中信期货
CITIC Futures

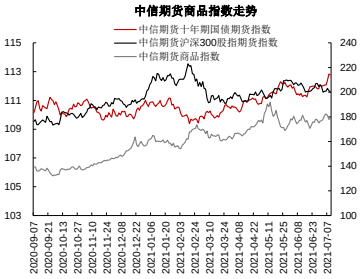
中信期货研究|量化研究报告

深度学习对基于限价订单簿的择时模型优化(一)

投资咨询业务资格:
证监许可【2012】669 号

报告要点

限价订单簿是基于市场限价单的信息汇总，限价订单簿中基本包含以下信息:买卖双方价格及成交量，双方价格的排布按照递减及递增的顺序依次展示。一定程度上，限价订单簿的走向既能反映市场的微观结构，又能为投资者对未来价格变动预测提供信息基础，本篇报告基于深度学习网络的方法论对不同的神经网络生成的高频择时信号准确率进行纵向比较及优化，最后选出表现较好的策略模型。后期报告研究除模型优化外，将研究模型在商品及金融标的上的择时应用。



摘要:

本篇报告，我们根据牛津大学量化金融实验室提出的基于**限价订单簿**的**高频择时**策略算法，分析来自金融交易所(开源)的高频限价订单簿数据，该类数据反映了市场中某类标的资产价量的微观结构。来自金融交易所的订单簿具有**大规模、序列高频、价格-时间变化区间短**等特点。基于此类限价订单簿的时间序列数据作为神经网络的输入数据，可以为预测未来相关标的价格涨跌提供数据基础，本报告所使用的数据集中包含超过 400 万个限价订单数据，基于构建的深度神经网络模型，我们验证了不同神经网络在数据集上预测的表现，实验表明经过空间信息(**卷积神经网络**)以及时间序列信息提取(**循环神经网络 LSTM**)的神经网络表现(预测准确率)优于基础的多层感知机模型，后期报告将继续分析序列对序列(Seq2Seq)、基于注意力机制以及在引入强化学习算法后神经网络择时的优化表现。

本篇报告主要目的在于初步研究基于订单簿的模型的预测准确率，后期优化模型有机会将对商品及金融期货品种进行高频择时支持。

本篇报告逻辑:

第一部分:基于限价订单簿高频择时策略背后逻辑、报告订单簿数据结构剖析、报告订单簿数据部分可视化。

第二部分:各类神经网络简介、不同神经网络模型在订单簿数据上高频择时策略表现及解读。

第三部分:总结不同模型表现背后金融学逻辑、择时模型适用性性分析对后期高频策略支持展望。

风险提示:1)模型参数失效， 2)模型过拟合。

商品量化组

研究员:
魏新照
021-80401773
weixinzhao@citicsf。com
从业资格号 F3084987
投资咨询号 Z0016364

目 录

图表目录

一、限价订单簿结构剖析及择时策略背景

(一)限价订单簿数据剖析

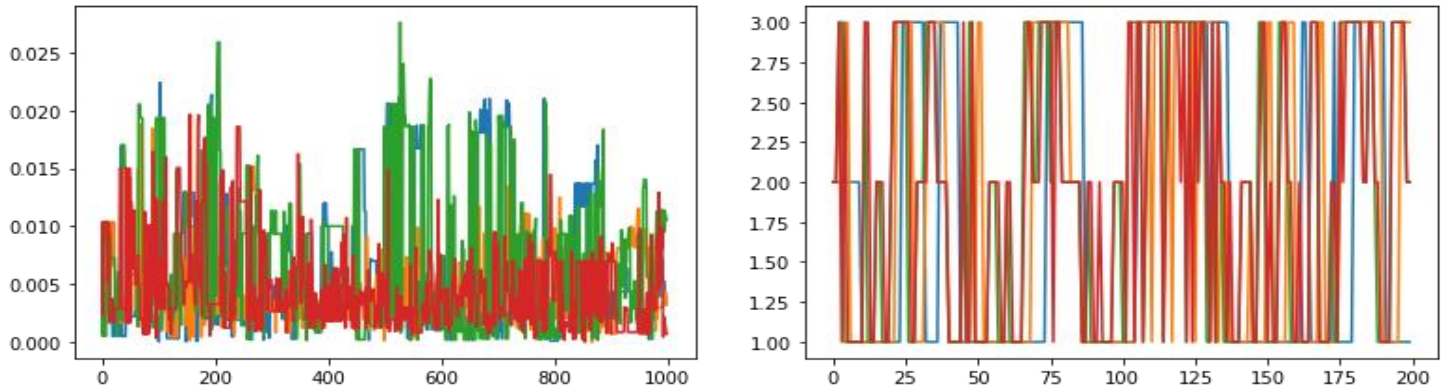
随着科技的发展，现代动态金融市场为金融投资者提供瞬时买低卖空的机会，在这种情况下，为了及时抓住转瞬即逝的交易机会，金融市场交易者必须及时且全面的了解市场动向，并在发现机会后快速采取行动以最小风险实现高利润。

此外，现代交易所每日自动化交易极大地增加了每天发生的交易量，交易所收集并整理这些大量交易所产生的信息并创建全面的交易日志进而呈现出完整的订单簿数据， 这些订单簿数据包含相应时间戳下买卖双方的价量数据，但是此类数据具有高噪声-信息比率，并且国内相关标的订单簿数据时间戳并非连续，因此要想产生有价值的信号，需要经过深度清洗并以此产生源于价量数据的衍生信息，并进一步可用于预测市场变化，而算法又可以使用这些信号在实时交易中做出正确的决策。

本篇报告所使用限价订单簿数据可视为多维面板数据(Multi-Dimension Panel Dataset)，相较于其他量化策略所使用的因子数据(单一时序或者截面数据)，订单簿数据除具有大规模、序列高频、 价格-时间变化区间短等特点，同时具有空间以及时间特性， 因此可以提取跟多相关信息。本篇报告采用的是经过深度清洗的订单簿数据，后期其他标的策略均可依据相应的格式清洗数据并得到相关模型。

本报告所用数据包含五只标的并采用了十天的交叉验证方法，因此使用者可以为每个标准化设置找到九个（交叉折叠）数据集以进行训练和测试。 每个训练和测试数据集都包含所有标的的信息。图表一显示了部分订单簿数据的可视化结果。

图表 1：订单簿部分数据可视化结果



资料来源：Fairdata

（二）模型策略简介

本片报告所着眼的核心前期在于应用深度学习中不同的网络对基于限价订单簿数据的高频择时策略的表现，报告的模型不仅着重分析多层感知机、卷积神经网络（CNN）、循环神经网络(LSTM）的单一深度学习方法，后期报告会分析序列对序列(Seq2Seq）、单头以及多头注意力机制(Attention）算法对单一神经网络进行优化，以上方法方法可用于从大规模高频限价单数据中预测未来的中间价格走势(后期报告将优化为微观价格)。报告同时着眼于模型对标的单步以及多步预测的精度结果。前期模型结果均为随机学习率，后期模型纵向对比对齐不同学习率总结模型。为方便对比，本报告神经网络层数均设定为三层。

模型的基本逻辑:以限价订单簿数据作为训练数据对未来时刻标的价格涨跌的分类问题。

二、模型策略表现对比

经过研究，本部分将主要类型神经网络在订单簿数据上的预测表现进行汇总对比。本部分的顺序依据多层感知机(MLP)、单独卷积神经网络(CNN)、单独循环神经网络(LSTM)、卷积神经网络嵌套循环神经网络(CNN+LSTM)、以及序列模型优化模型表现。

（一）多层感知机(MLP)

(1) 模型简介

多层感知机（Multilayer Perceptron，缩写 MLP）模型是一种前向结构的人工神经网络(维基百科)，清洗好的订单簿数据进入神经网络后进入各个单元，经过神经元内预先设置好的非线性激活函数得到本层神经元的输入(下一层神经元的输出)。具体计算顺序如下:

