中央财经大学
Central University of Finance and Economics

本科生毕业论文（设计）

基于动态时间规整的异常交易识别

学生姓名：钟明昊

学　　号：2019311632

学　　院：管理科学与工程

专　　业：投资学

指导教师：刘志东

日　　期：2023 年 04 月 26 日

内 容 摘 要

异常交易行为破坏了金融市场的完整性与有序性。为了更好地监督投资者交易行为，为相关部门提供方法参考，文章基于多元动态时间规整(DTW)，结合极值理论进行阈值划分，提出了一种识别异常交易的非参数化的方法。在异常交易方面，结合了国内外的实证研究，文章参考了深圳证券交易所中对异常交易行为的定义，并使用拉抬打压股价、虚假申报等指标作为异常交易的特征判断依据。在模型评价与实证分析中，研究分析了两个证监会披露的异常交易案例，并能够有效地分辨出异常交易存在日期内的交易序列异常。同时，文章结合了以往研究，对 DTW 方法进行了改进，将其拓展为多元 DTW，对不等长序列进行转化，且在计算 DTW 距离时引入了多个下界距离。实验结果表明，改进的 NN DTW 算法能够处理传统的 DTW 算法所无法处理的不等长多元 DTW 问题，在识别异常交易方面具有更好的性能表现。

关键词：异常交易　市场微观结构　动态时间规整　极值理论

ABSTRACT

Insider trading undermines the integrity and orderliness of financial markets. In order to better monitor investors' trading behavior and provide a references for authorities, this study proposes a non-parametric approach to identifying insider trading based on multiple Dynamic time warping (DTW), combined with extreme value theory for threshold delineation. In terms of insider trading, combining empirical studies at home and abroad, this study refers to the definition of insider trading behavior in the Shenzhen Stock Exchange and uses indicators such as pulling and suppressing share prices and false declarations as the basis for determining the characteristics of insider trading. In the model evaluation and empirical analysis, the study analyses two cases of insider trading disclosed by the Shenzhen Stock Exchange and is able to effectively discern insider trading sequences within the disclosed existence date. Meanwhile, this paper combines previous research and improves the DTW method by extending it to multivariate DTW, transforming unequal sequences, and introducing multiple lower bound distances in calculating the DTW distance. The

experimental results show that the improved NN DTW algorithm can handle the unequal-length multivariate DTW problem that cannot be handled by the traditional DTW algorithm, and has better performance in identifying insider trading.

KEY WORDS: Insider trading Market microstructure Dynamic time warping Extreme value theory

目　录

基于动态时间规整的异常交易识别

一、绪论

异常交易行为破坏了金融市场的完整性与有序性。证监会 2023 年 1 月 13 日发布的《证券经纪业务管理办法》[1]规定，投资者开展证券交易，应当遵守法律法规、中国证监会、证券交易场所的规定，不得进行违规交易，并主动避免异常交易。证券公司应当按照证券交易场所的规定加强异常交易监测，做好投资者交易行为管理。异常交易的定义依据所在国家而有所不同，按照《深圳证券交易所创业板股票异常交易实时监控细则》(下文称《交易监控细则》)[2]的定义，包括虚假申报、拉抬打压股价、维持涨(跌)幅限制价格、自买自卖或者互为对手方交易、严重异常波动股票申报速率异常等行为。随着我国股票市场逐渐发展，市场信息泄露问题愈发严重，机构投资者参与的内幕交易数量呈上升趋势。众所周知，旨在检测非法交易活动的成功的市场监督做法可以提高金融市场的质量[1][2]。

然而，精确识别异常交易活动是一个困难的问题，因为与合法交易量相比，确认的异常交易案例很少。此外，异常交易活动的模式取决于背景，通常需要结合账户组交易动机判断。目前识别异常交易数据的方法是相关从业人员根据历史业务经验，并人工设定识别准则参数的方法。明确的异常交易模式的统计方法能被参数方法有效辨别，然而，一系列问题限制了经典参数模型的应用。识别方法的有效性和准确性容易受到相关人员的业务能力以及历史数据中的小概率事件影响[3]。此外，由于标的资产的流动性和波动性随时间变化，有意识地试图掩盖异常交易策略的情况，依靠特征空间中交易之间的线性排列的传统统计距离测量方法往往不正确地测量相似性，意味着时间错位可能导致经典参数模型的失效。

目前，证券经纪商通过使用基于规则的市场监督系统来监测客户和原则账户的交易活动，以防止可疑的非法交易活动，从而遵守监管要求。这些系统的目的是保守地提出相关警报，以确定潜在的异常交易活动，供人工进一步审查。文章的基本前提是，异常交易活动与历史上的合法交易活动模式有很大不同。因此，从交易订单簿中体现的极端异常性是构成一个有效的市场监控系统中的相关警报的关键条件。

在这项研究中，最近邻动态时间规整模式识别算法与极值理论将被结合起来，共同构成一个旨在检测异常交易活动的自适应监控模型。监控模型的设计是基于时间序列异

[1] http://www.csrc.gov.cn/csrc/c101953/c6987676/content.shtml
[2] http://www.szse.cn/disclosure/notice/general/t20200612_578383.html

常检测技术，利用高频订单簿信息总结出的特征来确定异常交易活动序列，并找到多个证监会已披露出来的异常交易案例，检验方法的有效性。

相对于现有研究，文章的主要贡献体现在两个方面：其一，首次使用 NN DTW 研究股票异常交易的检测方法，文章设计了一个非参数性，基于单类机器学习方法的检测模型，即不对数据生成过程或异常内幕交易的结构进行假设，转而对一段时间内订单的整体异常情况进行检验；其二使用实证案例对模型进行评价与分析。

二、文献综述

（一）异常交易与其识别手段

非法交易活动在金融市场上普遍存在[4]，我国证监会将其称为异常交易。即使对于具体的定义存在法律规定层面的区分[5]，但是异常交易是金融市场微观结构理论一个最基本的假设。从根本上，这些事件是由信息不对称产生的，异常交易严重破坏了市场的公平，降低了资源配置的效率[6][7][8]。因此，各国都有明确的要求，对异常交易或非法交易进行监管与起诉[9]。基于对美国证券交易委员会(SEC)的执法研究表明，积极执法可以阻止非法活动并提高市场的流动性[10]。

过往的研究从多个方面给出了异常交易的作用因素和社会影响。1985 年，Kyle 提出了经典的顺序拍卖的内幕交易动态模型，用于研究价格的信息内容、投机市场的流动性特征以及私人信息对内部人士的价值，研究了单一风险中性内部人士、随机噪声交易员和竞争性风险中性做市商在一场内幕交易中的行为模式[11]。Steven Huddart 等人在多期理性预期框架中提出了内幕均衡交易策略，发现要求内部人士在事后公开披露其股票交易的法规使知情的寻租内部人士的交易决策复杂化，相对于 Kyle[11]，价格发现加速，内部利润更低[12]。此外，异常交易还可以反映在买卖价差上[13][14]。Kee.H 等人通过回归分析，指出尽管做市商可能无法在内幕交易发生时检测到内幕交易，但他们通过为内幕交易倾向较大的股票保持较大的价差来保护自己，证明了从交易数据层面识别异常交易的可能性[15]。

为了准确识别并监测异常交易，学者们提出并检验了很多方法。P.Collin-Dufresne 等人将 Kyle[11]的内幕交易模型扩展到包含噪声交易者的情况，说明了当波动更小的时候，知情人士进行异常交易的可能性就更大[16]。这类型的研究说明了异常交易出现的时机[17]，为进一步建立检测模型提供了思路，但检测精度未达到准确识别的标准。用于监测的检测模型通常使用高频数据对市场进行建模。比如 Foster 和 Viswanathan 就利用了高频数

据对投资者行为进行了研究，知情交易者信号之间的初始相关性对知情交易者的利润和价格的信息性有显著影响[18]。有的研究对高频数据的利用是基于传统的金融计量经济学方法，许多文献总结归纳出一些异常交易的特定交易模式[19]。也有更为复杂的，比如 J Olmo 等人引入了基于 U 统计类型过程的一致时序结构中断检验(CTSB)，通过扩展资本资产定价模型的截获结构变化来识别内幕交易[20]。而近年来，随着机器学习技术的发展，各种机器学习技术也被应用于证券异常交易检测。MF Esen 等人以美国市场为例，使用 k-mean 和分层聚类的异常值分析数据挖掘技术来检测可疑的内幕交易[21]。Christian T. Lundblad 等人从现有监管行动中提取典型内幕交易案例的几个关键经验特征，采用孤立森林算法(isolation forest)研究内幕交易检测[22]。对国内市场的研究中，运用机器学习识别异常交易的文献也不在少数。柴洪峰等人采用贝叶斯信念网络算法将一段交易操作序列与异常序列进行比对，判断交易序列的异常分数[23]。王欣等人则采用了分位数回归模型结合变点检验的方法[24]。而文章则采用 NN DTW 结合极值理论的异常交易识别方法。

（二）动态时间规整与极值理论

动态时间规整(Dynamic Time Warping, DTW)是一个在时间序列分类和聚类任务中表现出色的算法[25]，最早由日本学者 Itakura 提出[26]，通过构建观测值之间的非线性映射来找到两个序列之间的最佳排列，其基本思想是对查询序列的各个部分进行扭曲，以找到与参考序列的最佳匹配，对于两个序列的相似区域，任何时间错位都不影响他们之间的匹配情况度量。DTW 在时间序列相似性搜索任务中的优异表现已在以往文献中得到证实[27][28][29]，常见应用场景包括语音识别，手势识别，数据挖掘和信息检索等。但有别于常用做法，文章希望将 DTW 应用于证券交易情况下的高频时间序列异常分类。

