基于q阶正交模糊孪生支持向量机的股票预测

白玉景

（河北工程大学数理科学与工程学院　056002）

摘　要:股票的价格受经济、政治、公司经营状况和市场等多方面的复杂影响,且各因素自身具有模糊性和混乱性,再加之股票市场是一个非线性的系统,所以股票的价格数据存在着多噪声多离群点的特征,因而需要找到一种更好的方法解决该问题。为了在一定程度上解决了数据存在噪声的问题,本文提出了q阶正交模糊孪生支持向量机,可以更好的解决数据中存在噪声和离群点的问题,并将该模型应用于股票价格预测中,通过对算法分类效果和分类精度的分析,证实该算法在股票趋势预测中具有一定的作用。

关键词:q阶正交模糊孪生支持向量机;股票预测;q阶正交模糊集;噪声

一、引言

发行股票是上市公司融通资金的重要手段,股票市场的成熟程度影响着企业经营发展,进而影响经济的发展。股票市场不仅是上市公司融通资金的场所,买卖股票也是投资的重要手段,因此,自股票诞生之日起,对股市涨跌变化的预测方法就层出不穷。就中国市场而言,基本面分析及技术分析是主流方法,但预测的效果并不十分理想。加之经济大环境的阴霾,不论是机构投资者还是个人投资者,总体而言,在股市的投资收益并不理想,这更加吸引中国学者对股市走势预测的研究兴趣。

传统的股票价格研究是以简单的数学模型为基础进行研究的,金融学者们最初运用简单的线性模型来处理股票数据,如:简单自回归模型、简单滑动平均模型,利用单位根检验来验证时间序列的平稳性,将非平稳的时间序列进行差分运算转化。Adebiyi等人基于纽约证券交易所和尼日利亚证券交易所的股票数据尝试运用ARIMA模型预测股票价格,结果显示该模型具有短期预测潜力。但由于股票数据包含大量噪声及不确定性因素,随着预测周期的变长,线性模型的自身局限性不断凸显。科学家们尝试使用非线性模型进行研究,引入了神经网络、支持向量机等机器学习方法,并成功运用到股票预测之中。例如:张健等研究了人工神经网络在股票分析预测中的应用,并试图设计新的网络。张晨希等使用支持向量机预测上市公司股票走势,并证明优于传统神经网络。邹阿金等构建了新型的Legender神经网络,并证明可以很好地逼近非线性系统。李悦根据对支持向量机和股票价格预测问题的研究,提出了基于支持向量机的股票价格预测模型,预测模型在求解过程中将支持向量分类算法的决策函数连续化,从而能够获得较好的股票价格预测结果。李峥嵘等将Relief算法与加权支持向量机相结合对股票的价格涨跌进行预测研究,并以股票数据验证了模型的可行性和准确性。

支持向量机作为统计学习中重要的机器学习方法之一,相比于其它机器学习算法,支持向量机在解决小样本问题时具有很好的泛化能力。然而,当数据量较大时,求解支持向量机是很困难的。为了解决数据量大所带来的问题,研究人员在深入研究标准支持向量机的基础上提出了孪生支持向量机,将一个具有二类问题的支持向量机转化为求解两个较小规模的优化问题,在训练性能上得到了进一步的改善。然而,这些方法并没有考虑到不同数据样本点对最优超平面所产生的影响,而是同等对待所有的训练数据样本来构造最优分类面,为了解决这一问题,一些研究人员在孪生支持向量机的基础上,通过考虑不同数据样本点对最优分类面的影响,从而降低了噪声对最优分类面的影响,增强了其抗噪的特性。2015年, Gao等人将梯度下降法引入到模糊孪生支持向量机中,不仅具有传统孪生支持向量机的优点,而且适用于处理高维输入数据所带来的计算复杂等问题,缩短了计算时间。2013年,哈明虎等人提出了基于直觉模糊数和核函数的支持向量机。在高维特征空间中,利用核函数为每个训练点分配一个对应的直觉模糊数,然后引入直觉模糊数的得分函数来度量各训练点的贡献,最后根据每个训练点的得分,构造新的支持向量机,该方法降低了噪声和离群点对分类的影响。2019年, Rezvani等人将直觉模糊支持向量机的思想引入到孪生支持向量机中,提出了直觉模糊孪生支持向量机(Intuitionistic Fuzzy Twin Support Vector Machines, IFTSVM), IFTSVM中隶属度与非隶属度的引入降低了输入数据中噪声点与离群点对分类的影响,最小化了新的结构风险提高了分类的精度。模糊理论的提出,也为具有随机性和不确定性特点的金融时间序列建模提供了新的角度。2019年, Li等人等人改进了正交模糊集的得分问题,提出了一个新的得分函数,更进一步地解决了正交模糊集中排序以及偏好关系问题。本文所提出的q阶正交模糊孪生支持向量机(q-Rung Orthopair Fuzzy Twin Support Vector Machine,q-ROFTSVM)结合了处理不确定信息能力的模糊逻辑,以及具有较强的数据处理能力和泛化能力的孪生支持向量机两种技术,可有效发挥其各自的优势,并弥补其不足,集学习、识别、自适应及模糊信息处理于一体,提高整个分类系统的学习能力和表达能力,从而更好的对股票价格趋势进行预测。