①:假设输入订单簿数据为X。

②:订单簿数据进行计算得到相应预测值 Y，$Y_1^{[1]} = (W_1^{[1]})^T * X + b_1^{[1]}$。

③:经过激活函数产生下层神经网络的输入 $a_1^{[1]} = \sigma(Y_1^{[1]})$。

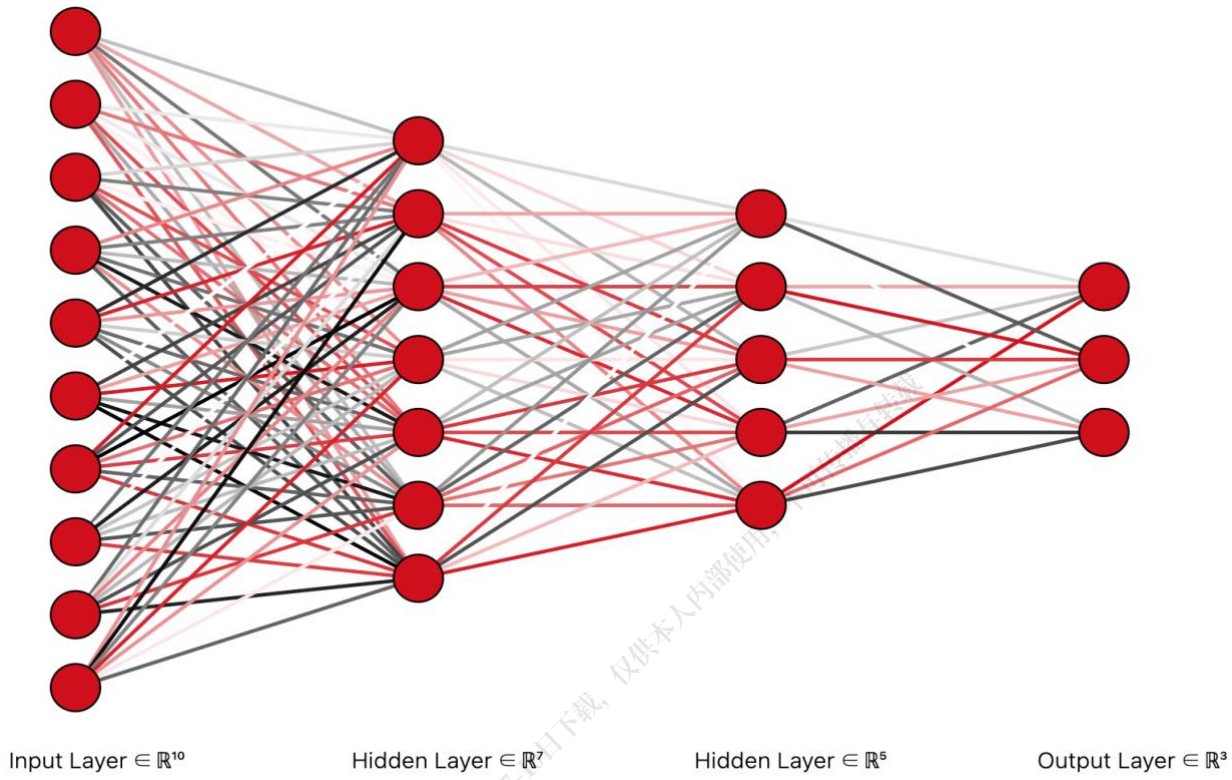
④:以此类推进入全连接层通过 $Softmax$ 函数输出最后分类的概率分布。

⑤:模型训练主要目标在于找到神经网络参数使得损失函数最小化。目前分

类问题常见损失函数为$Cross-entropy$损失函数。

$$L(y_{hat}-y)=-(y*logy_{hat})-(1-y)*\log(1-y_{hat})$$

图表 2：多层感知机模型缩略图

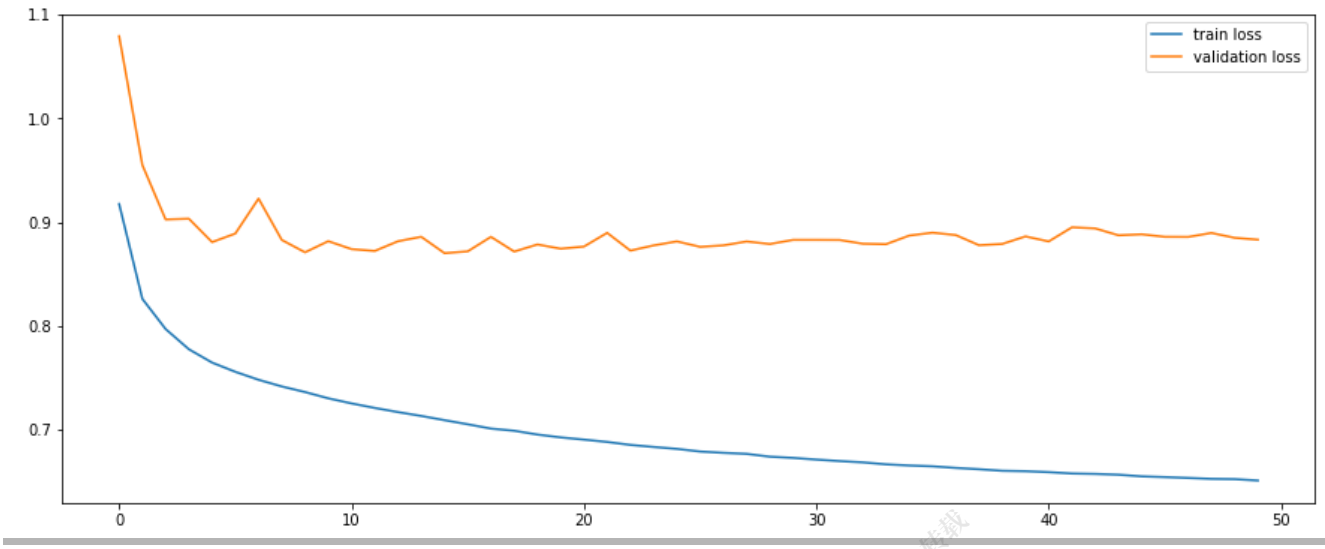


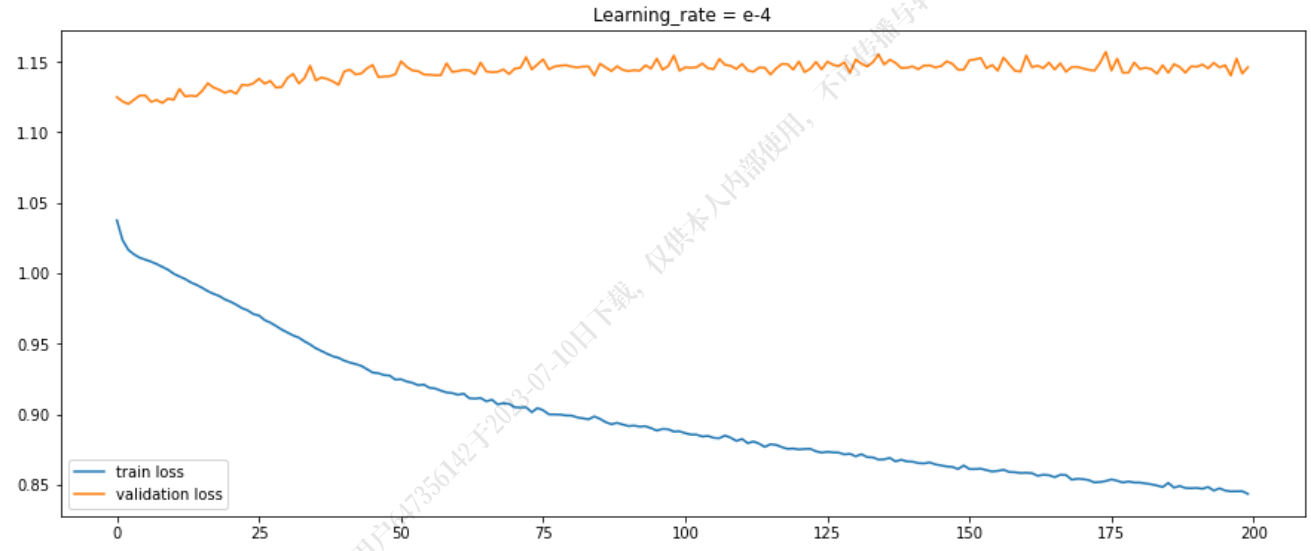
资料来源：中信期货研究所

(2) 模型表现

在训练模型时，我们着重考虑不同学习率在模型收敛以及在该学习率下最优模型的准确率：

图表 3：多层感知机模型随机学习率基本表现





学习率	最优模型表现(MLP)
e-5	30.67%
5*e-5	34.39%
e-4	33.27%
1.5*e-4	31.59%
2*e-4	**35.71%**

资料来源：Wind 中信期货研究所

(3) 模型分析

根据图表三中的模型表现， 多层感知机模型存在模型 1)过拟合、2)预测准确率低等缺点。模型过拟合原因在于训练数据与模型参数之间存在数量不对等问题。模型训练集共有 203800 观测序列，经过分析，模型参数量：**Total params: 4796483** 已经达到观测序列的 24 倍， 因此想要解决问题可以降低模型复杂度，或增加训练集数据量(今后存在优化可能)。模型准确率低原因在于多层感知机模型在训练过程中的全连接层会把数据拉成一维向量，这样会使模型获得信息大打折扣，既损失空间信息，又损失时序信息，因此后期优化因着眼于多提取空间以及时序信息。

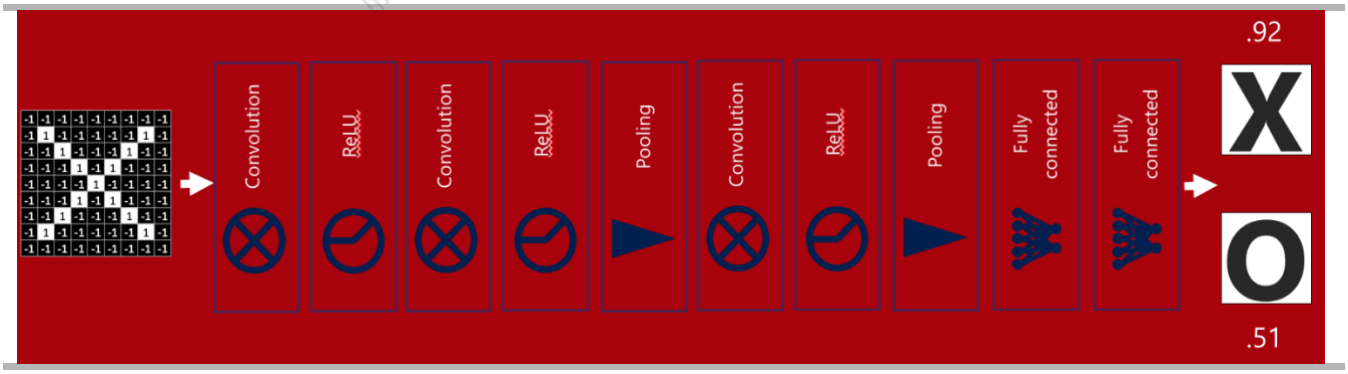
（二）卷积神经网络(CNN)

(1) 模型简介

卷积神经网络主要应用于计算机视觉领域，在图片处理中卷积神经网络能够起到重要作用。卷积神经网络的核心在于:1，能够有效地提取输入数据的底层和上层空间信息，2，在保留数据的信息的基础上能够将高维数据转化为较低维度的数据。相比于多层感知机模型，卷积神经网络模型的参数量有了大大减小，并且卷积神经网络易于做 GPU 的并行计算。

