

基差点价中机器学习模型的运用

-以有色、黑色金属为例

随着技术的进步和金融数据海量的增加，机器学习在金融中的应用范围不断拓展。当前机器学习模型被广泛地应用于风险管理、算法交易、投资组合管理以及合规与反洗钱监管当中。本文尝试在当前大宗商品交易普遍采用的基差点价贸易中引用机器学习模型，以期帮助点价方在价格波动之前采取措施，有效降低潜在风险。本文以上期所期货品种：铜、铝、锌、铅、镍等有色金属的日度期现货数据为样本，选取影响有色金属价格的因子特征，采用随机森林模型对未来基差涨跌进行训练和预测。

一、机器学习与基差点价模式的介绍

基差点价模式的优势

基差点价模式被广泛应用于石油、金属、农产品等大宗商品贸易。基差点价模式一般指，买卖双方约定以某一个期货品种在某一时间点的价格或者一段时间的均价为基准，加上一定程度的升贴水，来最终确定现货结算价格的贸易方式。由于这些商品的价格波动性较大，通过期货市场锁定价格变化的趋势，有助于贸易双方在价格上找到共同点。虽然这种贸易方式本质上仍属于现货贸易定价一部分，但是因为衍生品的加入使得企业可以在市场波动中有效对冲风险，降低价格波动对现货交易的影响。基差点价允许买方或者卖方选择合适的时机“点价”，确定现货价格，这有助于企业根据市场趋势锁定更有利的价格。点价模式的优势在于改变传统的定价模式，即双方不在决定价格上相互博弈，而是研判未来市场价格的相对价格变化。

机器学习优势

在基差点价模式中，机器学习模型可以发挥着重要的作用。机器学习的优势在于它能够处理大量的、多维度的数据。在基差点价的应用中，影响价格走势的因素众多，有基本面因素、技术面因素和宏观因素。同时基差存在较强的趋势和周期性行情，例如临近交割月时，现货和期货价格的趋同等。而机器学习模型，尤其是时间序列模型和非线性模型，能够较好的识别历史数据中的规律。同时机器学习模型通过特征选择或降维技术（如PCA），可以过滤掉无关或冗余特征，从而提升模型的训练效率和预测效果。

二、有色金属价格影响因素概述

影响有色金属的价格的因素，复杂多变。总体可以分为三种，第一宏观因素，包括利率、汇率、海外风险、能源价格、经济周期、货币政策和通胀等。这些因素通过影响整体市场和经济运行，对有色金属的价格产生间接但重要的影响。第一基本面因素，包括仓单、社会库存、供需变化等因素。这些因素直接反映了金属市场的供给和需求状况，是决定价格的核心力量。技术面因素：布林带、移动均线、动量指标、反转指标等。有色金属价格的变化受到以上多种不可控因素的影响，数据中往往存在大量的噪声，而传统的线性回归模型分析方法往往难以高效处理这些多维信息。但机器学习模型，特别是深度学习和集成学习（如随机森林和梯度提升机），能够更好地处理这些噪声数据，机器学习可以通过特征选择和特征工程优化数据输入，提取关键特征来提高预测准确性。本文通结合技术指标、基本面、和宏观经济数据，以期在高波动性和不确定的市场环境中，可以更好地识别市场的的变化。

三、模型表现

本文使用2011年1月-2024年9月有色金属的期现数据，包含铜、铝、锌、铅、镍等有色金属的日度数据。对有色金属的数据使用随机森林模型进行训练和预测，目标是预测下一个交易日的日基差。

随机森林是基于决策树的集成学习方法, 通过结合多个决策树的结果来提高分类效果和稳定性。核心思想是通过随机采样和特征子集选择中构造 N 个不同的决策树, 并且通过投票来输出结果。分类公式: 随机森林中的每棵决策树 $h_t(x)$ (其中 $t=1,2,3...T$ 是树的编号) 根据输入数据 x 给出预测类别。假设有 T 颗树, 输出类别为 y , 则分类的公式为:

$$\hat{y} = \text{Mode}\{h_1(x), h_2(x), \dots, h_T(x)\}$$

其中 Mode 表示众数函数, 即返回出现次数最多的类别。通过随机选择特征和样本构建多棵决策树, 能有效处理过拟合问题, 适用于复杂的基差预测任务。同时该模型可以通过特征重要性分析, 能有效找出哪些因素对基差变动影响最大, 从而优化决策, 提高基差预测的准确度。