极值理论(Extreme Value Theory, EVT)是次序统计理论的其中一个分支，处理概率分布的中极端值的理论，常用来分析概率的尾部分布，在水文，气候等风险管理相关领域有着广泛应用[30][31]，在金融风险管理中同样占据一席之地[32]。Fisher 和 Tippett 最早进行了极值理论方面的研究[33]，进一步 Gendenko 在 1943 年给出了三大类型定理的严格证明[34]，建立了著名的极值定理(Fisher-Tippett-Gnedenko 定理)，定理给出了极值理论中关于极阶统计量的渐近分布的一般结果，即适当重整化后的 i.i.d 随机变量样本的最大值只能收敛到广义帕累托分布的 3 种可能分布之一。

三、异常交易判别特征构造

（一）高频订单簿复现

文章所使用数据来源于深圳证券信息有限公司 [3]的数据服务产品，能够提供市场上逐笔委托的交易情况，包括委托提交，委托撤销以及交易执行的情况。原数据用 3 个表展示市场行情：证券行情快照档位表，逐笔委托行情表，以及逐笔成交表。由于目标是识别异常交易，故重点应当放在交易执行情况。因此通过将表与表之间的数据对应匹配，文章还原了逐笔成交情况以及对应时间点的买卖方最优报价情况。

算法 1: 订单簿复现建模

输入: 逐笔成交表 DataFrame **trade**， 逐笔委托行情表 DataFrame **order**， 证券行情快照档位表 DataFrame **snap_level.**

输出: 高频订单簿 DataFrame **OB**.

```
数据预处理
初始化当前档位 current_level
初始化订单簿 OB
initiate_side = 0
i = 0
合并 trade，order 两表，并以时间升序排序，记为 df
for row in df do
    if 当前交易信号时间 between time_i and time_{i+1} in snap_level then
        对 current_level 准确性进行验证并矫正
        i = i + 1
    if row from order then
        if 买方发起 then
            if 价格大于等于最优卖价 then
                if 集合竞价 then
                    current_level 对应档位挂单量增加
                else 连续竞价 then
                    更新 current_level 并进行交易操作
                    initiate_side = 0
            else
```

[3] http://www.szsi.cn/cpfw/sjfw/

```
                current_level 对应档位挂单量增加
        else 卖方发起 then
            if 价格小于等于最优买价 then
                if 集合竞价 then
                    current_level 对应档位挂单量增加
                else 连续竞价 then
                    更新 current_level 并进行交易操作
                    initiate_side = 1
    else row from trade then
        if 撤单 then
            current_level 对应档位挂单量减少
        else 成交 then
            if 集合竞价 then
                    更新 current_level 并进行交易操作
            else 连续竞价
                记录当前的 current_level， initial_side 与成交数据
                OB 新增一行数据
        else
            warning
    else
        warning
return OB
```

最终，用于构建特征的高频订单簿变量如下表所示。对于最终的订单簿，按照时间顺序排列，第 i 条成交记录，有以下原始指标

表 1 订单簿原始指标

符号	变量	说明
t_i	成交时间	
p_i	成交价格	
v_i	成交量	

Initiate side	交易发起方	0=买方发起；1=卖方发起
bid_{t_i}	t_i时刻的最优买价	
ask_{t_i}	t_i时刻的最优卖价	

（二）交易量特征

理论和实证文献指出，交易量与知情的市场参与者持有的私人信息程度相关[35][36]。而且依据深交所《交易监控细则》规定，异常交易情况之一的拉抬打压股价行为指大笔申报、连续申报、密集申报或者以明显偏离股票最新成交价的价格申报成交，期间股票交易价格明显上涨或下跌的行为；重点关注成交数量或者金额较大，以及成交数量占期间市场成交总量的比例较高的申报。依照这个标准，文章使用与每笔交易的原始交易量作为检测模型的一个特征，此外还计算了该笔交易占一段时间内市场成交账户交易总量的比例，作为衡量标准。计算成交量占比的时间窗口终点为该笔交易发生时间，时间窗口起点为该笔交易发生时间前 5 分钟。

$$\text{成交量占比} = \frac{v_i}{\sum_{z\in\omega_{t_i}^{(5)}} v_i} \tag{1}$$

其中$z \in \omega_{t_i}^{(5)}$表示 5 分钟的滑动时间窗口。

（三）价格冲击特征

《交易监控细则》中规定拉抬打压股价也属于一种异常交易行为，具体结果包括股票交易价格明显上涨(下跌)。自然而然，价格冲击特征也能够成为判别特征之一。在文章中，瞬时价格冲击被定义为 5 秒内对手方最优价格的对数变化量，如下面公式所示。

$$\text{瞬时价格冲击} = \begin{cases} \log\left(ask_{t_i+\delta}\right) - \log\left(ask_{t_i}\right), & Initiate\ side = 0 \\ \log\left(bid_{t_i+\delta}\right) - \log\left(bid_{t_i}\right), & Initiate\ side = 1 \end{cases} \tag{2}$$

其中 $ask_{t_i+\delta}(bid_{t_i+\delta})$ 表示第 i 次交易后第一次发生变化的对手方最优价格，当 5 秒内价格没有发生变化，则顺势价格冲击为 0。

如文章文献综述中所言，买卖价差有时候也能成为异常交易的判别特征之一[13][14]，反映了交易的长期价格影响。在这里将其归类为价格冲击特征，参照 Hendershott 等人的方法[37]，构建结果如下。

$$长期价格影响 = \begin{cases} 2(m_{t_i+k} - m_{t_i})/m_{t_i}, & Initiate\ side = 0 \\ -2(m_{t_i+k} - m_{t_i})/m_{t_i}, & Initiate\ side = 1 \end{cases} \quad (3)$$

其中$m_{t_i} = \frac{ask_{t_i} - bid_{t_i}}{2}$表示买卖价差。

（四）波动率特征

在实际操作层面，《交易监控细则》规定其中一种异常交易情况为：严重异常波动股票申报速率异常。表明股票的波动率也是衡量异常状态的特征。而在理论层面，Attanasio 在一个有知情和不知情的交易者的简单的两期模型中，证明了不对称的信息会导致不知情的交易者理性地但错误地对外生噪音做出反应，他们错误地认为这是一个基本的定价信号[38]。这种过度反应加剧了错误的价格运动，增加了波动性。从最近的市场经验上看，Kacperczyk 和 Pagnotta 发现，在市场上出现非法内幕人士的日子里，波动性会异常地高[39]。基于前文构建的订单簿数据，文章计算了 30 分钟内逐个交易移动窗口的日内实现波动率的计量。众所周知，在利用高频数据估算波动率时，我们的目的是计算综合波动率(Integrated Volatility, IV)，但是由于其不可观测，故转而计算已实现波动率(Realized Variance, RV)。

$$\widehat{RV}_i^{(\Delta)} = \sum_{z \in \omega_{t_i}^{(30)}} r_{i,j}^2 \quad (4)$$

$$r_{i,j} = \log(m_{i-1+\Delta j}) - \log(m_{i-1+\Delta(j-1)})$$

其中，$r_{t,j}$是短期收益率，$z \in \omega_{t_i}^{(30)}$表示 30 分钟的滑动时间窗口，$\Delta$表示的是取收益率时间的计算间隔。

众所周知，微观结构摩擦对日内已实现波动率的计算会产生影响[40]，文章参照 Patton 和 Sheppard 等人的研究，计算 2 种共 6 个微观结构噪声稳健的 RV 估计值的等权平均值作为文章 RV 估计值[41]，尽可能准确地估计 RV。第一种 RV 估计值是标准的已实现波动率 RV，并将计算间隔Δ取 3，9，30，60 秒的四种间隔，共计 4 个已实现波动率。第二种 RV 估计值的计算基于 Bandi 和 Russell 提出的无偏 RV 的计算方法[42]，主要的思想是将价格的观测值$\tilde{p}_i$视作真实的无摩擦价格p_i与摩擦系数ϑ_i的乘积，此处不再展开。文章

使用去除微观结构噪声的估计值$\widehat{RV}^{BR,\ bc}$，还有没有去除微观结构噪声估计值$\widehat{RV}^{BR}$作为剩下两种估计值。

（五）持续时间特征

《交易监控细则》中拉抬打压股价的描述里，连续申报、密集申报也属于有可能导致被判定为异常交易的因素，因此构建持续时间特征，计算了订单从发出到成交之间的持续时间；对于撤单的订单持续时间为订单发出到撤单之间的持续时间。

（六）判别特征小结

鉴于异常交易识别这一类问题的性质，标准的特征选择技术，如 Lasso 等惩罚性回归模型，不能用于确定检测的最佳特征集。相反，我们通过参考关于信息不对称的市场动态的献，结合我国的规章制度安排来定义一组相关的微观结构特征。按照大类进行设置的指标一方面能够保证了指标设置的合理性，也不会因为过分精细而导致过拟合问题。为了方便模型运算与分析，除了数据本身较小的价格冲击之外，将部分用到的指标归一化处理，文章一共设计的特征如表 2 所示。

表 2 判别特征小结

特征类型	特征	归一化操作
交易量特征	成交量	√
	成交量占比	√
价格冲击特征	瞬时价格冲击	
	长期价格影响	
波动率特征	已实现波动率 rv	√
持续时间特征	订单持续时间	√

四、异常交易识别方法

（一）异常评分框架

对于每只股票，文章将其每天的订单簿进行处理，构成一个日内成交序列，并按照每 15 分钟为划分，给出多个连续的逐笔交易窗口。每一笔交易都是一个p维特征向量$\boldsymbol{v}_j$，一个检测序列则表示为 $V_{id} = \{\boldsymbol{v}_j, j = 1,\ldots,m\}$，其中$i$表示时间窗口索引，$d$表示日期，$m$表示该时间窗口内的成交数。将所有处理后的检测序列作为模型所需的总信息集，集合表示为 $\{V_{id}, i = 1,\cdots,N_d\}_{d=1}^{D}$，其中$N_d$表示一天之内划分的时间窗口数量，$D$表示总天