卷积神经网络在分类问题的主要流程在于经理系列处理后， 模型提取输入数据的特征，最终获得输出类别的概率分布。在手写体是别的的应用如下表所示:

图表 4：卷积神经网络分类问题缩略图



资料来源：Brandon Rohrer's Blog

相比多层感知机模型,应用于订单簿数据的卷积神经网络旨在输入数据中中提取更多的空间信息，进而在预测准确率上取得提升。

相比标准神经网络，对于大量的输入数据，卷积过程有效地减少了 CNN 的

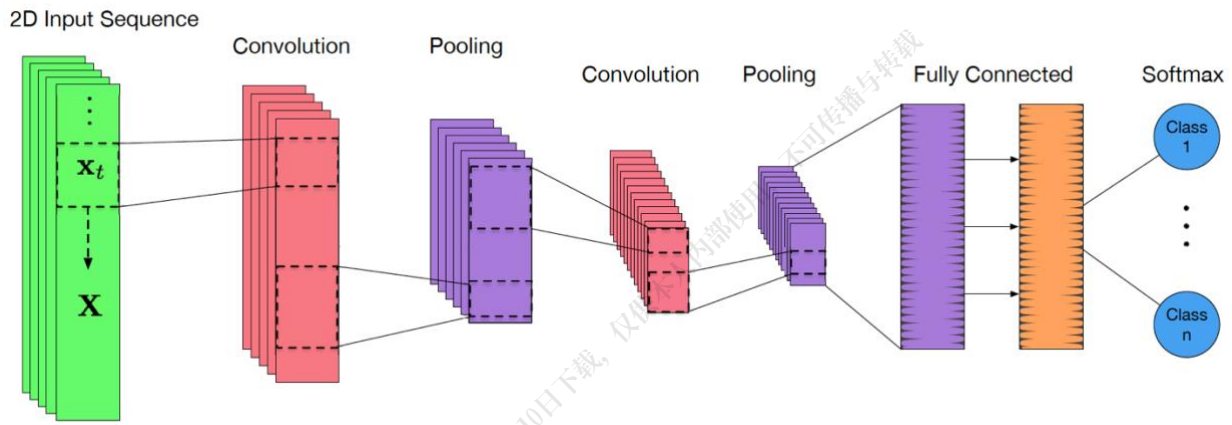
参数数量，原因有以下两点：

参数共享（Parameter sharing）：特征检测如果适用于图片的某个区域，那么它也可能适用于图片的其他区域。即在卷积过程中，不管输入有多大，一个特征探测器（滤波器）就能对整个输入的某一特征进行探测。

稀疏连接（Sparsity of connections）：在每一层中，由于滤波器的尺寸限制，输入和输出之间的连接是稀疏的，每个输出值只取决于输入在局部的一小部分值。（**摘自吴恩达深度学习**）

