证券研究报告—深度报告

金融工程

数量化投资

金融工程专题研究

2017 年 05 月 24 日

专题报告

递归神经网络 RNN—长短期记忆细胞（LSTM）的多因子预测

- **递归神经网络 RNN**

RNN 不同于传统神经网络的感知机的最大特征就是跟时间挂上钩，即包含了一个循环的网络，就是下一时间的结果不仅受下一时间的输入的影响，也受上一时间输出的影响，进一步地说就是信息具有持久的影响力。人们在看到新的信息的时候产生的看法或者判断，不仅仅是对当前信息的反应，先前的经验、思想的也是参与进去这次信息的推断的。

- **RNN 之长短期记忆细胞 LSTM**

LSTM 是一种经过精心巧妙设计的 RNN 网络，尽管 LSTM 和原始 RNN 总的来看都会三大层，即输入层、隐含层、输出层。但是 LSTM 和原始 RNN 在隐含层设计上有较大的差异，主要是 LSTM 是在隐含层具备特殊的 cell 结构。

- **多因子建模**

应用于 RNN 网络结构中时，与传统的多因子模型有一定的区别：

T+1 期的收益率仍然是训练的标签(label)，因子对应的是样本的特征(feature)，个股对应的是一个样本，但是，时间维度，在 RNN 中，是一个循环的过程，将过去 T-n 期的因子数据都要纳入 T+1 期收益率的预测之中。

- **训练结果**

在严格区分了训练集、测试集、样本外数据集之后，我们通过训练能够得到较高准确度的收敛结果，并且在样本外数据回测中，得到显著的超额收益。交叉检验的准确度接近 90%，样本外多空收益最近 12 个月的胜率则超过 90%。

相关研究报告:
《多因子系列研究报告之一：风险（Beta）指标静态测试》 ——2013-01-28
《金融工程机器学习专题：Adaboost 算法下得多因子选股》 ——2016-05-16
《金融工程机器学习专题：SVM 算法选股以及 Adaboost 增强》 ——2016-05-31
《金融工程机器学习专题：基于 KMeans 聚类的多因子特征检验》 ——2016-11-27

证券分析师：黄志文
电话：0755-82133928
E-MAIL：huangzw@guosen.com.cn
证券投资咨询执业资格证书编码：S0980510120059

联系人：陈镜竹
电话：0755-82130833-701336
E-MAIL：chenjz@guosen.com.cn

内容目录

图表目录

深度神经网络与投资

机器学习、深度学习是近些年在生活中出现频率最高的词语之一，在投资领域，也不断的有人尝试利用机器学习的逻辑，对投资决策作出建议。我们在前期的报告中，也曾经尝试了传统的分类器算法支持向量机（SVM）、增强算法 Adaboost 等机器学习的算法用在多因子选股之中，回测效果实际上是较为理想的。但是，对于投资领域而言，机器学习通常被质疑的问题在于，算法过于复杂，训练结果与投资逻辑关联较小，甚至被称为“黑箱”。

但是，“黑箱”并不应该称为机器学习进入投资领域的障碍，以多因子为例，我们不仅在回测检验中，发现了分类器算法有较好的适用性，同时，**从理论上来讲，传统多因子模型运用的核心算法——回归，与分类器算法都是机器学习理论中“有监督学习”的一个重要组成部分**，那么，利用分类器算法，以及其他学习算法，都应该是值得多数投资策略去尝试与探索的。

与深度学习而言，通过对深度神经网络的研究与实践，我们发现深度神经网络作为一个学习算法，它的原理与逻辑是十分严谨与清晰的，同时，通过对神经网络的理解，我们发现投资领域的数据结构多数都能够适应神经网络的要求。

当然，对于最直接的问题：能否利用神经网络，要机器自己识别 K 线图，自己做出判断，本篇报告的内容无法给出肯定的答案，但也不能否定其可能性，回答它需要更为深入、更为复杂的神经网络。本篇报告的目的是利用深度神经网络中的 RNN 递归神经网络的一些基本细胞结果，对多因子模型进行尝试，以检验深度神经网络在多因子、投资领域的适用性，使得投资者能够对神经网络有更为实践的理解，并能够在投资领域有所运用。