数。

紧接着，将总信息集分为 2 个子集，训练集X与测试集Y。测试集$X=\{V_{id}:i=1,\ldots N_d,\ d\in I_1\}$，其中$I_1\subset\{1,\ldots,D\}$，$X$由合法交易活动的数据组成；测试集$Y=\{V_{id}:i=1,\ldots N_d,\ d\in I_2\}$，其中$I_2=I_1^c$，并且$min(I_2)>max(I_1)$。测试集数据中的成交数据全部合法基于一个假设前提：证监会在该段时间内没有透露任何违规异常交易行为，且缺乏其他任何验证方法，因此文章认为交易数据全部合法。

异常交易的检测分数由 Y 和 X 的比较中得出。异常评分函数定义为$f(X,\ Y,\ \theta)$，其中 θ 是一组超参数。这个函数能够使用 NN DTW 算法为所有的 Y 序列生成异常得分，其参考对象为数据集 X。主要步骤分为以下两步：

首先，将异常评分函数应用于合法交易活动序列，生成一组训练样本的异常评分$\vartheta(X)=f(X,\ X,\ \theta)=\{\vartheta_{id},i=1,\ldots,N_d,\ d\in I_1\}$。令$F$表示$\vartheta(X)$的分布，它是一个连续分布。文章使用来自$\vartheta(X)$的顺序统计样本来估计$\vartheta(X)$广义帕累托极值分布的参数，该分布作为 F 的上尾的渐近有效近似，在置信度为α前提下，使用近似值$\tau(\alpha)$为异常得分阈值。

然后，将异常评分函数应用于测试集数据Y，生成一组测试样本异常分数$\vartheta(Y)=f(X,Y,\ \theta)=\{\vartheta_{id},\ i=1,\ldots,N_d,\ d\in I_2\}$。类标签向量 $\hat{y}=-sign(\vartheta(Y)-\tau(\alpha))$，表示异常判别结果。如果异常分数超过了阈值$\tau(\alpha)$，即如果$\hat{y}_{id}=-1$，则一个序列被标记为异常，代表潜在的异常交易活动。

（二）NN DTW

本小节中描述了如何将 DTW 与最近邻方法(nearest neighbor, NN)进行结合，解释了为何本方法被称为 NN DTW 算法。DTW 的序列匹配是一种允许时间轴的弹性移动的匹配方法，以适应相似但非相位的序列。DTW 常应用于一元时间序列，但多元序列同样能够应用 DTW[43]，对多元 DTW 的拓展主要有两种方式，独立 DTW 与非独立 DTW。独立 DTW 的方法为：对多变量序列中每个特征的时间序列应用 DTW 来构建特定特征的排列。然后，将与每个特征相对应的距离相加，以计算多变量序列之间的总距离。这种方法被称为独立 DTW，因为它假定特征之间相互独立，一个特征的特性不会影响另一个特征的弯曲路径。但是显然订单簿数据特征之间是非独立的，因此文章采用的是非独立 DTW，区别主要体现在利用高维欧氏距离作为点对之间的距离测量。大致原理如下：

设 **A** 是来自测试集的序列，**B** 是来自训练集的序列。**A** 和 **B** 表示两个具有 p 个特征的多元序列，**A** 包含 m 个观测值，**B** 包含 n 个观测值，即 **A** 是 $m\times p$ 维矩阵，**B** 为 $m\times p$ 维矩阵。为了使用 DTW 对准两个序列，本方法构建一个$n\times m$的矩阵，其中矩阵元素(i，

j)包含两个行向量$\{\boldsymbol{a}_i,\ i=1,\ \ldots,\ m\}$和$\{\boldsymbol{b}_j,\ i=1,\ \ldots,\ n\}$之间的欧氏距离$d(\boldsymbol{a}_i,\ \boldsymbol{b}_j)$，即

$$d(\boldsymbol{a}_i,\boldsymbol{b}_j)=\left\|\boldsymbol{a}_i-\boldsymbol{b}_j\right\|_2$$

其中，$\|\cdot\|_2$为欧氏范数。弯曲路径W被定义为一个在 3 个基本约束下的连续路径，其中 W 的第 k 个元素被定义为$w_k=(\boldsymbol{a}_i,\ \boldsymbol{b}_j)_k$，则

$$W=w_1,w_2,\ldots,w_k,\ldots,w_K,\ \ max(m,n)\le K<m+n-1$$

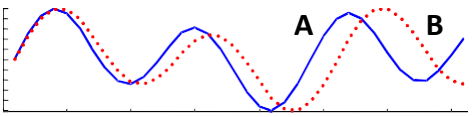
弯曲路径必须满足 3 个基本约束:

(1)边界条件：路径起始于点$w_1=(1,1)$，终止于点$w_K=(m,n)$，它表示两个序列的起始点和结束点对应匹配;

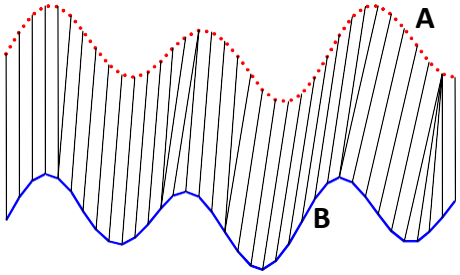
(2) 连续性：路径上的任意两个相邻点$w_k=(i_1,j_1)$和$w_{k+1}=(i_2,j_2)$满足条件$0\le|i_1-i_2|\le1$，$0\le|j_1-j_2|\le1$;

(3) 单调性：若$w_k=(i_1,j_1)$和$w_{k+1}=(i_2,j_2)$为路径上前后两个点，则须满足$i_2-i_1\ge0$，$j_2-j_1\ge0$。

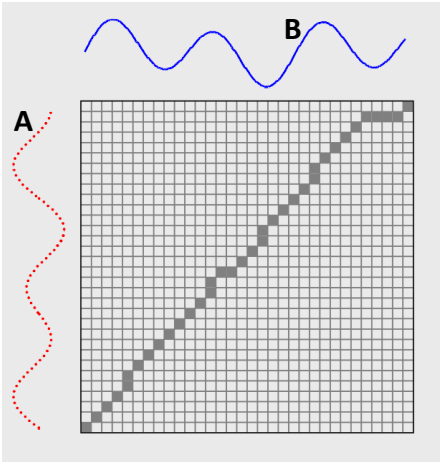
对于两个序列，弯曲路径有很多，每一条弯曲路径对应一种点对匹配关系。点对匹配关系中，点对距离之和的最小值即为 DTW 距离，对应的弯曲路径为最佳路径。如图 1 所示，在(b)图中，为了对齐两个序列，我们构建了一个弯曲矩阵，并寻找最佳的弯曲路径，用实心方块表示。在(c)图中，展示了得到的对齐结果[44]。



(a)时间序列 A 和 B



(c)点对匹配结果



(b)弯曲路径

图 1 弯曲路径与点对匹配结果

DTW 距离表示为

$$DTW(\boldsymbol{A},\boldsymbol{B}) = min\left\{\sqrt{\sum_{k=1}^{K} w_k}\right\} \tag{5}$$

但是在实际操作中，为了节约算力，加快运算速度，开平方的计算可省去而不影响结果，则

$$DTW(\boldsymbol{A},\boldsymbol{B}) = min\left\{\sum_{k=1}^{K} w_k\right\}$$

最佳弯曲路径可以通过动态规划求解，通过构建一个累积成本矩阵求得结果。用C^*表示累计成本矩阵，则初始条件是

$$C_{i,j}^* = \begin{cases} \infty, & 当i = 1 或 j = 1 \\ 0, & 当i = j = 1 \end{cases}$$

递推关系为

$$C_{i,j}^* = d(\boldsymbol{a}_i, \boldsymbol{b}_j) + min\{C_{i-1,j}^*, C_{i,j-1}^*, C_{i-1,j-1}^*\}, \qquad i = 2,\ldots,m;\ j = 2,\ldots,n$$

（三）DTW 优化

DTW 距离$DTW(\boldsymbol{A},\boldsymbol{B})$被用作标准 NN 算法中的统计距离测量。但是 NN DTW 算法的复杂性相当高，抑制了其在大规模时间序列分类问题上的应用，如文章研究的金融市场监控。为了解决这个问题，许多研究已经提出了来降低 NN DTW 算法的复杂性。在本章节中，将介绍多个 DTW 的优化方法，包括添加全局约束，计算下界距离，提早放弃策略三种。

1. 不等长时间序列至等长时间序列的转化

由于上文提到的这三种 DTW 优化方法只能处理相同长度的时间序列，严重限制了

其在实际情况中的运用，也迫使 DTW 放弃了其独特的测量不等长序列的优势。而按照实际时间分割的股票高频交易序列往往是不等长的，因此首先需要将其转化为等长时间序列。为了解决这个问题，Zhengxin Li[44]提出了一种序列拓展技术，并证实了拓展后的序列在计算下界距离时候，在紧密性与复杂性方面的优越性。文章依照 Zhengxin Li 文章中的方法对序列进行拓展，在此不再展开。

2. 弯曲路径的全局约束条件

除了 3.1 中提到的 3 个约束条件，弯曲路径通常需满足第四个约束条件，在文献中被称为弯曲窗口宽度约束(Warping Window Width constraint)。对于弯曲路径上的元素$w_k = (\boldsymbol{a}_i, \boldsymbol{b}_j)_k$的下标$i$和$j$，要求$j + \varphi \leq i \leq j + \varphi$，其中$\varphi$为一常数，表示弯曲限制的宽度，如图 2 的左图所示，表示为带状矩阵的宽度。这种约束被称为 Sakoe-Chiba 约束；如图 2 的右图所示，Itakura-Parallelogram 约束中，r 为 i 的函数，弯曲窗口为沿对角线方向的平行四边形。文中主要围绕 Sakoe-Chiba 约束条件进行讨论。

