（Boosting）

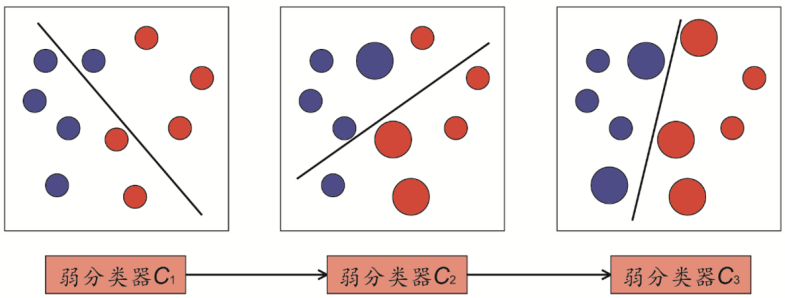
Boosting方法采用重赋权的方式迭代地训练基分类器。具体的操作是从训练集使用初始权重训练出第一个模型，再如图5.2.2中所显示，根据当前模型的错误数据更改训练集中样例的权值，训练第二个模型，以此类推，直到生成了n个弱学习器，再根据预测误差率将它们加权平均组合成一个强学习器，如图5.2.1中所示。相比于Bagging算法，Boosting算法能有效地降低模型的偏差。

图5.2.1 boosting算法的实现流程



资料来源：华泰证券交易所

图5.2.2 训练集权重调整过程



根据如何调整训练集的权重，以及如何组合基分类器，Boosting方法有很多不同的方式去实现，比较著名的有AdaBoost，GradienBoosting等算法。其中Adaboost是其中的成功代表，被称为数据挖掘十大算法之一[1]。

AdaBoost二元分类法的具体操作可分为以下几步：

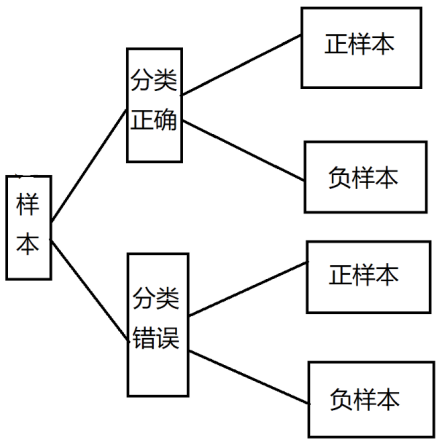
1. 我们令总数据集为T = {(x1 , y1) , (x2 , y2) ... (xn , yn)}，标记输出为{-1 , 1}，迭代次数为K
2. 初始化样本集的权值为W(1) = {w11 , w12 ... w1n}，其中w1i = 1/n，i = 1 , 2 ...n。
3. 执行多次迭代操作，令k = 1 , 2 ... K表示当前迭代数。

使用权重为D(k)的样本集对分类器进行训练，得到基本分类器Gk，并计算Gk(x)在该样本集上的分类错误率e(k) = P(Gk(xi) != yi) = $\sum_{i=1}^{n} w_{k,i} * I(G_k(x_i) != y_i)$。再根据这个分类错误率得到这个分类器在最终分类器中的重要程度 $\alpha_k = \frac{1}{2}\log\frac{1-e_k}{e_k}$，其中可以看出，$e_k$小于二分之一时$a_k$为正，且$a_k$随$e_k$的减小而增大，符合“越强的基分类器在最终分类器中作用越大”的思路。最后，我们还要根据当前分类器的错误情况更新训练集的权值分布W_{k+1}。 $w_{k+1,i} = \frac{w_{ki}}{Z_k}\exp(\alpha_k y_i G(x_i))$。其中$Z_k$是归一化因子，使$W_{k+1}$变为对应的概率化分布。通过这一函数，使得原本被分类错误的样本在下一次训练中具有更高的权重，相对的，正确的样本权重减小，使得AdaBoost方法能够“集中注意”在比较难以分类的样本上。

4. 最后按加权平均组合所有的基分类器 $f(x) = \sum \alpha_k * G_k(x)$

综上，我们可以看出AdaBoost在二元分类的问题上颇有建树，所以它常常可以和其他算法如决策树以及SVM等一起使用。而在龚利琴在2018年的研究中，认为简单地区分正确和错误在金融领域是不够的。在金融界中，人们对损失往往比对相同数额的收入更为敏感，所以对于分类正确和分类错误的样本继续进行一个正负样本的区分，如图5.2.3所示。进而使用付中良提出的多分类敏感问题代价敏感的AdaBoost算法[2]，实现Alpha策略选股[3]。持此之外，华泰证券[4]也对Boosting系列算法在金融预测方面的能力进行了评估，认为在同等预测水平下，Boosting的复杂程度要远低于Bagging，表现较为优异。

图5.2.3 分类方法



总结与展望

本文简要阐述了包括决策树，支持向量机，神经网络，集成学习等一系列经典人工智能算法的原理，并适当地介绍了一些基于这些算法进行改进的特殊模型，并解释了他们在股市预测领域的应用。在完成本片文献的过程中，本组组员们分工合作，每人对应其中一个研究方向进行较为有深度的学习,查找了大量文献,可谓是收获颇丰。除了对算法本身有了更深刻的理解之外，我们对于金融股市的预测也有了进一步的了解：金融大数据的特点是瞬息万变，不确定性高，而机器学习恰恰能够适应这样的数据，相比于使用传统的回归模型，机器学习不但能提高预测精度，还可以将更多文本数据，新闻事件等信息加入数据集中，大大拓展了数据类型。

然而，现如今人工智能在股市预测方面的应用尚在初级阶段，存在着对公式的过度依赖问题，不能够很灵活地利用模型，导致得到的预测结果泛化能力不足的问题出现。未来需要进一步提高模型预测的精度，挖掘更多不同类型的数据，另外，对于机器学习当中参数设定和预测结果的经济学理论解释也是未来的研究难题。

参考文献

[1] 吴玉霞,温欣.基于ARIMA模型的短期股票价格预测[J].统计与决策,2016(23):83-86.DOI:10.13546/j.cnki.tjyjc.2016.23.051.