基于卷积神经网络的订单簿优化模型流程图展示如下：

图表 5：卷积神经网络分类问题缩略图

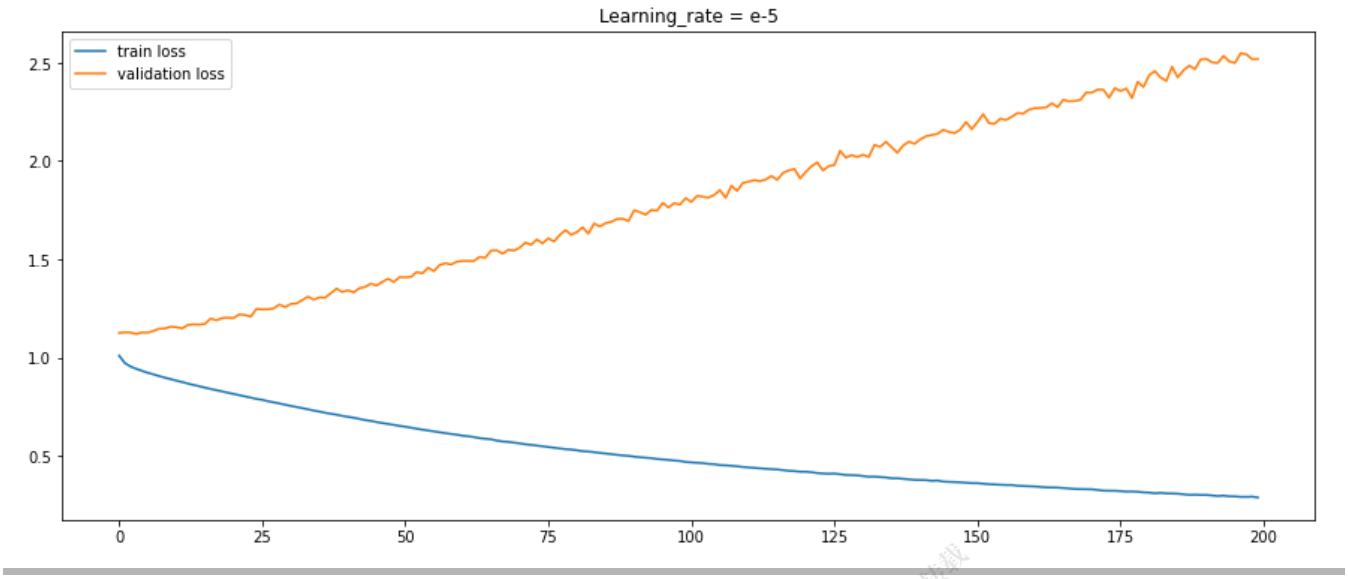


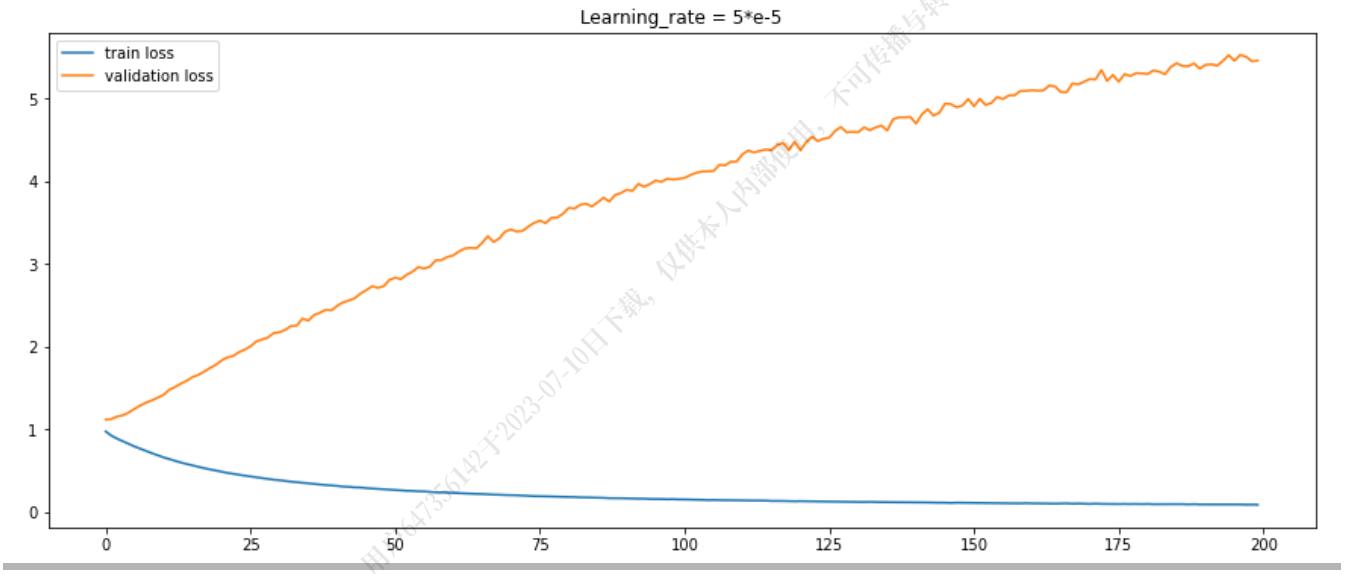
资料来源：牛津大学量化金融实验室

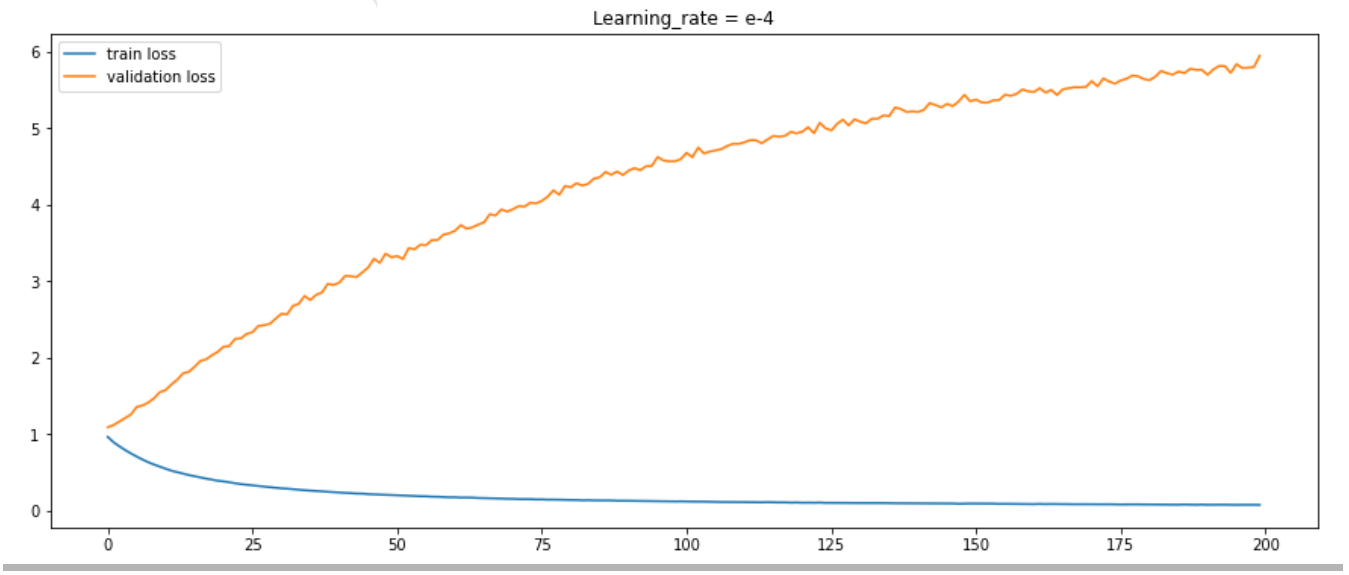
(2) 模型表现

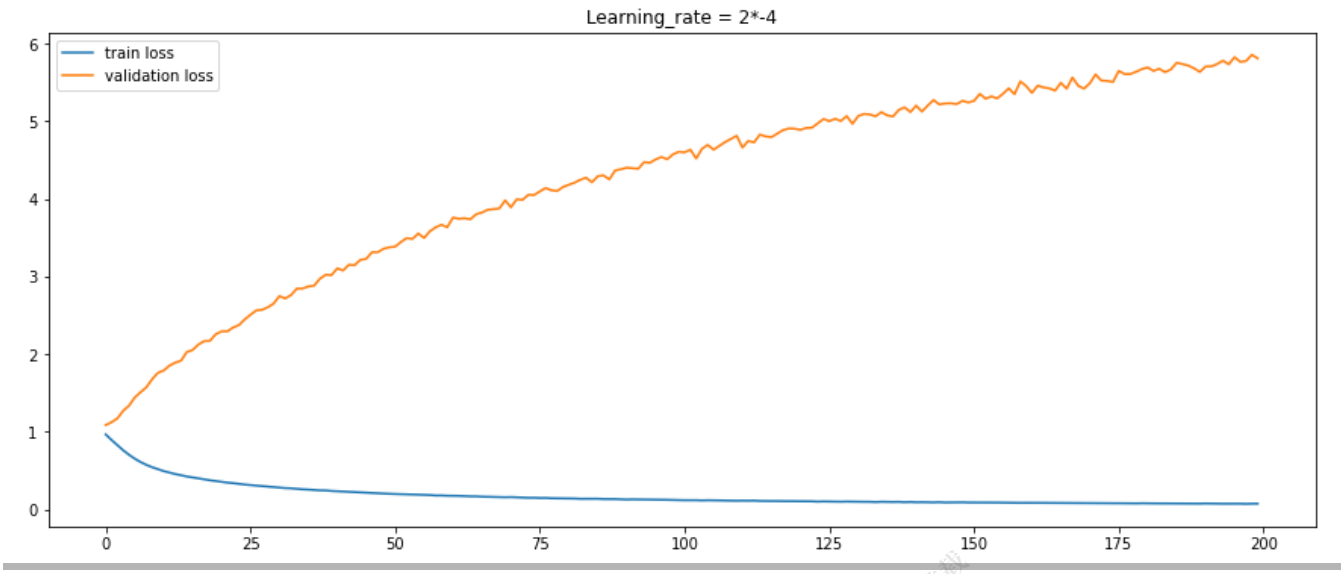
卷积神经网络在构造的神经网络架构的基础上选取不同的超参数(Learning rate)下的训练结果展示如下(报告前期已对相关其他参数进行优化)(前期优化参数包括:卷积核大小，步长等)：

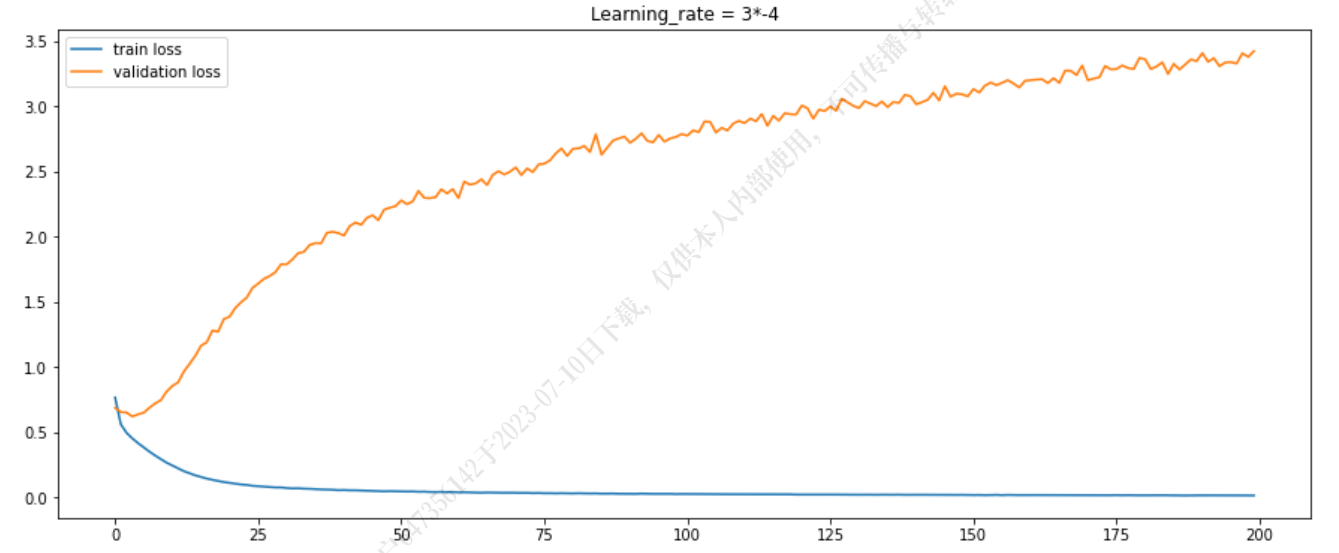
图表 6：卷积神经网络基本表现











学习率	最优模型表现(CNN)
e-5	45.53%
5*e-5	44.72%
e-4	43.57%
1.5*e-4	43.28%
2*e-4	**75.98%**

资料来源：中信期货研究所

(3) 模型分析

经过卷积神经网络的空间信息提取后， 可以发现模型预测的整体表现出明显提高。但整体准确率还是处于较低水平，因此需进一步挖据订单簿内在信息。卷积神经网络在处理图片时可以做到提取输入数据的空间信息，但是对时序信息，鉴于金融类信息高时序性的特点，卷积神经网络不能很好地提取相关信息。因此本报告将选取可以提取时序信息的神经网络，进而分析其表现。

（三）循环神经网络(LSTM)

如果说卷积神经网络可以有效地处理空间信息，那么本部分所描述的的循环神经网络（recurrent neural network，RNN）则可以更好地处理序列信息。 循环神经网络通过引入状态变量存储过去的信息和当前的输入，从而可以确定当前的输出。

(1) 模型简介

本片报告所使用的是含有隐状态的循环神经网络，每个序列输入对应下的隐状态都包含之前序列的相关信息，模型基于对应输出的隐状态预测未来价格涨跌。

以下图片为循环神经网络逻辑图:输入序列数据通过系列计算得到隐状态以及下一序列的输出。

通过输入数据 X，每个序列训练三份逻辑门:

输入门:$I_t = \sigma(X_t W_{xj} + H_{t-1} W_{hi} + b_i)$， $I_t \in \mathbb{R}^{n \times h}$

输出门:$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$， $O_t \in \mathbb{R}^{n \times h}$

遗忘门:$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f)$， $F_t \in \mathbb{R}^{n \times h}$

通过候选记忆元与上一时刻隐状态得出记忆元(记忆元中包含之前序列数据信息):

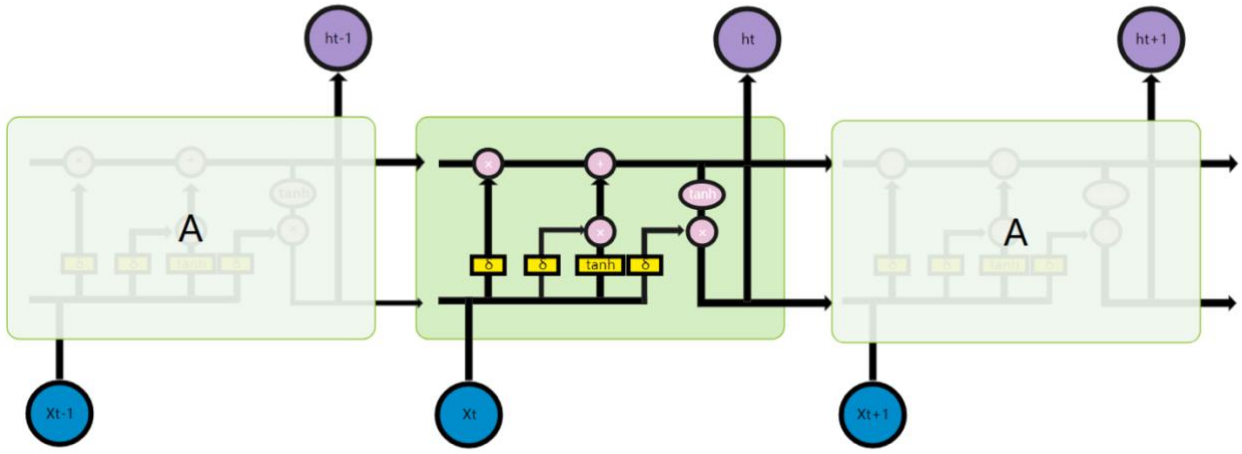
候选记忆元: $\widetilde{C_t} = tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c)$