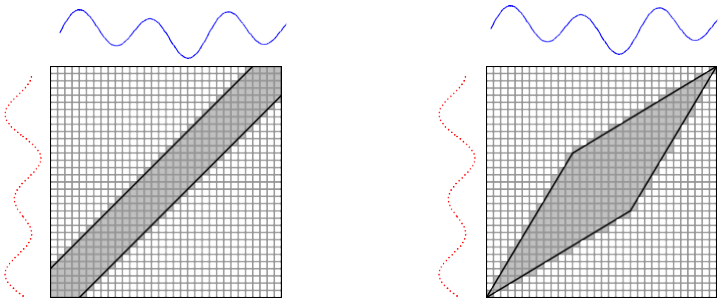


图 2 两种全局约束[45]

弯曲路径的全局约束条件有效地限制了累积距离矩阵和最佳弯曲路径的可信区域，直观地讲，全局约束限制了观察值可以向前或向后多远的时间弯曲，限制了最佳弯曲的搜索空间，也降低了 DTW 算法的复杂性，以此确保解决方案的现实性。

3. DTW 的下界距离与提早放弃策略

通过计算两个序列的下界距离，若目前的下界距离大于之前已经计算出来的最小的下界距离(以往文献中多称为 BSF 距离，*best_so_far_distance*)，则提早放弃匹配，是 DTW 效率优化的常用思路。一旦本次计算并不优于之前的计算则马上放弃，这种方法有效减少了计算次数。

文章使用三个下界函数和一个提早放弃技术(early abandoning techniques)进行 DTW 优化，序列下界函数的输出是一个小于真实 DTW 距离的标量值。为了达到优化计算速度的目的，下界函数除了必须比真正的 DTW 距离计算得更快，而且必须与真正的 DTW

距离相接近，这一属性被称为下限的 "紧密度"。本研究中使用的三个下限是以级联方式实现的，其复杂程度依次提高，同时紧密度也随之提高。下面将按照复杂度的升序来描述这些下界。

第一个下界距离被称为修正的 LB_Kim 下界距离[46]，它认为第一对观测值之间的欧氏距离和最后一对观测值之间的欧氏距离之和，修改后的 LB_Kim 下界表示为

$$LB_{Kim(\boldsymbol{A},\boldsymbol{B})} = \|\boldsymbol{a}_1 - \boldsymbol{b}_1\|_2 + \|\boldsymbol{a}_m - \boldsymbol{b}_m\|_2 \tag{6}$$

第二个更紧密但计算成本更高的下界，称为 LB_Keogh[45]。LB_Keogh 下界的原理是在匹配序列周围建立一个包络。LB_Keogh 下界表示为

$$LB_{Keogh(\boldsymbol{B},\boldsymbol{U},\boldsymbol{L})} = \sqrt{\sum_{q=1}^{m}\sum_{k=1}^{p}\begin{cases}\left(\boldsymbol{B}_{kq} - \boldsymbol{U}_{kq}\right)^2, & \boldsymbol{B}_{kq} > \boldsymbol{U}_{kq} \\ \left(\boldsymbol{B}_{kq} - \boldsymbol{L}_{kq}\right)^2, & \boldsymbol{B}_{kq} < \boldsymbol{L}_{kq} \\ 0, & else\end{cases}} \tag{7}$$

其中，

$$\boldsymbol{U}_{kq} = max(\boldsymbol{A}_{k-\varphi:k+\varphi,\ q})$$

$$\boldsymbol{L}_{kq} = min(\boldsymbol{A}_{k-\varphi:k+\varphi,\ q})$$

$\mathbf{U}_{\mathrm{kq}}$和$\mathbf{L}_{\mathrm{kq}}$为宽度依赖于弯曲窗口约束φ大小的上下包络序列，具体原理可以参照原文[45]，文章不再过多展开。

最后，如果 LB_Kim 和 LB_Keogh 都不能排除当前匹配，则计算 LB_Improved[47]。LB_Improved 利用 LB_Keogh 下界的信息来计算B′，即计算待验证的序列 **B** 在上下包络上的投影。

$$B'_{kq} = \begin{cases}\mathbf{U}_{\mathrm{kq}}, & \mathbf{B}_{\mathrm{kq}} > \mathbf{U}_{\mathrm{kq}} \\ \mathbf{B}_{\mathrm{kq}}, & \mathbf{L}_{\mathrm{kq}} < \mathbf{B}_{\mathrm{kq}} < \mathbf{U}_{\mathrm{kq}} \\ \mathbf{L}_{\mathrm{kq}}, & \mathbf{B}_{\mathrm{kq}} < \mathbf{L}_{\mathrm{kq}}\end{cases} \tag{8}$$

LB_Improve 的下界函数表示为

$$LB_{improved} = LB_{Keogh(\boldsymbol{B},\boldsymbol{U},\boldsymbol{L})} + LB_{Keogh(\boldsymbol{A},\boldsymbol{U}',\boldsymbol{L}')} \tag{9}$$

其中，

$$\boldsymbol{U}'_{kq} = max(\boldsymbol{B}_{k-\varphi:k+\varphi,\ q})$$

$$\boldsymbol{L}'_{kq} = min(\boldsymbol{B}_{k-\varphi:k+\varphi,\ q})$$

如果上述三个下界都小于 BSF 距离，就在计算 DTW 距离的过程中实施两种提前放弃策略。提前放弃规定了一些条件，当条件满足时，就有可能提前停止 DTW 距离的计算。第一个提前放弃策略为，在第 i 次迭代中获取完整 DTW 距离的部分值，即 $min\{C'_{i1},\ldots,C'_{im}\}$，若其超过了 BSF 距离，则可以放弃当前的候选序列比较，继续进行下一次的 DTW 迭代。第二个提前放弃策略额外需要计算余下部分的 LB_Keogh 下界，即当

$$min\{C'_{i1},\ldots,C'_{im}\} + LB_{Keogh(\boldsymbol{A}_{i:m,1:m},\ \boldsymbol{U}_{i:m,1:m},\ \boldsymbol{L}_{i:m,1:m})} \tag{10}$$

大于 BSF 距离时，放弃当前的候选序列比较。

下面的伪代码给出了 NN DTW 的异常值评分过程。

算法 2: NN DTW 异常评分

输入: 训练集数据 **X**，测试集序列 **A**.

输出: 测试集数据 **A** 的标量异常评分.

```
初始化 BSF_distance = inf
for i = 1，...，|X| do
    B = X_i
    if LB_Kim(A，B) < BSF_distance then
        if i = 1 then
            构造上下包络序列 U，L
        if LB_Keogh(B，U，L) < BSF_distance then
            构造序列 B'
```

```
9:              构造序列上下包络序列 U'，L'，B'
10:             if LB_Keogh(B，U，L) + LB_Keogh(A'，U'，L') < BSF_distance then
11:                 if DTW(A，B) 被提早放弃 then
12:                     DTW_distance_score = inf
13:                 else
14:                     DTW_distance_score = DTW(A，B)
15:             else
16:                 DTW_distance_score = inf
17:         else
18:             DTW_distance_score = inf
19:     else
20:         DTW_distance_score = inf
21:     if DTW_distance_score < BSF_distance then
22:         BSF_distance = DTW_distance_score
23: anomaly_score = BSF_distance
24: return anomaly_score
```

（四）极值理论下的阈值确定

本节描述了如何利用极值方法中的POT(Peaks-Over-Threshold)方法确定异常得分的阈值，方法描述了一个随机过程在异常大或小的数值下的渐进行为。该理论可用于构建基础分布的尾部的近似分布，而不需要对基础分布的整体进行参数化的假设。

对于一组训练集样本的异常分数$\vartheta(X) = \{\vartheta_{id}, i = 1, \ldots, N_d, d \in I_1\}$，假设这些样本都是独立同分布的，且他们的未知的分布函数设为 F。对于一个给定的阈值 u，定义F_u为超过该阈值的观察值X的分布，则

$$\begin{aligned} F_u(y) &= Pr|X - u \le y|\, X > u| \\ &= \frac{Pr\left|X - u \le y,\ X > u\right|}{\Pr(X > u)} = \frac{F(y+u) - F(u)}{1 - F(u)} \end{aligned} \tag{11}$$

其中$y = x - u > 0$表示观测值超出阈值部分的大小，称为超出量，对于$x > u$，下面的表达式描述了尾部分布的一个性质

$$F(x) = 1 - [1 - F(u)][1 - F_u(y)] \tag{12}$$

基于Pickands[48]和Balkema等人[49]提出的分布尾部估计方法，在一个较大的阈值下，尾部分布$F_u(\mathrm{y})$收敛于广义帕累托分布族$G_{\xi,\sigma}(y)$(generalized Pareto distribution, GPD)，则

$$F(x) = 1 - [1 - F(u)]\left[1 - G_{\xi,\sigma}(y)\right] \tag{13}$$

其中

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \dfrac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}}, & \xi \neq 0,\ y > 0 \\ 1 - \exp\left(-\dfrac{y}{\sigma}\right), & \xi = 0,\ y > 0 \end{cases} \tag{14}$$

其中ξ是形状参数，σ是尺度参数 [4]，且$1 + \frac{\xi y}{\sigma} > 0$。对于不同的值，GPD 对应着不同

的分布形式，正如文献综述中提及到的 3 种可能分布[34]，在此不过多展开。为了对这里的$F(x)$进行估计，POT 方法进一步将上述式子中的$1 - F(u)$非参数地近似于超过阈值 u 的上方尾部区域的观测值的比例，对于比较大的阈值 u，则 POT 的尾部估计为

$$\hat{F}(x) = 1 - \frac{N_u}{N}\left(1 + \frac{\xi z}{\sigma}\right)^{-\frac{1}{\xi}} \tag{15}$$

其中，N_u表示样本中大于阈值的样本点的数目，N 表示样本大小。

在文章的股票异常评分中，对于给定的置信水平α，我们仅希望有$1 - \alpha$的概率使得异常评分高于 u，从而限定异常评分的严格度。因此，可以通过计算分布函数的反函数获得α的估计量$\hat{Q}_\alpha$，公式为