递归神经网络(RNN)之 LSTM

神经网络原理介绍

我们知道传统神经网络是神经元按一定层次结构连接起来的，一个重要特点就是传统感知机的应用。感知机是有两层神经元构成，输入层和输出层。输入层功能是接受外界信号后传递信号给输出层，输出层是 M-P 神经元，亦称“阈值逻辑单元”(threshold logic unit). 由于感知机只有输出层有激活函数处理功能，所以学习能力十分有限，难以解决非线性问题。所以进一步发展为多层功能神经元，这就出现了隐含层(hidden layer). 传统神经网络的学习算法中最成功的莫过于误差逆传播算法(error BackPropagation)，简称 BP 算法。BP 算法是基于梯度下降(gradient descent)的策略，以目标的负梯度方向进行调整，调整步长我们称之为学习效率η，学习效率η控制着 BP 算法的每一轮迭代的更新步长，若太大则容易产生震荡，若过小，迭代会太慢。有时为了更精细化调节，向后反馈时输出层到隐含层或者隐含层到更低一层隐含层的学习效率可以使用不同的值。

BP 具体操作如下：先将原始训练个例放入输入层，然后逐层将信号往前传播，这一过程称为前馈，最终到达输出层并产生结果；然后计算输出层的误差，计算输出层梯度项，再将误差逆向传播至隐含层，隐含层根据误差和梯度项计算隐含层误差和梯度项，再对连接权和阈值进行调整。上面过程（前馈+后馈）不断重复，知道达到终止条件，一般 BP 算法的目标是最小化训练集上的累积误差。有学者证明【Homik et al.,1989】,只需一个包含足够多的神经元的隐层，多层前馈网络就可以

任意逼近任何复杂度的连续函数，但目前存在的问题在于隐含层的神经元个数设置，实际操作中通常还是使用试错法来确定最优隐含层神经元个数。

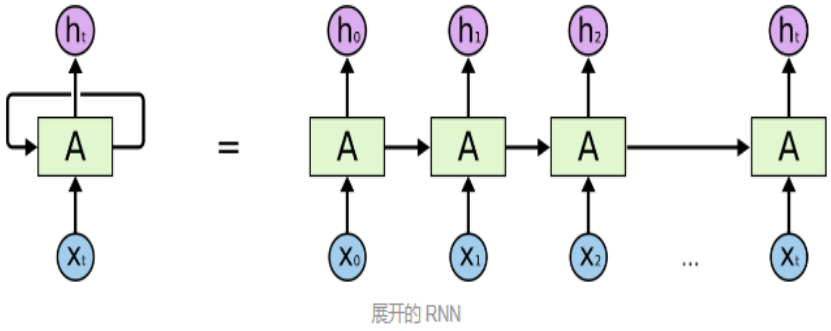
正由上面的学者所述，BP 神经网络强大的表达能力，使得它较容易出现过拟合问题（即在训练集上训练误差不断减小，但是在测试集上误差却在不断增大，目前比较好的解决这个问题有两种常用方法，第一种是“提前停止策略”，即如发现训练误差虽然在减小，但是测试误差开始增大，这时候我们应该提前终止迭代，取此时的权重和阈值作为最终调整结果。第二种方法是正则化，简单地说就是在原目标函数基础上网络复杂度的调整项，即加入连接权和阈值的 2 范数。对于调整后的目标函数中经验误差和网络复杂度的权重选择，一般是采用交叉验证(Cross Validation)来进行估计。

递归神经网络 RNN 简介

RNN 不同于传统神经网络的感知机的最大特征就是跟时间挂上钩，即包含了一个循环的网络，就是下一时间的结果不仅受下一时间的输入的影响，也受上一时间输出的影响，进一步地说就是信息具有持久的影响力。放在实际中也很容易理解，人们在看到新的信息的时候产生的看法或者判断，不仅仅是对当前信息的反应，先前的经验、思想的也是参与进去这次信息的推断的。人类的大脑不是一张白纸，是包含许多先验信息的，即思想的存在性、持久性是显然的。

