支持向量机（SVM）是一种二分类模型，其基本模型为定义在特征空间上的、间隔最大的线性分类器，其中的核技巧（核函数）将数据向高维空间映射，得以用线性分类学习方法高效解决非线性问题，使它实质上成为非线性分类器。它的学习策略为间隔最大化，可形式化为求解凸二次规划的问题，但其易出现局部最优解问题[1]。

股票数据非线性、高噪声、波动性较强且需要对多特征数据进行分析。针对这样的特性，在股市预测中，SVM的独有优势是能利用核函数，通过非线性映射将股票数据映射到高维空间，从而利用线性函数对股票的涨跌进行分类。

支持向量机中虽然存在大量已构造出的核函数可供使用，研究者可以根据实际需要选择适合的核函数进行应用。但是，如何核函数选取的问题目前还没有太好的解决方案。同时，当该方法应用于大规模训练样本时会耗费大量的计算机内存和运算时间，它在大数据时代股票预测中的发展空间因此受到一定程度的限制。

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。支持向量机与包括时间序列算法（Arima）、近邻传播（AP）、遗传算法（GA）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。

**1.近邻传播-支持向量机（AP-SVM）**

近邻传播（AP）于2007年由Frey和Dueck在Science期刊中提出，其最不同于以往建立数据簇的方法在于将每个实值数据点都视为潜在的聚类中心，使得模型对数据的初始值不敏感，效率提高，方差减小。它用负欧几里得距离表示样本点相似度（s(i,k)），原理是通过数据交换更新相似度矩阵中每个点的吸引度信息（r(i,k)，数据k适合作为i中心点的程度）、归属度信息（a(i,k)，数据i选择k作为中心点的程度），再对样本点的吸引度信息和归属度信息求和，检测其选择聚类中心的决策，经过若干次迭代后收敛成若干聚类[7]。

胡迪（武汉理工大学）创新性地将近邻传播与支持向量机结合运用于A股预测。该模型的创新点在于很好地利用了相似走势股票与待预测股票的相关性，使用近邻传播（AP）算法挑选出与待预测股票价格变化相似度较高的其他股票, 然后将目标股票和与其价格变化相似的其他股票一起作为输入数据，有效摆脱了传统股市预测易出现的、仅针对单支股票分析的局限性，据此训练了一个支撑向量机[3]。

其模型建立过程如下：

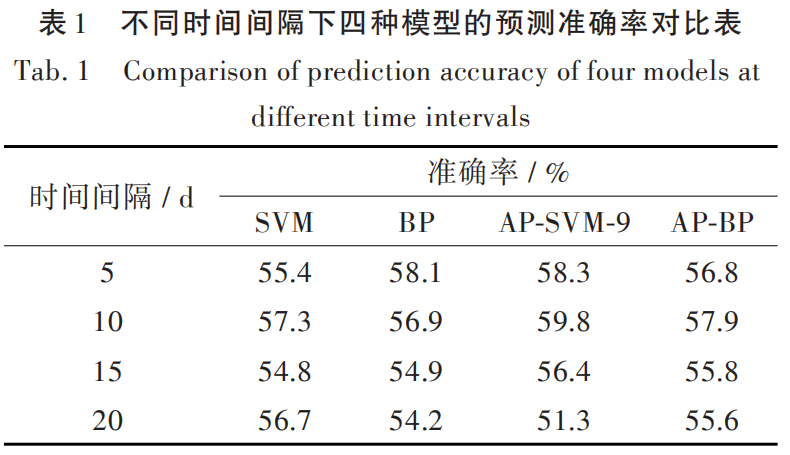
（1）特征选择

利用概率统计中相关系数的计算公式，计算A股主要指标：开盘价、收盘价、最高价、最低价、成交量和成交额的相关性，剔除相关性强的冗余特征值，最终保留收盘价和成交量作为两个输入特征。

（2）聚类相似度度量构建股票矩阵簇

对收盘价和成交量两个特征分别进行二范数归一化（使数据向量落在一个球面上，从而能用负欧几里得距离刻画数据间相关性），据此构建在这两个维度形成聚类的矩阵簇，利用相似走势的股票相关性