[4] 在阈值较小的情况，还应该存在一个位置参数μ，本方法目的为极值判断，阈值设置大，省略了位置参数。

$$\hat{Q}(\alpha)=\hat{F}^{-1}(\alpha)=u+\frac{\hat{\sigma}}{\hat{\xi}}\left(\left(\frac{N}{N_u}(1-\alpha)\right)^{-\hat{\xi}}-1\right),\quad \alpha>F(u) \tag{16}$$

然而，使用$\hat{Q}(\alpha)$作为异常评分阈值仍留下了改进的空间，因为超过分位数$\hat{Q}(\alpha)$的信息被忽略了。换句话说，异常分数略大于$\hat{Q}(\alpha)$的序列将被标记为异常，尽管它们通常与略低于$\hat{Q}(\alpha)$的异常分数没有太大区别。一个鲁棒性更高的异常得分阈值将考虑训练样本异常得分超过$\hat{Q}(\alpha)$的部分的平均值$E[\vartheta_{id}-\hat{Q}(\alpha)|\vartheta_{id}>\hat{Q}(\alpha)]$。这部分尾部样本称为条件尾部期望(Conditional Tail Expectation, CTE)。为了推导这一数据量，还需要用到另一个极值理论的定理

$$E[\vartheta_{id}-u|\vartheta_{id}>u,i=1,\dots,N_d,d\in I_1]=\frac{\sigma}{1-\xi},\quad \xi<1 \tag{17}$$

对于考虑到 CTE 的u′ > u，超出量$y'=\{\vartheta_{id}-u'|\vartheta_{id}>u'\}$同样收敛于 GPD 且

$$\begin{cases}\xi'=\xi\\ \sigma'=\sigma+\xi(u'-u)\end{cases}$$

当$\xi<1$时，最终基于 CTE 的异常评分阈值为[50]

$$\hat{\tau}_\alpha=\hat{Q}(\alpha)+E\left[\vartheta_{id}-\hat{Q}(\alpha)|\vartheta_{id}>\hat{Q}(\alpha),\ i=1,\dots,N_d,d\in I_1\right]=\hat{Q}(\alpha)+\frac{\hat{\sigma}+\hat{\xi}\left(\hat{Q}(\alpha)-u\right)}{1-\hat{\xi}} \tag{18}$$

其中，$\alpha>F(u)$

异常得分的阈值可以理解为训练集中最不寻常的序列的平均异常得分，而置信度 α 定义了异常训练样本异常得分在 F 上的分位数。按照财务风险的惯例，在本研究中α被设置为 0.95，即异常得分阈值是至少根据最大的 5%的训练样本异常得分的平均行为来设置的。

最后，为了得到最终的阈值，还有三个参数需要确定，分别是ξ，σ和u。对于ξ和σ的估计使用到了极大似然估计法，对于 k 个阈值为 u 的超出量$\{y_i,\ i=1,2,\dots,k\}$，计算其似然函数为

$$L(\xi,\ \sigma)=\begin{cases}-k\cdot\log(\sigma)-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\left(1+\frac{\xi y_i}{\sigma}\right), & \xi\neq 0\\ -k\cdot\log(\sigma)-\sigma^{-1}\sum_{i=1}^{k}y_i, & \xi=0\end{cases} \tag{19}$$

其中，结果需要满足$1+\frac{\xi y_i}{\sigma}>0$。

而 GPD 的使用要求确定一个 u 的初始阈值，这需要在偏差和方差之间进行权衡。如果 u 设置太低，那么异常得分阈值的估计将受到偏差的影响，因为极值理论的尾部渐进基础可能被违反。相反，如果 u 太高，只有少量的观测值被用于参数估计，导致方差过高。为了选择合适的 u，文章参照了 Bader 等人[51]的自动顺序测试程序。

本方法的核心思路为假设检验。首先将训练样本异常分数中的 70 和 95 百分位的异常值作为候选阈值的上下界，并且以 1 为增量构建候选阈值集合$\{\tilde{u}_1,\tilde{u}_2,\ldots,\tilde{u}_n\}$。每个候选阈值与一个超出量样本$\tilde{y}_j=\left\{\tilde{\vartheta}_{id}-\tilde{u}_j\middle|\tilde{\vartheta}_{id}-\tilde{u}_j,\ i=1,\ldots,N_d,\ d\in I_1\right\},\ j=1,\ldots,\mathrm{n}$ 对应。第 j 个原假设$H_j^{(0)}$表示为：样本$\tilde{y}_j$满足 GPD。然后构建 Anderson-Darling 统计量进行 AD 拟合优度检验。对公式(9)使用概率积分变换$\hat{\theta}_n,z_{(i)}=G(y_{(i)}|\hat{\theta}_n)$，并排序生成次序统计量$\{z_{(i)}|z_{(1)}<z_{(2)}<\cdots<z_{(n)}\}$。AD 检验的统计量为

$$A_n^2=-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log\left(z_{(i)}\right)+\log\left(1-z_{(n+1-i)}\right)\right] \tag{20}$$

基于计算得到的统计量，进而可以计算每一对$(\tilde{u}_j,\tilde{y}_j)$的 p 值$p_j$。其中，由于缺乏直接可得的 p 值表，AD 检验样本关于 GPD 的拟合优度的 p 值表格需要通过蒙特卡洛模拟完成，这里我们参照了 Bader 等人[51]在 R 中提供的 p 值表。

一般而言，按顺序从$(\tilde{u}_1,\tilde{y}_1)$开始进行 AD 检验，直到某一对$(\tilde{u}_j,\tilde{y}_j)$的原假设$H_j^{(0)}$被接受，则该$\tilde{u}_j$为我们确定的阈值 u。然而，依据 Bader 等人[51]指出，除非 AD 检验精度很高，否则从小往大进行检验时，原假设被接受的情况可能偶然发生在低阈值$H_k^{(0)}$，因此原本应该有效的，大于$\tilde{u}_k$的阈值被舍弃了。当然也可以从阈值$\tilde{u}_l$开始下降，直到有一个原假设被拒绝，但这将导致 I 型错误率增加。为了避免这一问题，文章采用了 G’Sell 等人提出的前进停止规则(ForwardStop Rule)[52]，在进行有序假设检验的情况下，可以将错误发生率控制在β，基于 AD 检验的最优阈值指标为

$$\hat{k} = \max\left\{k \in \{1, \dots, n\}: -\frac{1}{k}\sum_{j=1}^{k}\log(1-p_j) < \beta\right\} \tag{21}$$

最优阈值$u = \tilde{u}_k$，参照 Bader 等人文中的设置，$\beta = 0.1$。

整体而言，在阈值确定这一板块，文章使用的阈值评分估计程序基本上属于非参数方法范畴，适用于不同的样本输入，能够自动计算所有参数比如ξ，σ，自动输出对应的阈值$\hat{\tau}_\alpha$，唯一需要确定的超参数为置信水平α和β，且这两个超参数拥有经验值 0.95 和 0.1。

五、实证分析

（一）实证案例描述

对 NN DTW 模型的有效性评估使用到了 2 个证监会披露案例，案例搜集选取标准为：证监会通过《中国证监会行政处罚决定书》披露公示的有关证券二级交易方面的违规行为，这类行为通常反映了一种或多种异常交易行为。进一步，为了配合文章对《交易监控细则》的研究，优先选取深圳交易所上市企业相关案例。为了提高评估准确性，优先选取对异常交易发生时间节点披露更为详细的案例。同时为了进一步说明模型识别的有效性，文章还将综合总市值，所在行业等因素给出一个或多个可比企业的股票高频数据异常识别结果进行比较。

最终，确定评估用案例的第一个为《中国证监会行政处罚决定书(夏传武)》[5]中有关卓翼科技(002369)的内幕交易事件，横向比较股票为拓邦股份(002139)和美格智能(002881)。披露案例主要内容如下：2019 年 3 月 13 日，夏传武作为公司 2018 年发行股份购买资产事项的内幕信息知情人，在公司拟终止本次重组事项的内幕敏感期内，通过大宗交易卖出其持有的“卓翼科技”无限售流通股 11552730 股，占“卓翼科技”总股本的 1.9919%。依照《交易监控细则》，在异常交易发生期间存在重大价格变动，大致属于“成交数量占期间市场成交总量的比例较高的申报”的拉抬打压股价一类异常。

[5] http://www.csrc.gov.cn/csrc/c101928/c1714744/content.shtml

图 3 发生异常交易事件前后一段时间“卓翼科技”股票走势图

评估用的第二个案例为《中国证监会行政处罚决定书(王威)》[6]中有关东方能源(000958)(如今股票更名为电投产融)的内幕交易事件，横向比较股票为富春环保(002479)。披露案例主要内容如下：2019 年 3 月 22 日期间，属于内幕信息敏感期内，操纵涉案账户卖出其他股票，持续交易“东方能源”，其交易金额、持仓量均较内幕信息敏感期有所放大；同时，上述交易和与另一位内幕信息知情人通话时间高度吻合，并与上市公司停牌时点邻近，交易行为明显异常。依照《交易监控细则》，是典型的内幕交易案例。