二、q阶正交模糊孪生支持向量机(q-ROFTSVM)

(一)隶属度与非隶属度的分配

1、样本隶属度的确定

使用在高维特征空间中训练样本点到类中心的距离来设

置隶属函数，类似于直觉模糊孪生支持向量机的设置方法。对于每一个训练样本点，隶属度可以表示成：

$$\mu(x_i)=\begin{cases}1-\dfrac{\left\|\phi(x_i)-C^{+}\right\|}{r^{+}+\delta} & y_i=+1\\ 1-\dfrac{\left\|\phi(x_i)-C^{-}\right\|}{r^{-}+\delta} & y_i=-1\end{cases}$$

其中，δ＞0为可调节参数，r^{+}，r^{-}分别表示正类与负类训练样本的半径，C^{+}，C^{-}分别表示正类与负类训练样本的类中心，$\phi(x_i)$表示高维特征空间中的输入样本。

2、样本点非隶属度的确定

训练样本点非隶属的程度越大，则该样本属于此类的程度就越大小，训练点的非隶属程度是由其邻域内异类点的数量与该邻域内所有点的数量之间的比例来确定的。

$$\nu(x_i)=\frac{\left|\left\{x_j\mid\left\|\phi(x_i)-\phi(x_j)\right\|\le\alpha,y_j\ne y_i\right\}\right|}{\left|\left\{x_j\mid\left\|\phi(x_i)-\phi(x_j)\right\|\le\alpha\right\}\right|}$$

其中，α＞0为可调节参数。

q-ROFTSVM相较于IFTSVM取消了非隶属度的设置中隶属度的限制问题，重新对隶属度和非隶属度做了作了定义与约束，约束条件为：

$$0\le\mu^{q}(x_i)+\nu^{q}(x_i)\le1$$

其中，q≥1为正整数，q的选取为满足上述约束条件的最小正整数值。

3、训练样本点得分函数的确定

当非隶属度为0时，$s_i=\mu_i$；

当隶属度≤非隶属度时，$s_i=0$；

其它情况时，

$$s_i=\frac{((1-\nu_i)^{q}+1-\nu_i^{q})^{1/q}}{((1-\nu_i)^{q}+1-\nu_i^{q})^{1/q}+((1-\mu_i)^{q}+1-\mu_i^{q})^{1/q}}$$

（二）q阶正交模糊孪生支持向量机模型构建

假设训练样本点的集合为：$T=\{(x_1,y_1),(x_2,y_2),...,(x_i,y_i)\}$，$i=1,2,\ldots,N$，训练样本分为两类，A代表正类样本点，B代表负类样本点。

1、线性模型构建

两类样本点将通过两个非平行的超平面进行分类：

$\omega_1\cdot x_i+b_1=0$ 和 $\omega_2\cdot x_i+b_2=0$

本文通过在最大化间隔的思想下增加正则化项最小化结构风险来构建模型。原始问题如下：

$$q\text{-FOFTSVM1:}\quad\min_{\omega_1,b_1,\xi_2}\ \frac{1}{2}\left\|A\omega_1+e_1b_1\right\|^{2}+\frac{1}{2}c_1\left\|\omega_1\right\|^{2}+c_3s_2^{T}\xi_2$$
$$s.t.\quad-(B\omega_1+e_2b_1)+\xi_2\ge e_2,\xi_2\ge0 \qquad (3\text{-}1)$$

$$q\text{-FOFTSVM2:}\quad\min_{\omega_2,b_2,\xi_1}\ \frac{1}{2}\left\|B\omega_2+e_2b_2\right\|^{2}+\frac{1}{2}c_2\left\|\omega_2\right\|^{2}+c_4s_1^{T}\xi_1$$
$$s.t.\quad(A\omega^{(2)}+e_1b^{(2)})+\xi_1\ge e_1,\xi_1\ge0 \qquad (3\text{-}2)$$