（3）对簇内股票进行样本标签制作，训练SVM。



**2.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。该模型用时间序列算法捕捉股票价格的线性规律，用支持向量机捕捉股票价格的非线性规律，从而创新性地用基于小波分解的时间序列算法-支持向量机组合模型对股票收盘价格这一典型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

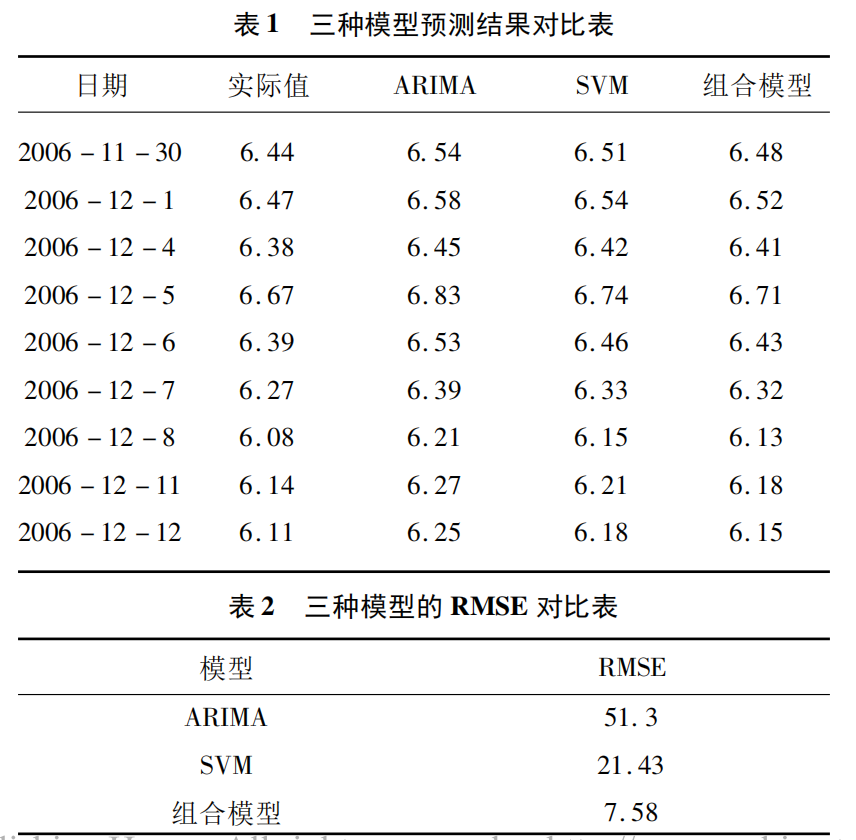
首先利用Mallat算法（二抽取是该算法的原理，信号数据会在每层分解后比分解前的长度减半，而总输出数据长度与输入数据长度保持一致。经Mallat算法分解后的信号可进行二插值重构），对非平稳时间序列进行分解和重构得到低频和高频信息，分别对应线性部分和非线性部分。

（2）线性部分平稳化处理后用于构建ARIMA模型

高频信号重构后可近似看作平稳时间序列，可建立 ARIMA 模型预测股票价格历史数据。

（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长期趋势具有非线性，可采用 SVM 模型对其非线性规律进行建模和预测。该模型中SVM训练使用的核函数为高斯函数，在MATLIB7.0平台下自编程调用LIBSVM工具箱来实现SVM建模，采用5折交叉验证，经gridre－gression.py 自动搜索确定模型最优参数[4]。