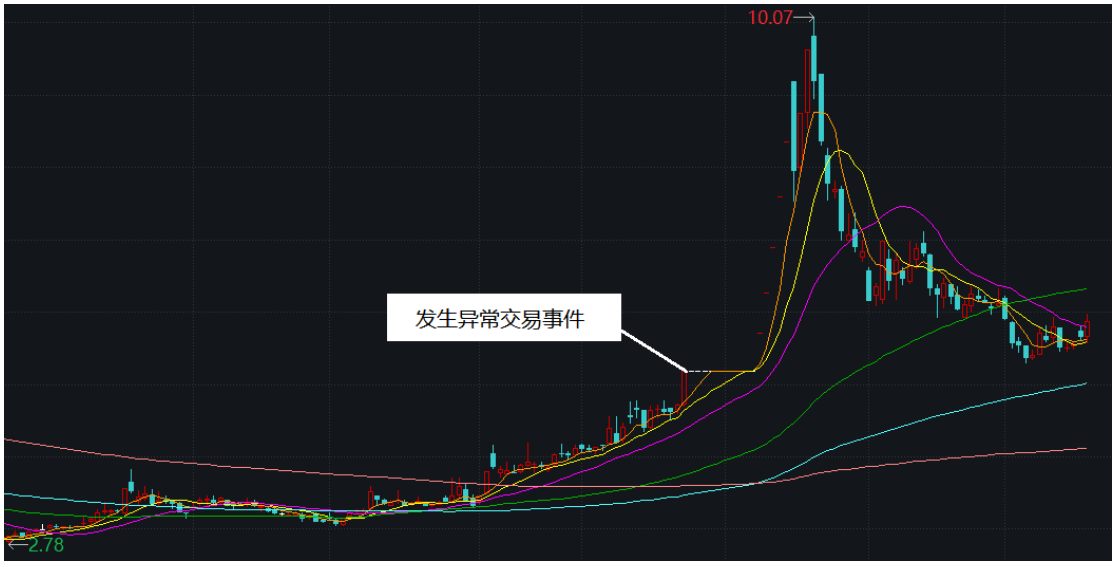


图 4 发生异常交易事件前后一段时间“东方能源”股票走势图

[6] http://www.csrc.gov.cn/csrc/c101928/c1936092/content.shtml

对于这两个案例，异常交易披露日期为 1 天。对于训练集，文章认为，训练集样本使用异常交易日之前 21 个交易日，约为 1 个交易月的交易数据是可行的。因此总共 5 只股票的 21+1 日，共 110 天的交易活动数据成为了文章的评估案例数据。基于前文提及过的前提假设，默认训练集数据中不存在其他异常交易数据。

（二）模型训练效果评价

1. 评价方法的选取

在机器学习算法按照样本和数据是否自带标签，可以分为有监督、半监督和无监督的机器学习。而对于异常检测算法而言，大部分异常检测算法都是无监督的。NN DTW 属于异常检测算法，鉴于评价用的数据集本身不带标签，同样属于无监督机器学习模型。经典的评价方法如 F1 分数、AUC-ROC 曲线在文章中难以实现。且无监督学习算法的好坏评估，比较复杂的，并没有一个确切的标准。对于这种无监督机器学习模型，常用的评价方法内部数据结构和外部人工标签。对于内部数据结构差异，往往使用到了相似度分析，如 k-mean 的方法。文章使用与可比企业的 NN DTW 识别结果的区别说明。而外部人工标签，有研究致力于将无监督学习转化为监督学习，比如 Shebuti 等人[53]，使用了顺序集成的方法不断迭代，希望得到可信度较高的标签。然而采用迭代的方法生成人工标签对算力要求比较大，这种方法暂不考虑。

2. NN DTW 识别结果

本节展示了研究案例的异常值评分，图中每个点都代表了一个逐笔交易窗口，纵轴表示 NN DTW 模型得出的分数，在横轴方向上对样本点进行了排序，使得样本异常得分的整体分布形式得到直观展示。每张图中的横线表示依据极值理论得出的阈值$\hat{\tau}_{\alpha}$。