记忆元:$C_t = F_t \odot C_{t-1} + I_t \odot \widetilde{C_t}$

最后得出用于预测的隐状态:$H_t = O_t \odot \tanh(C_t)$

由公式可知: 只要输出门接近，我们就能够有效地将所有记忆信息传递给预测部分， 而对于输出门接近，我们只保留记忆元内的所有信息，而不需要更新隐状态。

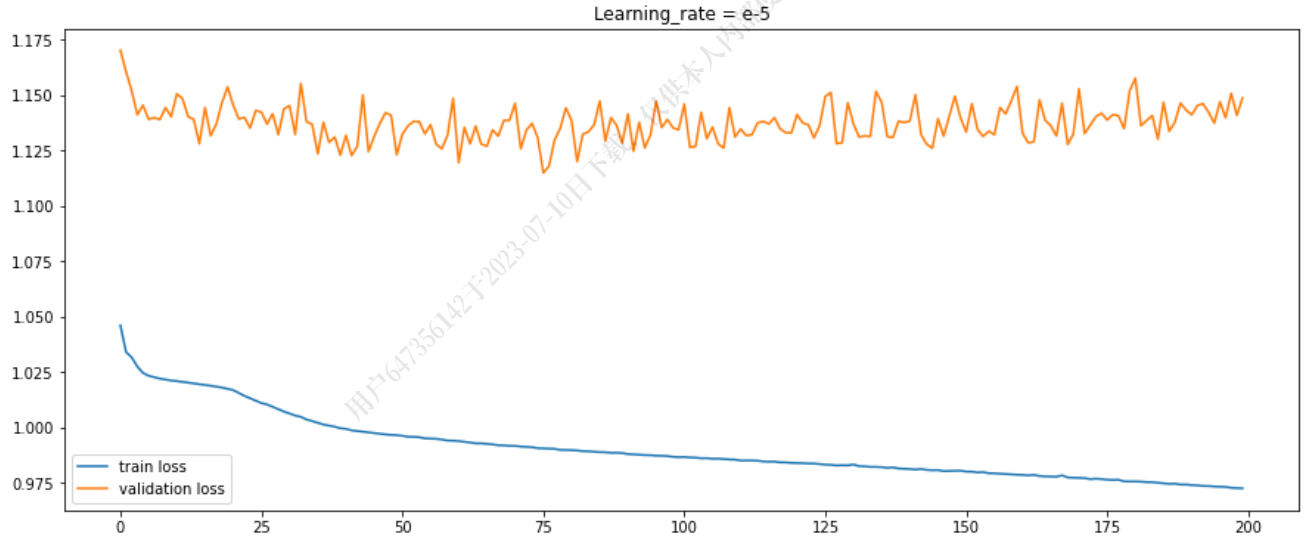
图表 7：循环神经网络逻辑图

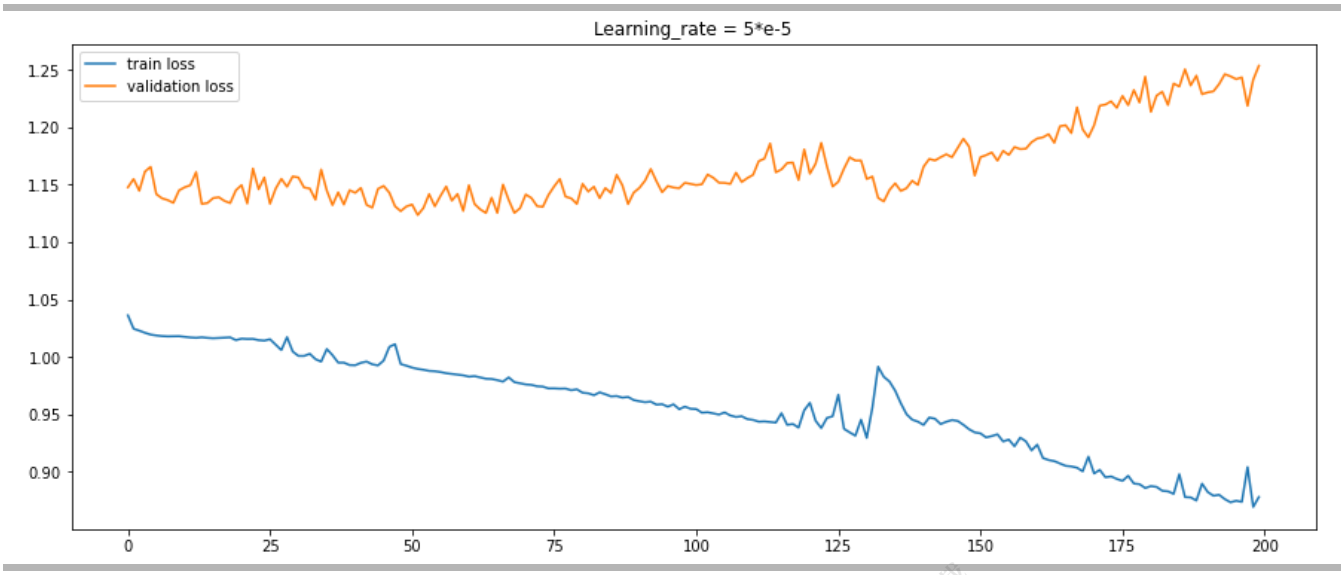


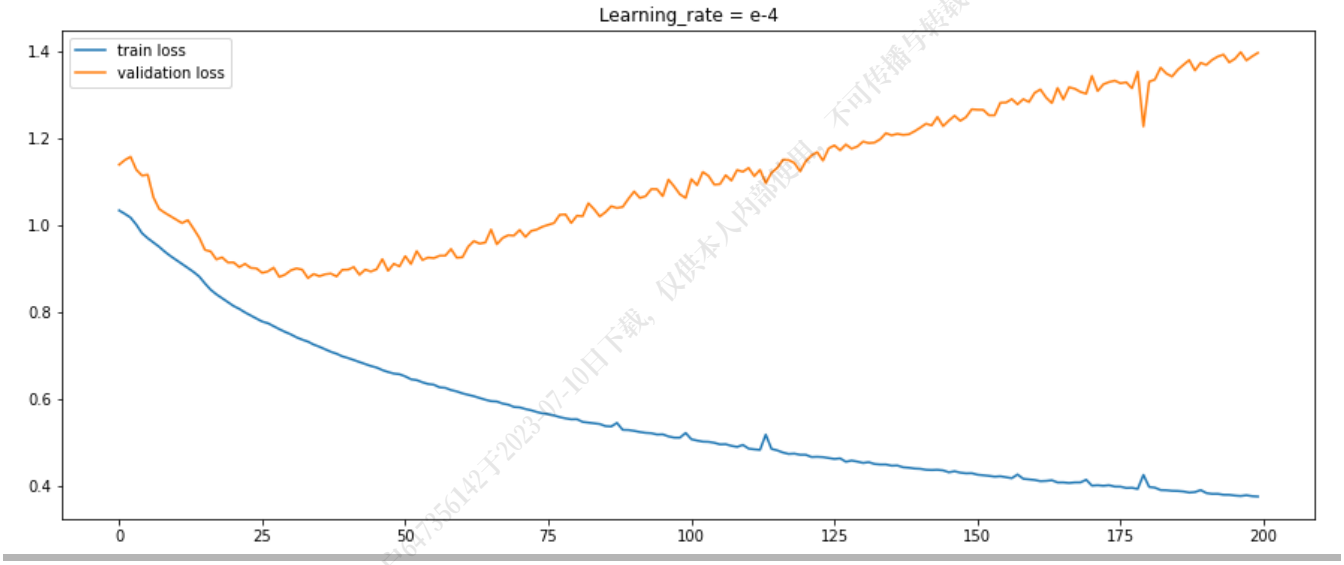
资料来源：中信期货研究所

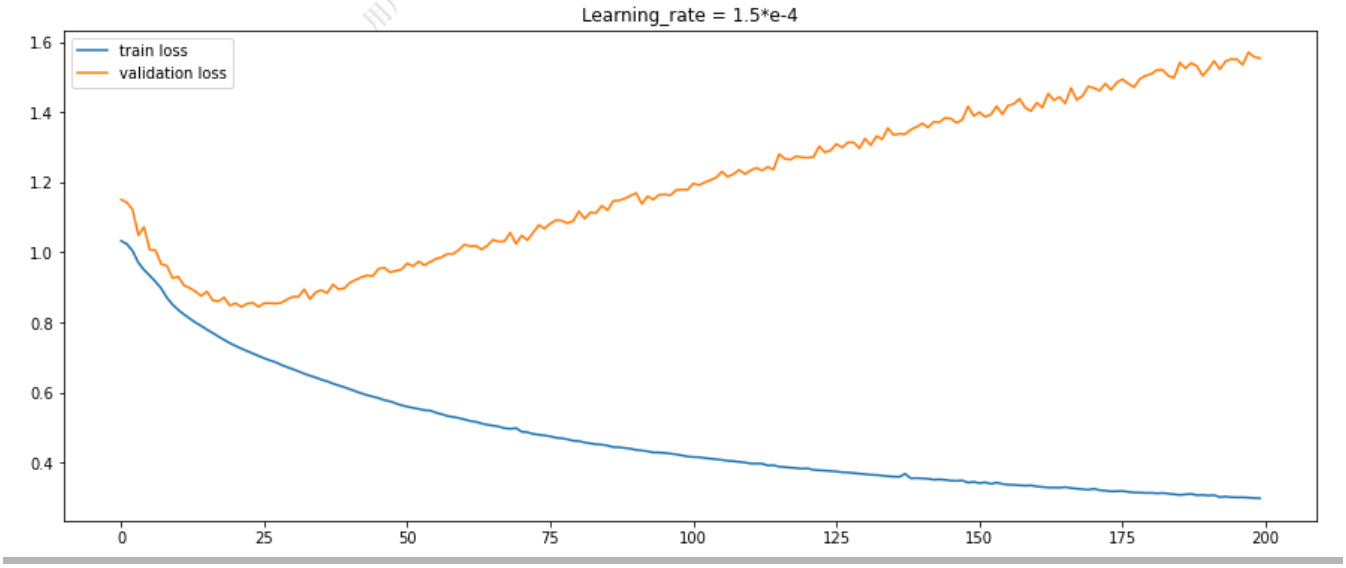
(2) 模型表现

图表 8：循环神经网络(LSTM)模型表现汇总

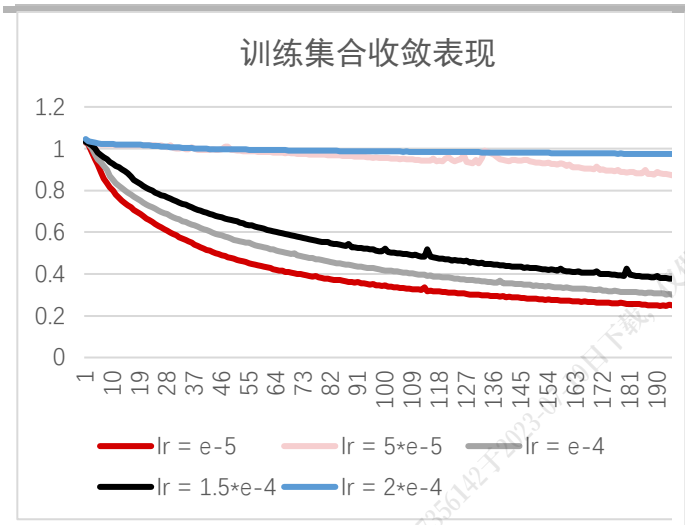


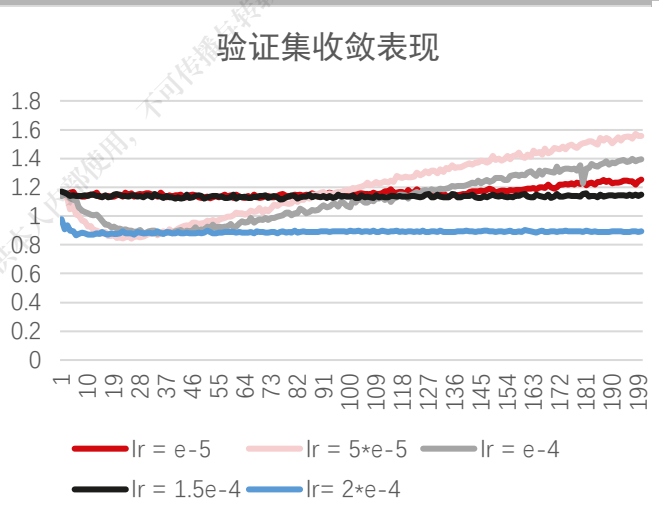






学习率	最优模型表现(LSTM)
e-5	42.33%
5*e-5	48.32%
e-4	52.77%
1.5*e-4	60.54%
2*e-4	**63.93%**





资料来源：中信期货研究所

（3） 模型分析

经过训练后可以发现两点:1，循环神经网络相比多层感知机和卷积神经网络的平均表现有了显著提高(个学习率下最优表现平均提升 5%-10 左右)， 2，学习率选取在 5*e-5~2*e-4 范围内训练收敛趋势明显， 3， 在金融类信息分析的背景下序列信息相比空间信息对预测准确率的提升有更加明显的贡献程度。

但是单独的卷积神经网络或者循环神经网络的最佳模型所实现的预测准确率均存在较大进步空间。报告下一部分将进行空间与序列信息的综合提取进而对预测准确率的提升有所帮助。