**3.遗传算法-支持向量机（GA-SVM）**

遗传算法最早于1969年由Holland教授（美国Michigan大学）提出, 后经DeJong、Goldberg等人归纳总结而成为一类模拟进化算法。它是一类借鉴自然界生物进化和遗传机理而发展起来的自适应全局优化随机搜索算法，其本质是一种不依赖具体问题的、自组织、自适应的直接搜索方法，其底层思想来源于达尔文的进化论、魏茨曼的物种选择学说和孟德尔的群体遗传学说，具有坚实的生物学基础。它作为经典高性能模型，计算、建模、寻优的方法渐趋成熟[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。该模型利用遗传算法（GA）的全局自动寻优能力先智能地找到SVM的最佳参数, 再剔除了冗余特征, 提高了预测的命中率, 大大减少了计算量和预测时间[5] [6]。

其模型建立过程如下：

（1）选定原始特征向量

每一个特征被定义为一位基因，用一个长度为特征个数的二进制串表示所有特征，该二进制串为染色体（若染色体某位为1，则表示该位特征被选中，反之则表示该位特征被屏蔽，每一个染色体代表不同的特征子集）。

（2）对参数进行优化，获得最优尺度参数σ和惩罚因子Ｃ

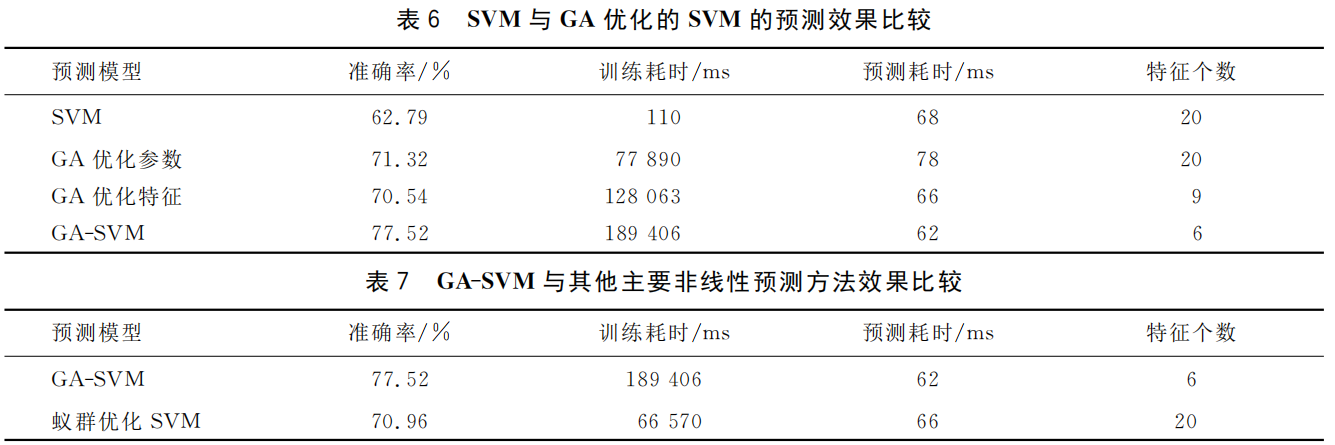
首先按经验设定尺度参数σ和惩罚因子Ｃ的范围，此后进行离散化及二进制编码。如σ∈(0,10)，步长为10/1024，得到尺度参数σ的二进制串为Ｘ＝x1x2x3…x9x10惩罚因子Ｃ∈(0,100)，步长为100/1024，得到Ｙ＝y1y2y3…y10，于是染色体为XY＝x1x2…x10y1y2…y10。遗传算法寻优结束后根据参数与二进制编码间的映射关系ｆσ＝fσ·Ｘ，Ｃ＝fＣ·Ｙ进行解码，将染色体转换成实际的尺度参数σ和惩罚因子Ｃ。

（3）对特征向量进行优化，获得最佳特征向量集

（4）用改进后的SVM对测试集进行测试，得到准确率。

根据股票数据特点，该股票采用径向基核函数。

将所训练模型实际运用于股票预测，并与统计模型、时间序列模型方法、神经网络进行对比实验后使用均方根误差 （RMSE）作为模型的评价指标证明，GA-SVM相较其他几种方法除训练耗时较长外均具有优势。



[1]李航，统计学习方法，北京:清华大学出版社，2012:107-146.

[2]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

[3]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.

[4]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.

[5]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.

[6]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.

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