其中，c_1,c_2,c_3,c_4 为正的惩罚参数，ξ_1,ξ_2 表示松弛变量，e_1,e_2 为合适维度的列向量。

为了方便计算，可以通过构造拉格朗日函数的方式来得出公式(3-4)和(3-5)的对偶问题。

$$L(\omega_1,b_1,\xi_2,\alpha,\beta)=\frac{1}{2}\left\|A\omega_1+e_1b_1\right\|^{2}+\frac{1}{2}c_1\left\|\omega_1\right\|^{2}+c_3s_2^{T}\xi_2$$
$$+\alpha[(B\omega_1+e_2b_1)-\xi_2+e_2]-\beta\xi_2$$

其中，α,β 是拉格朗日乘子，并且结合KKT条件，可以得到：

$$\begin{pmatrix}A^{T}\\e_1^{T}\end{pmatrix}\begin{pmatrix}A & e_1\end{pmatrix}\begin{pmatrix}\omega_1\\b_1\end{pmatrix}+\begin{pmatrix}B\\e_2\end{pmatrix}\alpha=0 \qquad (3\text{-}3)$$

令 $H_1=[A\ \ e_1]$，$H_2=[B\ \ e_2]$，$u_1=\begin{pmatrix}\omega_1\\b_1\end{pmatrix}$，

等式(3-3)可以重新表述为：

$$H_1^{T}H_1u_1+H_2^{T}\alpha=0\Rightarrow u_1=-(H_1^{T}H_1)^{-1}H_2^{T}\alpha$$

为方便求解，我们引入单位矩阵I，使得：

$$u_1=-(H_1^{T}H_1+c_1I)^{-1}H_2^{T}\alpha$$

同理可得：

$$u_2=-(H_2^{T}H_2+c_2I)^{-1}H_1^{T}\beta$$

由上述推导我们可以得出(3-1)和(3-2)的对偶问题分别为：

$$q\text{-FOFTSVM1:}\quad\min_{\alpha}\ \frac{1}{2}\alpha^{T}H_2(H_1^{T}H_1+c_1I_1)^{-1}H_2^{T}\alpha-e_2^{T}\alpha$$
$$s.t.\quad0\le\alpha\le c_3s_2 \qquad (3\text{-}4)$$

$$q\text{-FOFTSVM2:}\quad\min_{\beta}\ \frac{1}{2}\beta^{T}H_1(H_2^{T}H_2+c_2I_2)^{-1}H_1^{T}\beta-e_1^{T}\beta$$
$$s.t.\quad0\le\beta\le c_4s_1 \qquad (3\text{-}5)$$

当求出最优解时，两个非平行的最优超平面也就确定了。最后，当一个新的输入数据出现时，我们可以通过下列式子可以判断出新的数据属于正类还是负类：

$$f(x)=\arg\ \min_{i\in1,2}\left\{\frac{\left|\omega_1^{T}x+b_1\right|}{\left\|\omega_1\right\|},\frac{\left|\omega_2^{T}x+b_2\right|}{\left\|\omega_2\right\|}\right\}$$

2、非线性模型构建

为了解决非线性分类问题，我们引入核函数将训练样本点映射到高维特征空间中进行分类。两个非平行超平面分别为：

$$K(x,C^{T})\omega_1+b_1=0,K(x,C^{T})\omega_2+b_2=0$$

其中，$C=[A;B]$ 表示所有训练样本，$K(x_1,x_2)=(\phi(x_1),\phi(x_2))$ 为核函数。

对于非线性问题，原始问题为：

$$K\text{-}q\text{-FOFTSVM1:}\quad\min_{\omega_1,b_1,\xi_2}\ \frac{1}{2}c_1\left\|\omega_1\right\|^{2}+\frac{1}{2}\left\|K(A,C^{T})\omega_1+e_1b_1\right\|^{2}+c_3s_2^{T}\xi_2$$
$$s.t.\quad-(K(B,C^{T})\omega_1+e_2b_1)+\xi_2\ge e_2 \qquad (3\text{-}6)$$

$$K\text{-}q\text{-FOFTSVM2}:\quad \min_{\omega_2,b_2,\xi_1}\ \frac{1}{2}c_2\|\omega_2\|^2+\frac{1}{2}\|K(B,C^T)\omega_2+e_2b_2\|^2+c_4s_1^T\xi_1$$
$$s.t.\quad (K(A,C^T)\omega_2+e_1b_2)+\xi_1\geq e_1$$