（四）卷积神经网络连接循环神经网络(CNN_LSTM)

（1） 模型简介

相比卷积神经网络在限价订单簿上的择时表现， 循环神经网络的表现可以说有了明显提高， 但是模型的泛化能力收到模型参数与训练数据不平衡以及特征挖掘不充分双向影响的影响，本部分将综合报告前几部分的模型进行进一步优化， 在先提取空间信息的基础上进行序列信息的提取。

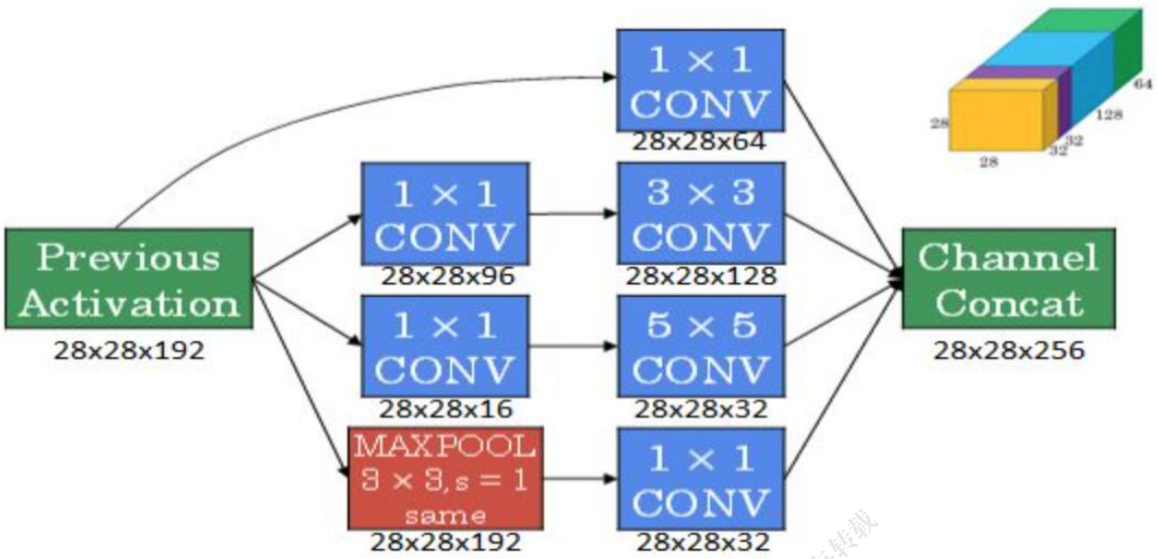
但是前期卷积神经网络模型仅仅依赖一类卷积核函数作为空间信息提取的手段， 这不免对订单不得空间信息挖掘不充分，在本部分的嵌套模型中，卷积神经网络部分将应用多层卷积核对订单簿空间信息进行充分提取，在此引入 Inception Module 作为丰富空间信息提取的手段。Inception Module 在提升空间信息挖掘的同时也可以极大地减少计算机的计算成本。这里借用**吴恩达深度学习**中的例子作为讲解:

假设输入数据为 28x28x192，存在 32 个滤波器 5x5x192，得到输出 28x28x32 的因此需要计算 28x28x32 个数据，对于每个数都需要计算 5x5x192 次，因此整体一轮下来计算量达到 8x28x32x5x5x192 = 1.2 亿次。

但是如果在卷积前先进行一次 1x1 的**瓶颈层**（Bottleneck layer）计算，计算开销将极大缩小。例如同样的输入维度， 先进行一次 1x1 卷积把 192 的通道数减少到 16，再进行 5x5 卷积得到 28x28x32 的输出计算量将变为 28x28x192x16 + 28x28x32x5x5x15 = 1.24 千万，相比 1.2 亿次的计算，这样的计算量减少将近 90%。因此，只要合理构建瓶颈层，就可以既显著缩小计算规模，又不会降低网络性能。