3.1 主流模型和算法的分类和简介

当前主流的机器学习算法中, 主要分为三类: 回归算法, 主要是针对连续性数据。第二类是分类算法, 主要用于离散型数据的预测, 第三类, 聚类算法, 用于无监督学习的数据分组, 例如市场客户分群。之前大部分对于基差的预测中, 主要将基差变化看为连续型变量, 采用多元线性回归对于未来价格进行预测, 但是由于线性回归的强假设, 导致预测结果不尽如人意。因此我们将基差未来走势简化为涨跌问题, 减少预测难度。将未来基差的预测当成分类问题, 根据时间序列中的基差值进行分类, 判断当前值相对于前一个时间点的基差是否上涨或下跌, 如果后一个时间点的基差值比前一个时间点高, 则标记为上涨(1), 反之则标记为下跌(0), 并以此作为目标变量。

随机森林算法是一种非线性模型, 之所以叫森林, 是因为将多个决策树的结果集成为森林。通过对于多个决策树的预测结果进行投票, 可以避免单个决策树受到极端值影响。每一个决策树根据样本特征(库存、动量、经济好坏等)对未来基差的涨跌进行判断, 最终对于 N 个决策树的结果采取投票形势确定最终的基差涨跌判断。该模型的决策思路即将每个决策树的预测结果进行投票, 未来涨跌的判断主要根据获得投票数最多的预测结果来进行。

3.2 样本数据的划分

样本数量总体有 14661 行数据, 34 列因子。在随机森林模型训练中, 我们使用 `scikit-learn` 中常见的 `train_test_split` 函数, 将数据集划分为训练集和测试集。按照 25% 的数据作为验证集, 75% 数据作为训练集。在划分训练集和验证集中, 通过打乱数据样本, 随机地划分数据样本。

标准化处理: 因为期货市场中容易出现极端的价格波动或者市场事件, 因此训练数据集中可能存在一些极端值。这些异常值会显著影响机器学习模型的训练效果, 因此需要极端值处理。另外不同因子特征的取值范围往往相差很大, 比如交易所库存特征可能是数量变化(取值在万到十万吨之间), 而技术指标例如动量特征可能是百分比变化(取值在 0 到 1)。这样的量纲差异可能导致模型在训练过程中对某些特征给予过多或过少的权重。因此需要统一量纲。

训练集划分、验证集划分: 使用过去 10 年的数据作为训练集, 因为 10 年的数据提供了足够长的历史信息, 确保模型能捕捉长期趋势和周期性。使用最近 4 年的数据作为验证集。验证集用于调整模型的超参数, 评估模型在未见过的数据上的表现, 避免过拟合。

决策树的设定: 这里设定为 200 棵树。因为更多的决策树可以带来更稳定的预测, 但劣势是训练时间也会相应增加。

3.3 评价指标

1. 准确率 (Accuracy = 0.5766)。准确率表示模型预测正确的样本占总样本的比例。在随机森林模型中，交叉验证后的准确率为 57.66%。这意味着在整个数据集中，大约 57.66% 的样本被正确分类。

2. AUC (AUC = 0.5998)。是另一个衡量模型性能的指标，它关注模型在区分不同类别 (上涨/下跌) 时的能力。AUC 取值范围为 0 到 1，数值越大，模型的区分能力越好。AUC = 0.5998 表明，当前模型的预测效果有一定能力，特别是区分上涨和下跌。

为了提高模型性能，后期准备作以下更多尝试：1、调整模型超参数：可以尝试扩大时间范围，尝试其他参数组合。2、增加特征处理：增加更多基于经济和市场规律的因子帮助提高模型的预测能力。3、数据预处理：进一步检查数据质量，比如平衡数据集或处理异常值。

四、本文总结

结合机器学习，在基差点价策略中引入随机森林算法不仅能够有效识别市场风险，还能提高策略优化的能力。通过精确的预测和动态调整，企业能够更好地应对价格波动，降低成本，减少市场价格波动带来的不确定性，在竞争中保持优势。总体来看，利用历史数据训练机器学习模型，可以相对准确的给出未来一段时间的基差相对走势。这可以辅助拥有定价主动权的一方在期货交易中把握基差的变化规律，在点价期内获取寻找较优势的价格，从而制定更有效的交易策略。

本文写作时间 2024 年 9 月 26 日

华安期货 投资咨询业务资格 证监许可[2011]1776

刘德勇 分析师 咨询从业资格：F03094242/Z0020048

初审：李伟 F0283072 /Z0010384

复审：夏雨辰 F3031745/Z0014542

终审：闫丰 F0251054/Z0001643

免责声明

本报告中的信息均来源于公开可获得资料，华安期货研究院力求准确可靠，但对这些信息的准确性及完整性不做任何保证，据此投资，责任自负。本报告不构成个人投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需要。客户应考虑本报告中的任何意见或建议是否符合其特定状况。