（3-7）

同样为了方面计算，可以类似于线性模型中转化为对偶问题进行求解。

首先可以得到：

$$u_1^*=-(S_1^TS_1+c_1I)^{-1}S_2^T\alpha,u_2^*=-(S_2^TS_2+c_2I)^{-1}S_1^T\beta$$

其中，

$S_1=[K(A,C^T)\ e_1]$，$S_2=[K(B,C^T)\ e_2]$，$u_1^*=\begin{pmatrix}\omega_1\\b_1\end{pmatrix}$，$u_2^*=\begin{pmatrix}\omega_2\\b_2\end{pmatrix}$。

同理线性模型的推导我们可以得出(3-6)和(3-7)的对偶问题分别为：

$$K\text{-}q\text{-FOFTSVM1}:\quad \min_{\alpha}\ \frac{1}{2}\alpha^TS_2(S_1^TS_1+c_1I_1)^{-1}S_2^T\alpha-e_2^T\alpha$$
$$s.t.\quad 0\leq\alpha\leq c_3s_2$$

（3-8）

$$K\text{-}q\text{-FOFTSVM2}:\quad \min_{\beta}\ \frac{1}{2}\beta^TS_1(S_2^TS_2+c_2I_2)^{-1}S_1^T\beta-e_1^T\beta$$
$$s.t.\quad 0\leq\beta\leq c_4s_1$$

（3-9）

当求出最优解时，两个非平行的最优超平面也就确定了。

最后，当一个新的输入数据出现时，我们可以通过下列式子可以判断出新的数据属于正类还是负类：

$$f(x)=\arg\ \min_{i\in1,2}\left\{\frac{|K(x,C^T)\omega_1^T+b_1|}{\sqrt{\omega_1^TK(A,C^T)\omega_1}},\frac{|K(x,C^T)\omega_2^T+b_2|}{\sqrt{\omega_2^TK(B,C^T)\omega_2}}\right\}$$

三、数据选取与预处理

（一）数据来源

本文所用股票数据均来自于软件大智慧，选取了浦发银行、中信证券、白云机场等5支股票2018年12月-2020年8月的价格数据，以每日的开盘价、最高价、最低价、收盘价、成交额、成交量，5日、10日、20日、30日、60日、120日价格均线，共计12个指标作为特征属性。

（二）数据预处理

首先对股票数据设定标签，若第二天的开盘价大于等于今天的收盘价，则判定今天股票下一日价格是上涨的，数据的标签为1，若第二天的开盘价小于今天的收盘价，则判定今天股票下一日的价格是下跌的，数据的标签为0；然后对每一支股票的数据进行归一化，将所有属性的值转化成[0，1]之间，特征属性的值均值为1方差为0的数，然后将每一支数据分成两份，其中80%作为训练数据集，20%作为验证数据集。

对于模型所用的核函数，本文选择最常用的高斯径向基核函数：

$$K(x,z)=\exp(-\|x-z\|^2/2\sigma^2)$$

其中σ为参数。

我们在解决最优化问题时使用MATLAB优化工具箱中的quadprog函数，由于没有明确的寻找参数σ和惩罚参数c最优值的方法，所以采用“启发式”方法进行搜寻。主要的思想是利用交叉验证和网格搜索的方法，从一个较大的取值范围内找到使交叉验证所得结果最好的一组参数，对于不同的数据，我们分别对参数值进行寻优。

四、实验分析

首先我们选取600000浦发银行的股票数据进行分析，收益率趋势如下图所示：

从浦发银行收益率折线图中，我们可以看出，收益率在零附近波动，且波动幅度较大，因此在这些数据中易存在不易区分的噪声数据，将数据的特征属性的值代入到模型中进行实验，首先对参数进行寻优，通过交叉验证和网格寻优，我们选取