本部分中所实现的网络嵌套是，前部分所实现的即用 Inception Module 提取的多层空间信息。

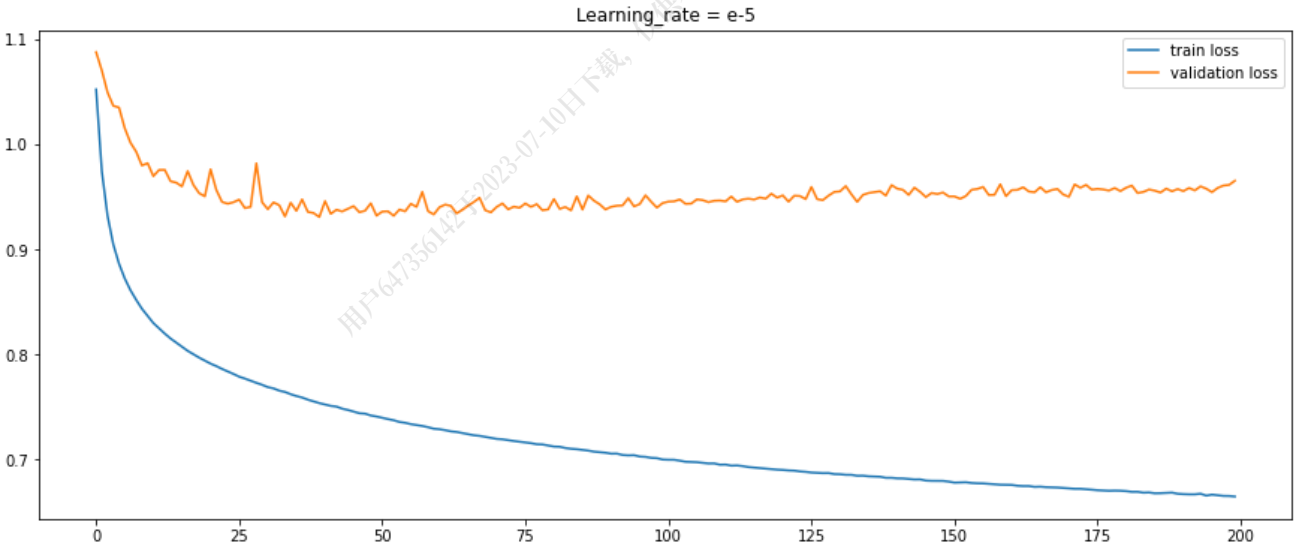
图表 9：Inception Module 逻辑图

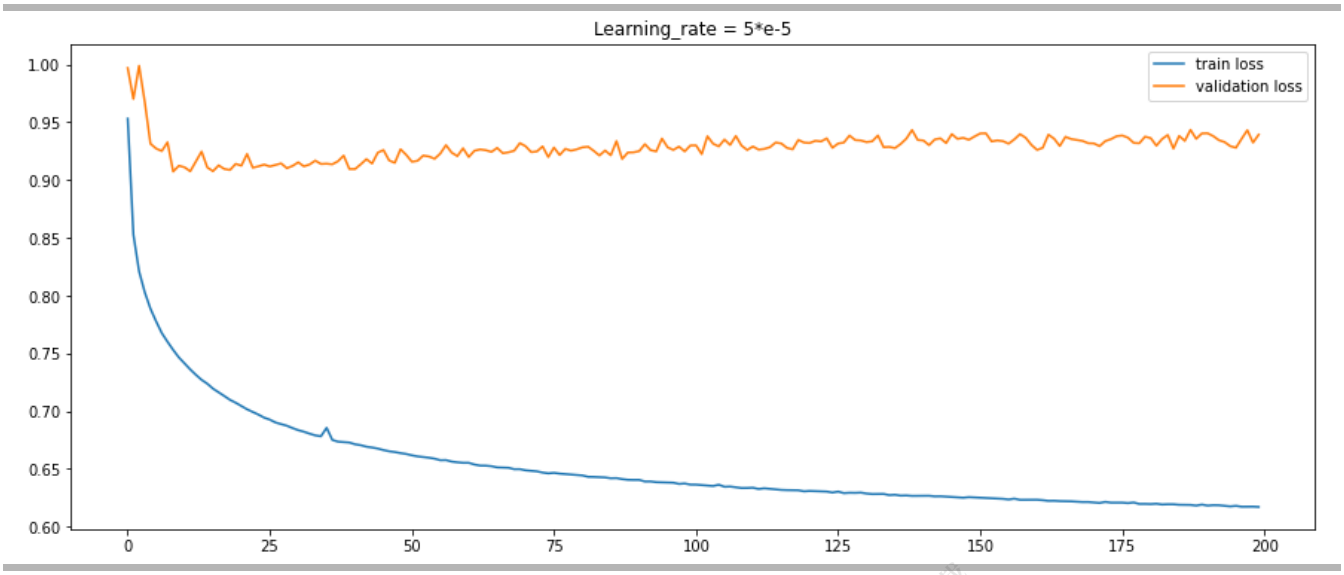


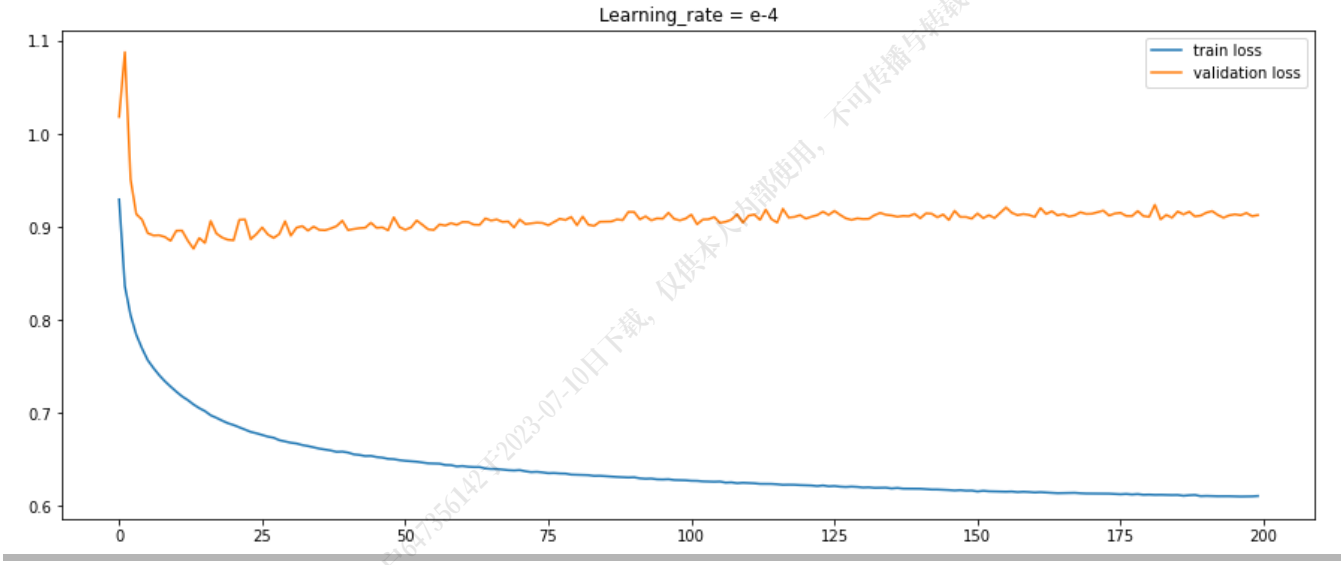
资料来源：吴恩达深度学习

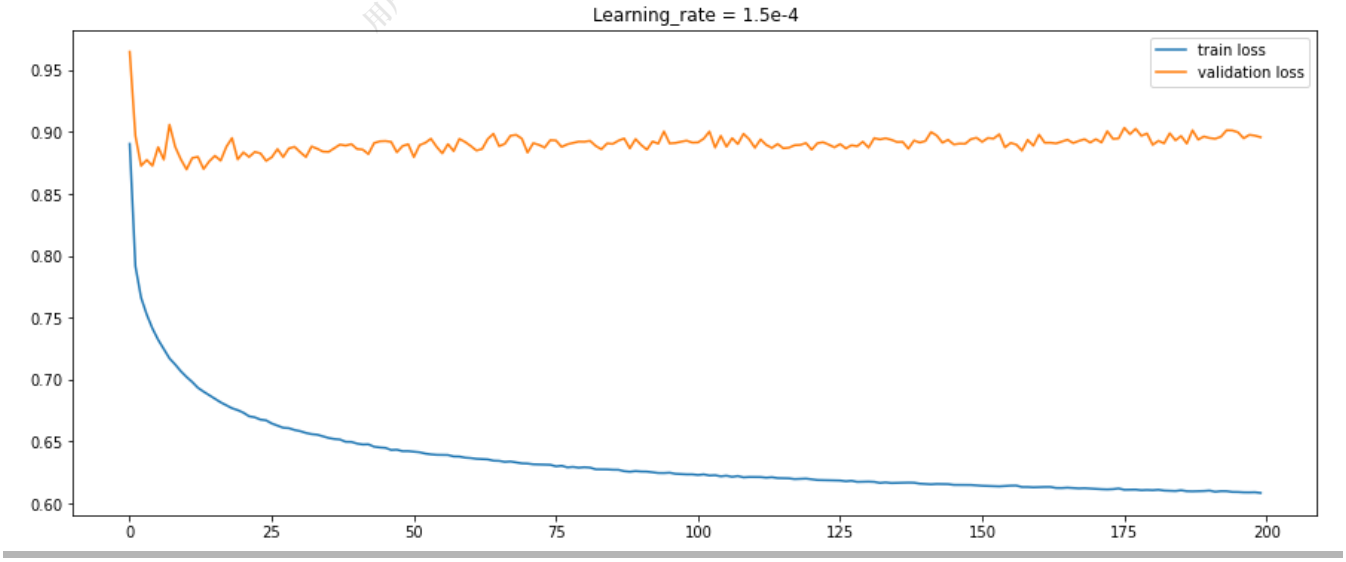
(2) 模型表现

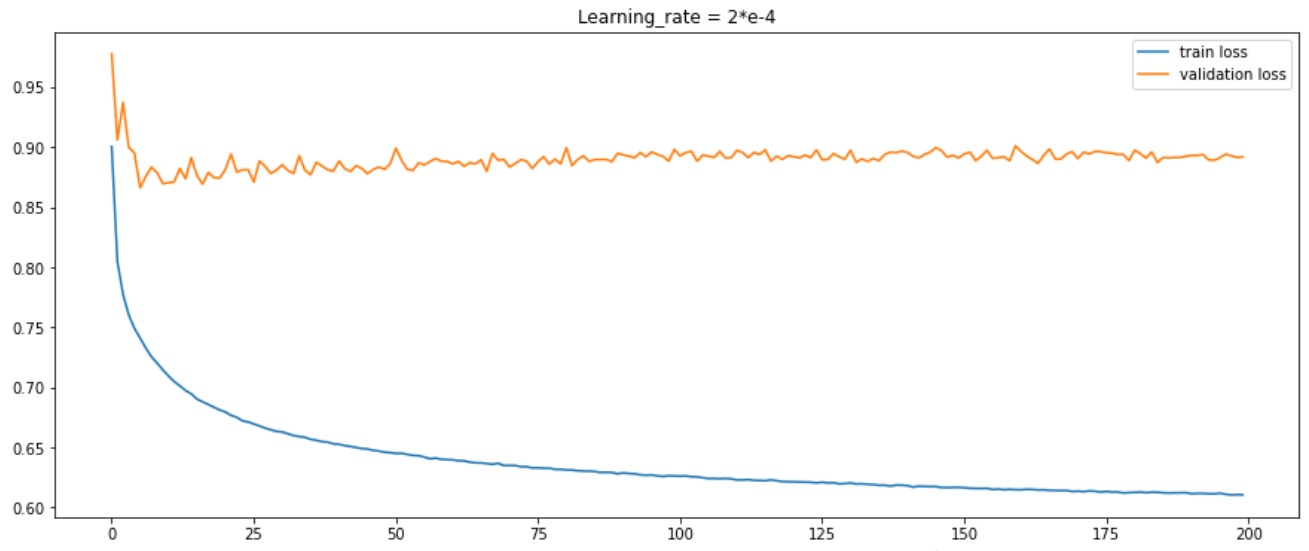
图表 10：CNN+LSTM 模型表现汇总



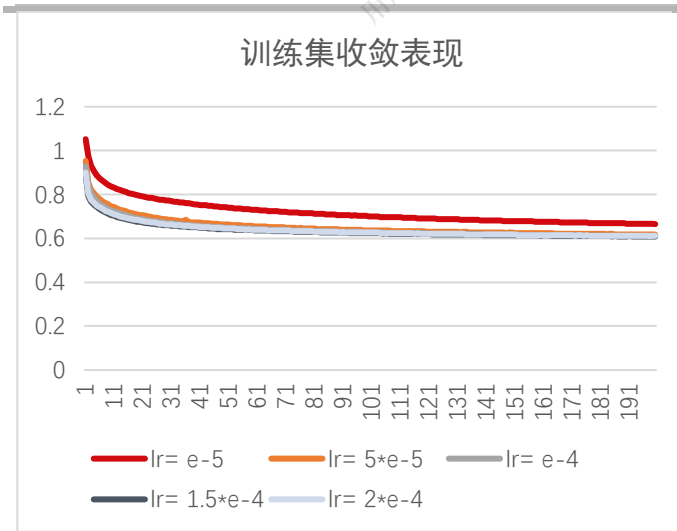


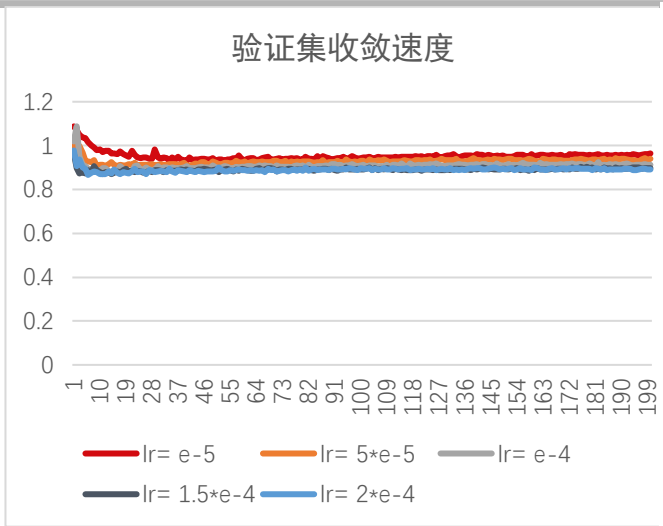






学习率	最优模型表现(CNN_LSTM)
e-5	68.77%
5*e-5	70.04%
e-4	74.39%
1.5*e-4	**75.70%**
2*e-4	75.11%





资料来源：中信期货研究所

3） 模型分析

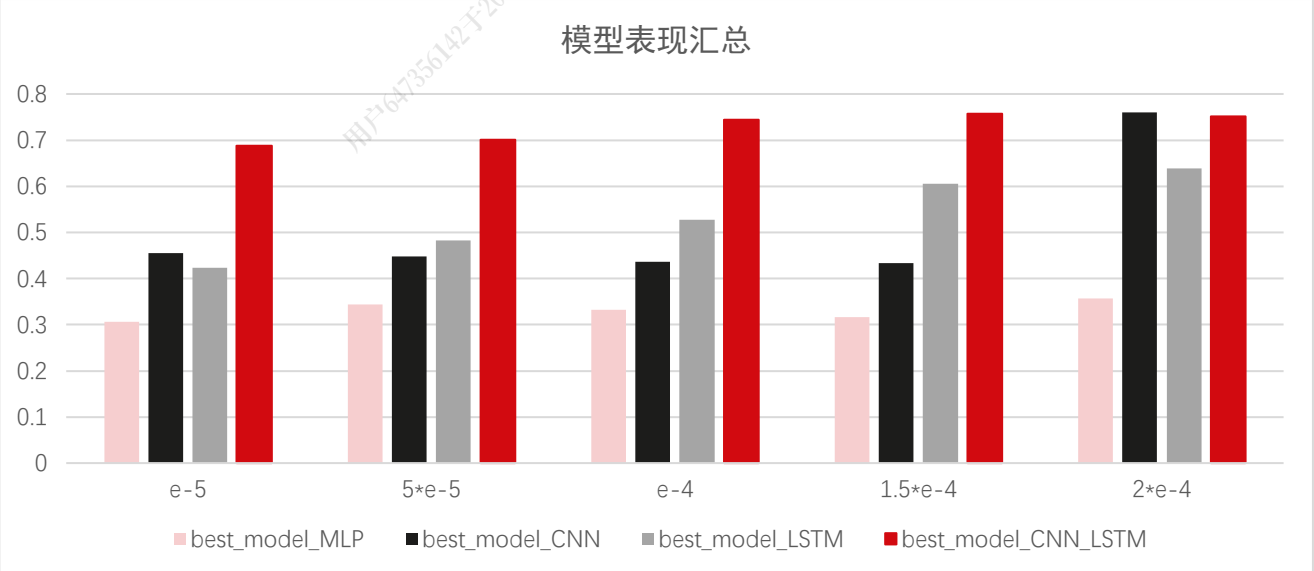
经过卷积神经网络嵌套循环神经网络之后，模型相较有了两方面的进步，1:相比单独的多层感知机、卷积神经网络以及 LSTM，嵌套模型的准确率有较大幅度的提升且提升过程都相对稳定。进一步来说，在不同学习率下的最优模型表现的平均值、最大值也分别有 30%的增长。这样的模型表现侧面表示订单簿空间信息及序列信息都会对未来价格预测起到正向作用，且序列信息的边际贡献率高于空间信息。两种信息同时提取亦会产生叠加效果，进而提升模型准确率。2:在经过多轮训练后，模型依旧在验证集上有稳定的输出(相较单独模型后期的过拟合表现明显)。

三、模型表现总结及后期优化方向

（1） 模型对比总结

经过对基础多层感知机模型的逐步优化可以看出在限价订单簿上单独对空间和序列信息抽取都相较于单纯的维度变换模型在预测力上都有明显的提升，但是在预测准确率的绝对值上都未达到理想效果，这样的准确率在高频领域存在绝对劣势，经过选取相同的五个学习率之后，并在空间和时间序列信息的综合提取后，准确率相较于单独且独立的空间、时间的信息提取都有较大且明显的提升。

图表 11：四类模型表现汇总



资料来源：中信期货研究所

（2） 后期研究方向及适用性分析

现阶段报告在特征提取以及神经网络模型的构造上仍然存在明显缺陷。后期

报告将着重从一下几个方面进行细致研究:

① 现阶段神经网络仅仅着眼于未来的单步预测，鉴于高频交易快进快出的特点，下阶段将继续对多步预测进行研究。

② 下阶段研究报告将加大全新算法的引进，相对仅仅引入时间以及空间信息的算法、后期一方面在数据分析上加大订单量及相关信息的引入，另一方面将进一步引入 Attention 机制以及基于订单簿数据的的强化学习算法，期待在预测力上有更优质表现。