[2] Alt, R., Puschmann, T. (2012). The rise of customer-oriented banking - electronic markets are paving the way for change in the financial industry. Electronic Markets, 22(4), 203-2015.

[3] 顾文涛,王儒,郑肃豪,杨永伟.金融市场收益率方向预测模型研究——基于文本大数据方法[J].统计研究,2020,37(11):68-79.DOI:10.19343/j.cnki.11-1302/c.2020.11.006.

[4] 赵明清，吴圣强. 基于微博情感分析的股市加权预测方法研究[J]. 数据分析与知识发现，2019，3(2) : 43 - 51.

[5] Deepak, K., Kumar, S.P., & Rajit,V. (2021).A systematic review of stock market prediction using machine learning and statistical techniques[J]. Materials Today: Proceedings

[6]刘利，何先平． 基于遗传算法和模糊决策树的时间序列预测模型［J］．计算机工程与设计，2008 (19) : 5044 － 5046.

[7]王领,胡扬．基于C4.5决策树的股票数据挖掘［J］．计 算机与现代化，2015(10): 21－24

[8]王禹，陈德运，唐远新．基于Cart决策树与 boosting 方法的股票预测［J］．哈尔滨理工大学学报，2019,24(6) : 98 －103.

[9]林晓明. 金工: 人工智能选股之随机森林模型[R]. 华泰证券研究报告,2017。

[10]邓晶，李路． 参数优化随机森林在股票预测中的应用［J］．软件，2020，41(1) : 178 － 182.

[11]李航，统计学习方法，北京:清华大学出版社，2012:107-146.

[12]Deepak, K., Kumar, S.P., &Rajit,V.(2021).A systematic review of stock market prediction using machine learning and statistical techniques[J]. Materials Today: Proceedings

[13]胡迪,黄巍.基于AP-SVM组合模型的股票价格预测[J].武汉工程大学学报,2019,41(03):296-302.

[14]程昌品,陈强,姜永生.基于ARIMA-SVM组合模型的股票价格预测[J].计算机仿真,2012,29(06):343-346.

[15]张伟,李泓仪,兰书梅,张洁.GA-SVM对上证综指走势的预测研究[J].东北师大学报(自然科学版),2012,44(01):55-59.DOI:10.16163/j.cnki.22-1123/n.2012.01.014.

[16]葛继科,邱玉辉,吴春明,蒲国林.遗传算法研究综述[J].计算机应用研究,2008(10):2911-2916.

[17]Frey B J, Dueck D. Clustering by passing messages between data points[J]. science, 2007, 315(5814): 972-976.

[18]WhiteH.Economicpredictionusingneuralnetworks: The case of IBM daily stock returns[J]. Earth Surface Processes & Landforms，1988(5)

[19]蔡红，陈荣耀. 基于 PCA - BP 神经网络的股票价格预测研究[J]. 计算机仿真，2011，28(3) : 365 - 368.

[20]王文波，费浦生，羿旭明. 基于 EMD 与神经网络的中国股票市场预测[J]. 系统工程理论与实践，2010，30(6) : 1027 - 1033.

[21]杨进，陈亮. 基于小波神经网络与 ARIMA 组合模型在股票预测中的应用[J].经济数学，2018，35(2):62-67.

[22]郭怡然，王秀利. 基于 BP 神经网络的股市大小盘风格 轮动预测[J]. 计算机仿真，2019，36(3) : 239 - 242.

[23]綦方中，林少倩，俞婷婷. 基于 PCA 和 IFOA - BP 神经 网络的股价预测模型[J]. 计算机应用与软件，2020，37(1) : 116 - 121 + 156.

[24]陈卫华. 基于深度学习的上证综指波动率预测效果比较研究[J]. 统计与信息论坛,2018,33(5):99-106.

[25]彭燕，刘宇红，张荣芬. 基于 LSTM 的股票价格预测建 模与分析[J]. 计算机工程与应用，2019，55(11) : 209

[26]欧阳红兵，黄亢，闫洪举. 基于 LSTM 神经网络的金融时间序列预测[J]. 中国管理科学，2020，28(4) : 27 -35.

[27]张永安，颜斌斌.一种股票市场的深度学习复合预测 模型[J]. 计算机科学，2020，47(11) : 255 - 267.

[28]任君，王建华，王传美，等. 基于 ELSTM - L 模型的股票预测系统[J].统计与决策，2019，35(21) : 160 - 164.

[29]张世杰. 大数据下的机器学习在股市预测中的应用[J]. 贵阳学院学报:社会科学版, 2021, 16(4):6.

[30] Zhou Z H, Yang Y, Wu X D, Kumar V, The Top Ten Algo-rithms in Date Mining New York, USA: CRC Press, 2009, 127149

[31] 付中良. 多分类问题代价敏感AdaBoost算法[J]. 自动化学报, 37(8):973-983, 2011

[32] 龚利琴. 基于AdaBoost算法的Alpha组合研究[D]. 郑州大学,2018.

[33] 林晓明. 金工:人工智能选股之Boosting模型[R]. 华泰证券研究报告,2017