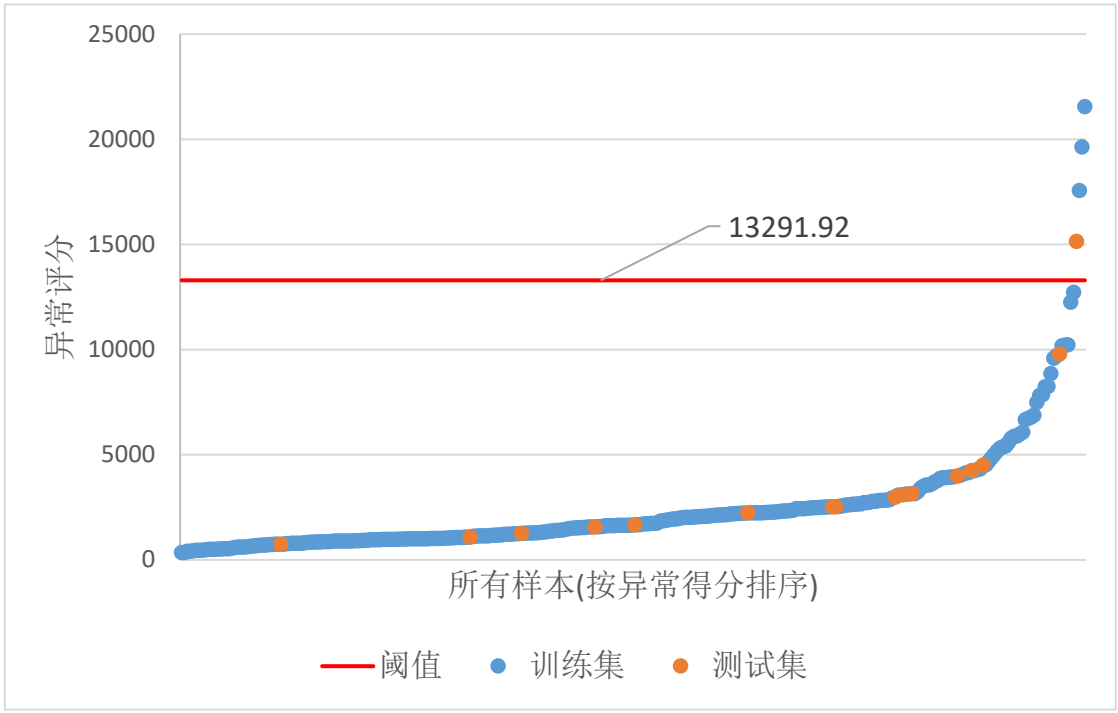


图 5 卓翼科技(002369)异常值评分

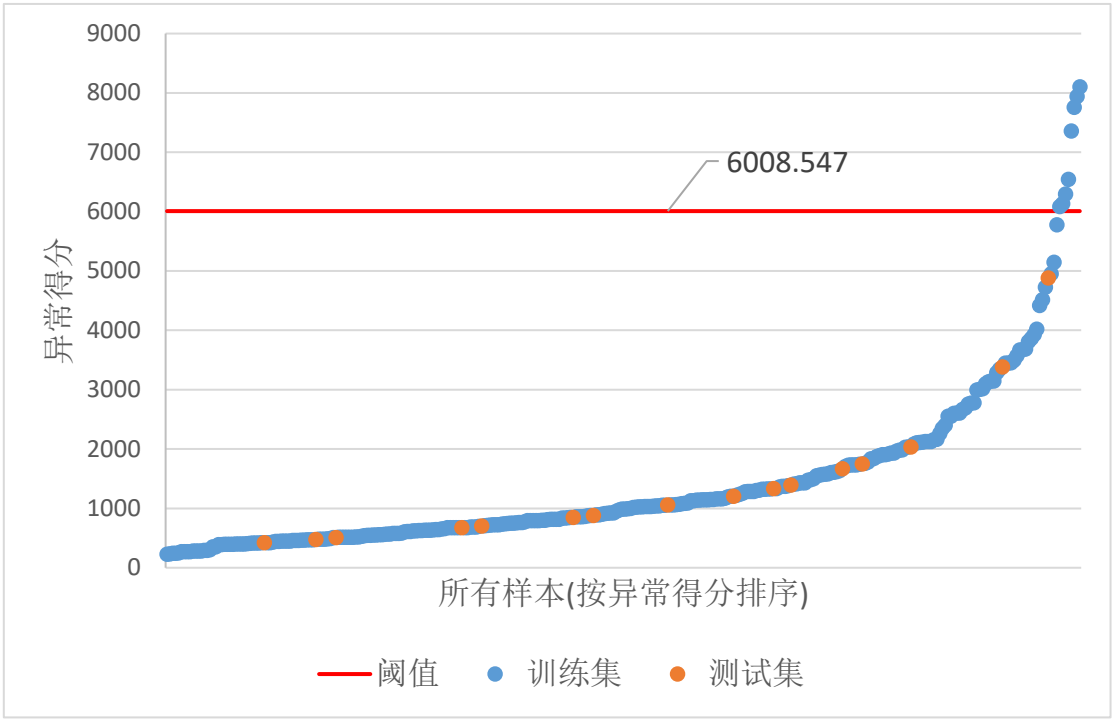


图 6 拓邦股份(002139)异常值评分

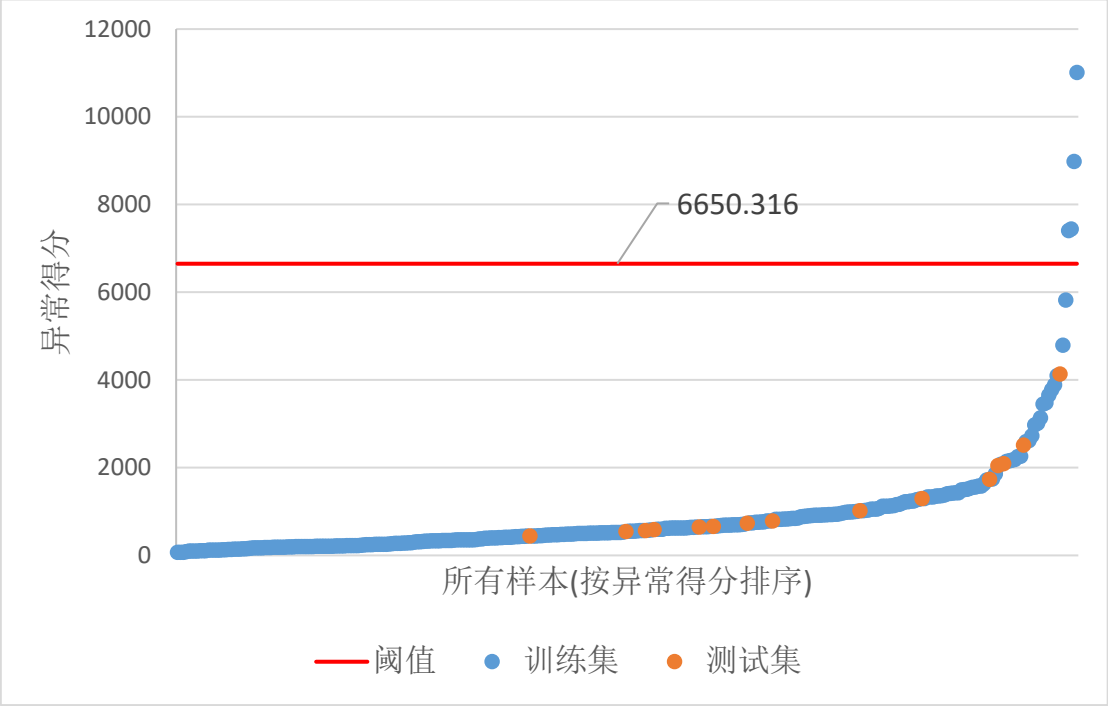


图 7 美格智能(002881)异常值评分

如图 5 所示，基于卓翼科技 2 月 12-3 月 12 日的交易数据确定正常交易序列的判定阈值为 13291.92。在训练集中，仅有 3 个时间窗口的 NN DTW 得分数值大于该阈值，被认定为“异常”。依据我们的假设，认为这三个时间内可能发生了《交易监控细则》所列示的“值得注意”的市场交易行情，但若不存在异常交易动机则不被视作异常交易。而在测试集中，3 月 13 日上午开盘 15 分钟的这段时间内，出现了得分为 15142.114257 的交易序列，超过阈值。由于没有办法从证监会获得准确的异常交易标签，认为模型有可能成功识别出异常交易。而进一步的证明需要结合可比企业的识别结果说明。如图 6，7 所示，相比于训练集得出的阈值，测试集中没有出现得分超过阈值序列。可以认为，一般情况下，每日股票高频数据的异常得分存在超过按照历史数据计算得出的阈值的情况不是常见的，说明卓翼科技的 3 月 13 日股票交易序列存在于异于平日的情况。

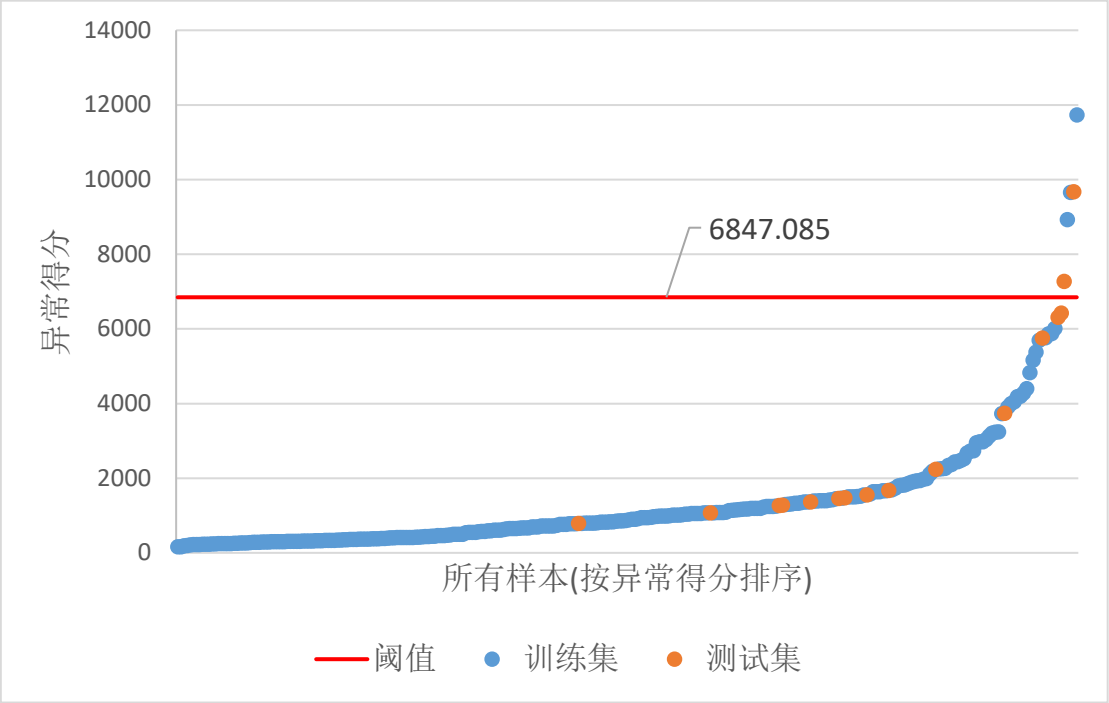


图 8 东方能源(000958)异常值评分

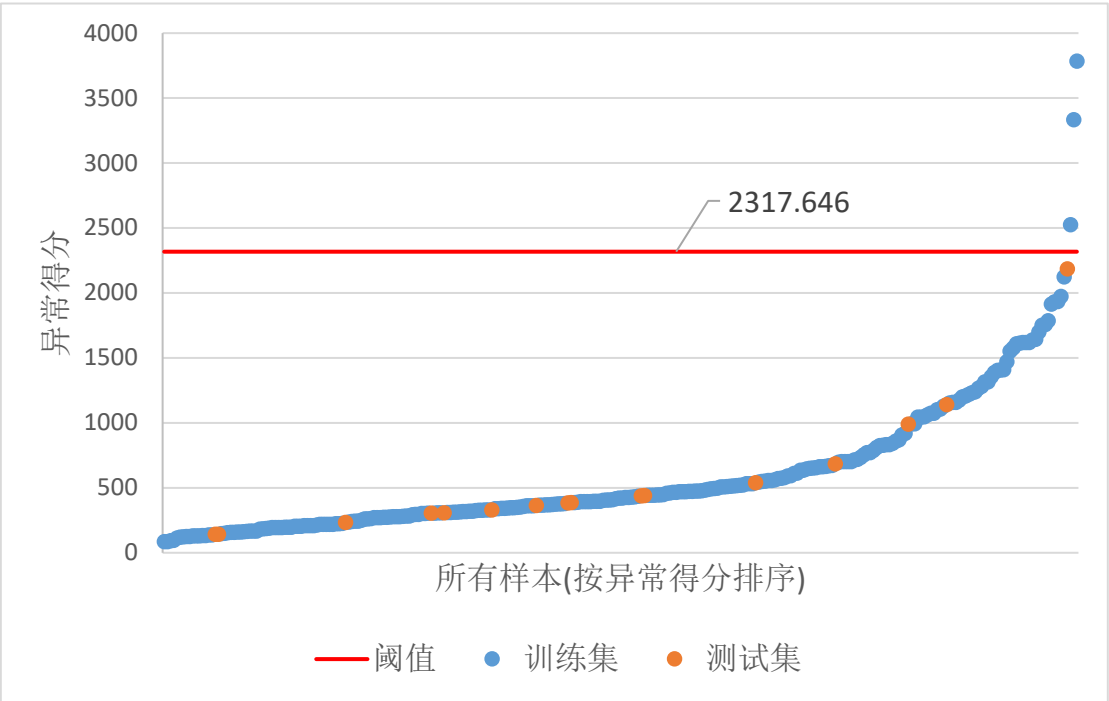


图 9 富春环保(002479)异常值评分

类似地，如图 8 所示，基于东方能源训练集数据得出的序列异常得分阈值为 6847.085，在 3 月 22 日单日期间，共有 2 段时间序列 NN DTW 异常评分超过阈值，而可比企业富春环保在该日则没有评分超过历史阈值的异常序列。

六、总结与展望

鉴于金融市场中不对称信息的普遍存在，准确和及时地识别异常交易是极其重要的。在这项研究中，建立了一个基于机器学习 DTW，与极值理论的金融市场监控模型，有助于异常交易监管人员尽早发现可疑的异常交易。作为一个无监督机器学习模型，本模型不需要获得非法交易的历史例子来进行训练，为进一步优化证券行业标准市场监控系统的提出了可行方案。

这项识别模型需要构建滑动时间窗口的股票交易序列作为被评分对象，在将其每天的订单簿进行处理之后，按照每 15 分钟为标准，划分出多个日内逐笔成交序列。指标构建方面，本研究结合了国内外的实证方法，与中国的政策要求。按照大类进行设置的指标一方面保证了指标设置的合理性，也不会因为过分精细而导致过拟合问题。异常评分的生成基于 NN DTW 方法。在两个交易序列中具有相似特征的交易可能以不同的顺序执行，这个方法有效避免了交易时间错位而可能导致的匹配失败。然后，基于极值理论的阈值划分，考虑了数据的尾部分布。相比于直接使用百分数进行阈值划分，不需要进行参数化的假设，减轻了阈值划分的工作量。相较其他单类机器学习模型如 IsolationForest，本模型优点在于可解释性更高，能够直观理解股票被标注异常的原因。

在模型评价与实证分析中，这项研究分析了 2 个证监会披露的异常交易案例，数据涉及到 110 日的股票交易高频数据。从结果上看，2 个案例都能有效分辨出异常交易存续日期内的交易序列异常。通常情况下，证监会不将公布异常交易订单簿的具体订单异常状态，加上文章利用行情快照挡位数据倒推了一部分没有展示出来的逐笔限价转市价订单，由此没有办法准确为订单簿中的交易订单打上二值标签，因此对逐条订单验证的方法设置了阻碍。

这项研究在若干方面存在改进之处。首先，文章的机器学习研究由于受限于硬件算力，对评估案例的选取限制较大。选择这两个评估案例的原因之一为受限于电脑算力大小。对于作者的笔记本电脑而言，处理 21+1 个交易日的数据已经花费时间超过 0.5 小时/轮。鉴于 DTW 算法的复杂度为$O(n^2)$，对于疑似存在异常交易日期区间为 7 天的案例，若训练集数据量也相应延长至 7 倍，训练时间可能达到 24 小时。更有甚者，异常交易存续期间长达半个月，则更没有办法开展研究。此外，受限于无法获得带标签的数据，无法计算 I 类错误率和准确率等指标，模型较难进行评价。对于国内开展识别股票异常交易的公开研究凤毛麟角，缺少参考也是问题之一。

在未来的研究中，本项研究可以从下列几个方法进行改进与拓展。首先，需要完善对模型的评价部分，按照一般机器学习模型评价框架，在完成新模型的提出之后，还需要将其与其他文献中的模型进行比较，从准确率、误判率等指标方面说明新模型的优越性，在算力与时间条件允许下，可以更广泛地选取实证案例；其次，可以优化采样技术，使用 bagging 等方法进一步提高识别效率；最后，由于研究初衷是为现有监管机构进行异常交易分析提供可行方案，本项研究仅针对内地市场地股票进行了实证分析，而没有对更广泛意义下的跨市场，跨品种的异常交易行为进行研究。在未来的研究中，如果能够有针对账户组的交易数据，NN DTW 方法或许能够应用于识别更复杂的，跨品种与跨市场的异常交易情况。

参考文献

[1] Nuno F , Ferreira M A . Insider Trading Laws and Stock Price Informativeness[J]. Review of Financial Studies, 2009(5):1845-1887.