③ 现有的深度学习模型训练成本相比其他模型的训练有更高的成本， 想要进行更高泛化预测能力需要更高的成本， 后期报告将在有条件的基础上进行部分模块的迁移学习。

④ 高频择时模型的准确率直接影响数据以及后期交易信号的输出， 后期报告将着重分析交易信号的产生以及下单延迟的优化。

⑤ 目前国内市场中逐渐增多，部分产品并不具备大量交易数据样本数据，后期根据牛津大学量化研究实验室对小样本类标的产品的处理方式对模型进行改善。(论文题目 Transfer Ranking in Finance: Applications to Cross-Sectional Momentum with Data Scarcity)。

适用性上，后期进一步优化的模型训练方式可以在交易信号生成以及标的风险管理预警上起到支撑作用，具体应用需要深入研究。

免责声明

除非另有说明，中信期货有限公司拥有本报告的版权和/或其他相关知识产权。未经中信期货有限公司事先书面许可，任何单位或个人不得以任何方式复制、转载、引用、刊登、发表、发行、修改、翻译此报告的全部或部分材料、内容。除非另有说明， 本报告中使用的所有商标、服务标记及标记均为中信期货有限公司所有或经合法授权被许可使用的商标、服务标记及标记。未经中信期货有限公司或商标所有权人的书面许可，任何单位或个人不得使用该商标、服务标记及标记。

如果在任何国家或地区管辖范围内，本报告内容或其适用与任何政府机构、监管机构、自律组织或者清算机构的法律、规则或规定内容相抵触，或者中信期货有限公司未被授权在当地提供这种信息或服务，那么本报告的内容并不意图提供给这些地区的个人或组织，任何个人或组织也不得在当地查看或使用本报告。本报告所载的内容并非适用于所有国家或地区或者适用于所有人。

此报告所载的全部内容仅作参考之用。此报告的内容不构成对任何人的投资建议，且中信期货有限公司不会因接收人收到此报告而视其为客户。

尽管本报告中所包含的信息是我们于发布之时从我们认为可靠的渠道获得，但中信期货有限公司对于本报告所载的信息、观点以及数据的准确性、可靠性、时效性以及完整性不作任何明确或隐含的保证。因此任何人不得对本报告所载的信息、观点以及数据的准确性、可靠性、时效性及完整性产生任何依赖，且中信期货有限公司不对因使用此报告及所载材料而造成的损失承担任 何责任。本报告不应取代个人的独立判断。本报告仅反映编写人的不同设想、见解及分析方法。本报告所载的观点并不代表中信期货有限公司或任何其附属或联营公司的立场。

此报告中所指的投资及服务可能不适合阁下。我们建议阁下如有任何疑问应咨询独立投资顾问。此报告不构成任何投资、法律、会计或税务建议，且不担保任何投资及策略适合阁下。此报告并不构成中信期货有限公司给予阁下的任何私人咨询建议。

深圳总部

地址：深圳市福田区中心三路 8 日卓越时代广场（二期）北座 13 层 1301-1305、14 层

邮编：518048

电话：400-990-8826

传真：(0755)83241191

网址：http://www.citicsf.com