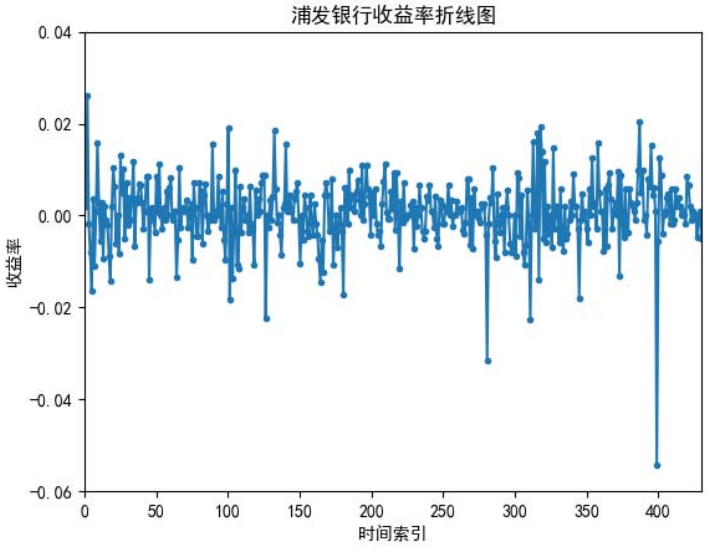

图4-1　浦发银行收益率折线图

σ=0.1，惩罚参数c=100，可以得出以下结果：

表1 浦发银行股票数据模型预测结果对比

股票数据	TSVM	IFTSVM	q-ROFTSVM
浦发银行	61.24%	71.65%	75.37%

可以看出，本文所提出的模型在浦发银行股票数据上的应用效果相对比较好。为了验证本文提出模型的优劣性，我们再随机选取五只股票数据进行预测，预测结果如下：

表2 股票数据模型预测结果对比

股票数据	TSVM	IFTSVM	q-ROFTSVM
科大讯飞	74.27%	77.31%	78.25%
中信证券	77.52%	82.17%	82.95%
三一重工	79.84%	84.5%	80.62%
白云机场	68.75%	72.66%	76.56%

结果分析：由表2可以看出，q-ROFTSVM模型的预测准确率均达到75%～85%之间，且与参照模型相比，准确率有所提高。为了避免单只股票的随机性，本文选取了不同行业的4支股票进行对比实验，可以看出，其中有3支股票相对于已有模型准确率有一定提高。因此，说明本文的改进是有效的。

五、结论与建议

本文考虑到不同样本点对模型效果的影响不同，将q阶正交模糊集与孪生支持向量机相结合，将提出的q阶正交模糊集孪生支持向量机应用于股票价格涨跌预测中，并通过对浦发银行等股票进行实证分析，验证了该模型的准确性和有效性。

股票市场受宏观、微观、政治环境等各方面因素的影响，复杂性很高。本文所提出的模型可以在一定程度上帮助投资者判断股票价格的未来趋势。本文所选取的为常用的12个指标属性，但对于股票数据来说，影响股票价格的指标属性远远超过12个，而且不同行业的股票价格也受整个社会大环境以及公司运营情况的影响，因此，对股票价格涨跌预测本文可作为一个参考。

【参考文献】

[1]Ayodele Adebiyi A, Aderemi O.Adewumi Charles K.Ayo.Stock price prediction using the ARIMA model[C].2014 UKSim-AMSS 16th International Conferenceon Computer Modelling and Simulation. Cambridge:IEEE, 2014

[2]张健，陈勇.人工神经网络之股票预测[J].计算机工程，1997（2）

[3]张晨希，张燕平，张迎春等.基于支持向量机的股票预测[J].计算机技术与发展，2006（6）

[4]邹阿金，罗移祥.Legender神经网络建模及股票预测[J].计算机仿真，2005（11）

[5]李悦.基于支持向量机的股票价格预测[D].南开大学，2013

[6]李峥嵘，韦增欣，祝人杰.基于Relief-WSVM的股票预测研究[J].中国管理信息化，2020（11）

[7]Cortes C, Vapnik V.Support vector networks[J]. Machine Learning, 1995, 20（3）

[8]V.N.Vapnik, Statistical learning theory[M].1998

[9]Jayadeva, Khemchandani R, Chandra S.Twin support vector machines for pattern classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007（5）

[10]Gao BB, Wang JJ, Wang Y, Yang CY.Coordinate descent fuzzy twin support vector machine for classification[J].2015 IEEE 14th International Conference on Machine Learning and Applications（ICMLA）, 2015

[11]Ha MH, Wang C, Chen JQ.The support vector machine based on intuitionistic fuzzy number and kernel function[J].Soft Computing, 2013（4）

[12]Rezvani S, Wang X, Pourpanah F.Intuitionistic fuzzy twin support vector machines[J].IEEE Transactions on Fuzzy Systems, 2019（11）

[13]Zadeh L A.Fuzzy sets[J].Information and control, 1965（03）

[14]Atanassov K T.Intuitionistic fuzzy sets[J].Fuzzy Sets and Systems, 1986, 20（01）

[15]Yager R.R., Abbasov A.M., Pythagorean membership grades, complex numbers and decision-making. International Journal of Intelligent Systems, 2013（05）

[16]Yager R.R.Generalized orthopair fuzzy sets[J]. IEEE Transactions on Fuzzy Systems, 2017（05）

[17]Li H, Yin S, Yang Y.Some preference relations based on q-rung orthopair fuzzy sets[J].Int J Intell Syst.2019