[2] Cumming D , Johan S , Li D . Exchange trading rules and stock market liquidity[J]. Journal of financial economics, 2011(3):99.

[3] 刘宇. 银行异常交易识别的决策树及其改进算法应用研究[D].武汉理工大学,2018.

[4] Banerjee, Ajeyo, Eckard, et al. Why Regulate Insider Trading? Evidence from the First Great Merger Wave (1897-1903).[J]. American Economic Review, 2001.

[5] Tālis J. Putniņš. Market Manipulation: A Survey[J]. Social Science Electronic Publishing, 2012, 26(5):952-967.

[6] Bagehot W,Treynor J． The only game in town[J]． Financial Analysts Journal, 1971,27: 12-14,22．

[7] Kyle A S．Continuous auctions and insider trading[J]．Econometrica,1985,53: 1315-1335．

[8] Easley D,Hara M O． Price,trade size,and information in securities markets[J]． Journal of Financial Economics,1987, 19: 69-90．

[9] Bhattacharya U , Daouk H . The World Price of Insider Trading[J]. Journal of Finance, 2002, 57(1):75-108.

[10] Guercio D D , Odders-White E R , Ready M J . The Deterrent Effect of the Securities and Exchange Commission's Enforcement Intensity on Illegal Insider Trading: Evidence from

Run-up before News Events[J]. Journal of Law & Economics, 2017, 60(2):269-307.

[11]Kyle A S . Continuous Auctions and Insider Trading[J]. Econometrica, 1985, 53(6):1315-1335.

[12]Huddart S , Hughes J S , Levine C B . Public Disclosure and Dissimulation of Insider Trades[J]. Econometrica, 2001, 69(3):665-681.

[13]Glosten L R , Milgrom P R . Bid, ask and transaction prices in a specialist market with heterogeneously informed traders[J]. Journal of Financial Economics, 1985, 14(1):71-100.

[14]DAVID, EASLEY, MAUREEN, et al. Time and the Process of Security Price Adjustment[J]. The Journal of Finance, 1992.

[15]Kee, H, Chung, et al. Insider Trading and the Bid-Ask Spread[J]. Financial Review, 1998.

[16]Collin-Dufresne P , Fos V . Insider Trading, Stochastic Liquidity and Equilibrium Prices[J]. Econometrica, 2016, 84(4):1441-1475.

[17]Demarzo P M , Fishman M J , Hagerty K M . The Optimal Enforcement of Insider Trading Regulations[J]. 1996(3).

[18]Foster F D , Viswanathan S . Strategic Trading When Agents Forecast the Forecasts of Others[J]. Journal of Finance, 1996, 51(4):1437-1478.

[19]Tavakoli M , Mcmillan D , Mcknight P J . Insider trading and stock prices[J]. International Review of Economics & Finance, 2012, 22(1):254-266.

[20]Olmo J , Pilbeam K , Pouliot W . Detecting the presence of insider trading via structural break tests[J]. Journal of Banking & Finance, 2011, 35(11):2820-2828.

[21]Esen M F , Bilgic E , Basdas U . How to detect illegal corporate insider trading? A data mining approach for detecting suspicious insider transactions[J]. Intelligent Systems in Accounting, Finance and Management, 2019, 26(2):60-70.

[22]Lundblad, Christian et al. Detecting Insider Trading in the Era of Big Data and Machine Learning.[J]. Electronic Journal, 2022.

[23]柴洪峰, 李锐, 王兴建,等. 基于数据挖掘的异常交易检测方法[J]. 计算机应用与软件, 2013, 30(1):6.

[24]王欣, 尹留志, 方兆本. 异常交易行为的甄别研究[J]. 数理统计与管理, 2009(4):7.

[25]Ding H , Trajcevski G , Scheuermann P , et al. Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures[J]. Proceedings of the Vldb Endowment.

[26]Sakoe H , Chiba S . Dynamic Programming Algorithm Optimization for Spoken Word Recognition[J]. IEEE Transactions on Acoustics Speech and Signal Processing, 1978, 26(1):43-49.

[27]Aach, J. and Church, G. (2001). Aligning gene expression time series with time warping algorithms. Bioinformatics. Volume 17, pp 495-508.

[28]Ding, H., G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment 1 (2), 1542–1552.

[29]Gavrila, D. M. & Davis,L. S.(1995). Towards 3-d modelbased tracking and recognition of human movement: a multi-view approach. In International Workshop on Automatic Face- and Gesture-Recognition.

[30]Katz R W , Parlange M B , Naveau P . Statistics of extremes in hydrology[J]. Advances in Water Resources, 2002, 25(8):1287-1304.

[31]R, C, H, et al. A goodness-of-fit test using Moran's statistic with estimated parameters[J]. Biometrika, 1989.

[32]Poon S H , Rockinger M , Tawn J . New Extreme-Value Dependance Measures and Finance Applications[J]. HEC Research Papers Series, 2001.

[33]Fisher R A , Tippett L . Limiting forms of the frequency distribution of the largest or smallest member of a sample[J]. Mathematical Proceedings of the Cambridge Philosophical Society, 2008, 24.

[34]Gnedenko,B.V.. Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire.[J]. *Annals of Mathematics,* 1943, (44): 423

[35]Easley D , O'Hara M . Price, trade size, and information in securities markets[J]. Journal of Financial Economics, 2006, 19(1):69-90.

[36]He H , Wang J . Differential Information and Dynamic Behavior of Stock Trading Volume[J]. Social Science Electronic Publishing.

[37]Hendershott T , Jones C M , Menkveld A J . Does Algorithmic Trading Improve Liquidity?[J]. CFS Working Paper Series, 2008, 66(1):1-33.

[38]Attanasio O P . Asset price volatility and information structures[J]. Economics Letters, 1990, 33(2):159-164.

[39]Kacperczyk, Marcin, Pagnotta, et al. Chasing Private Information[J]. Cepr Discussion

Papers, 2015.

[40]Zhang L , Mykland P A , Ait-Sahalia Y . A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data[J]. National Bureau of Economic Research, Inc, 2003(472).

[41]Pattern A J , Sheppard K . Optimal combinations of realised volatility estimators[J]. International Journal of Forecasting, 2009, 25(2):218-238.

[42]Bandi F M , Russell J R . Microstructure Noise, Realized Variance, and Optimal Sampling[J]. Review of Economic Studies, 2008, 75(2):339-369.

[43]Shokoohi-Yekta M , Hu B , Jin H , et al. Generalizing DTW to the multi-dimensional case requires an adaptive approach[J]. Data Mining & Knowledge Discovery, 2017, 31(1):1-31.

[44]Li Z . Exact Indexing of Time Series under Dynamic Time Warping[J]. 2020.

[45]Keogh E , Ratanamahatana C A . Exact indexing of dynamic time warping[J]. Knowledge and Information Systems, 2005, 7(3):358-386.

[46]Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2013.

[47]Lemire D . Faster retrieval with a two-pass dynamic-time-warping lower bound[J]. Pattern Recognition, 2009, 42(9):2169-2180.

[48]Pickands J. Statistical inference using extreme order statistics[J]. Ann.Statist, 1975,3:119-131.

[49]Balkema A A , Haan L D . Residual Life Time at Great Age[J]. The Annals of Probability, 1974, 2(5):792-804.

[50]Stuart Coles.An Introduction to Statistical Modeling of Extreme Values[M]. London: Springer, 2002.

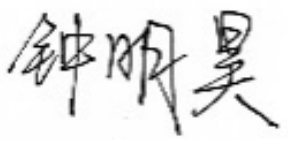
[51]Bader J . AUTOMATED THRESHOLD SELECTION FOR EXTREME VALUE ANALYSIS VIA ORDERED GOODNESS-OF-FIT TESTS WITH ADJUSTMENT FOR FALSE DISCOVERY RATE[J]. The Annals of applied statistics, 2018, 12(1).

[52]Wager, Stefan, Tibshirani, et al. Sequential selection procedures and false discovery rate control[J]. Journal of the Royal Statistical Society, 2016, 78 (2), 423–444.

[53]Shebuti, Rayana, Leman, et al. Less is More: Building Selective Anomaly Ensembles: ACM, 10.1145/2890508[P]. 2016.

中央财经大学本科毕业论文（设计）原创性声明

本人郑重声明：所提交的毕业论文（设计）《基于 DTW 的异常交易识别》，是本人在指导老师的指导下独立进行研究工作所取得的成果。除文中已经注明引用的内容外，不含任何其他个人或集体已经发表或撰写过的作品成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。对文章研究/设计做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果，如违反有关规定或上述声明，愿意承担由此产生的一切后果。

作者签名：  　　2023 年　04　月　26 日