举个例子，你要对某电影中各个时点发生的事件类型进行分类：温馨、烂漫、暴力等等，如果利用传统神经网络是很难做到这一点的，但是 RNN 因为具备一定的记忆功能，可以较好处理这个问题。

图 1：递归神经网络及其展开形式



资料来源：国信证券经济研究所整理

其中 A 代表隐含层，X 为输入信号，h 为输出信号。

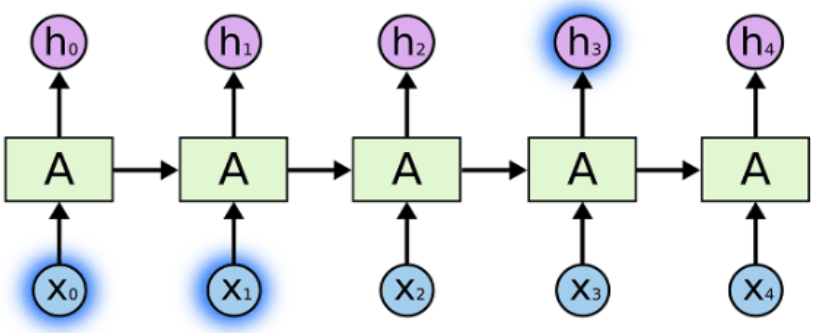
从图中我们也可以看出，RNN 是具备链式结构特征的。递归神经网络因为该循环结构而具有一定的记忆功能，可以被用来解决很多问题，例如：语音识别、语言模型、机器翻译等。但是它并不能很好地处理长时依赖问题，这一问题在（Yoshua Bengio，1994）这篇论文中阐释得很明白。文章指出，最直接的的原因是原始 RNN 模型也是采用 BP 算法进行权重和阈值的调整优化，梯度消失问题依然得不到解决，虽然由于记忆功能的存在使得该问题比传统神经网络有所缓解。但是类似于人类的记忆，人总是会忘事的，即在后面的时间步难以走不回过去了，过去的时间步传递到现在也效果甚微了。所以这使得难以习得远距离的影响。

长期依赖问题（long term dependencies）

RNN 的一个核心思想是，既然网络结构是时间列表特征的，那么可以将以前的信息用到当前的任务中来，例如，在语义推断中，通过前面的话来猜测接下来的话。如果 RNN 真的能够这样做的话，那么它们将会极其有用。但是事实真是如此吗？我们来看下面的例子。

考虑一个语言模型，通过前面的单词来预测接下来的单词。如果我们想预测句子“the birds are flying in the sky”中的最后一个单词，我们不需要考虑上下文信息，就可以得到答案，很明显下一个单词应该是 sky。在这种情况下，当前要预测位置(sky)与相关信息(birds 和 fly)所在位置之间的距离相对较小，RNN 可以被训练来使用这样的信息。

图 2：处理相关信息较近时候的 RNN

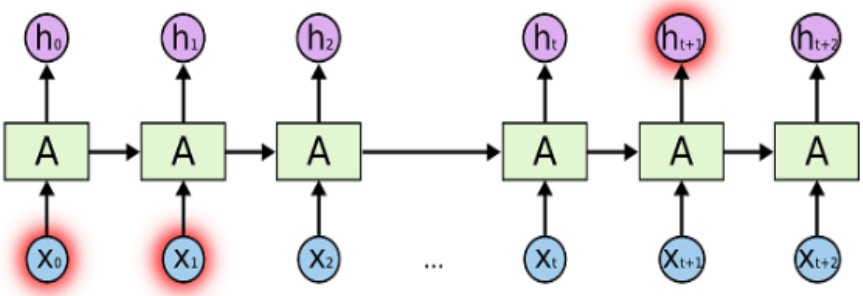


资料来源：国信证券经济研究所整理

但是如果当前位置和相关信息位置距离很远时候，RNN 就会遇到困难了。比如“I grew up in China, when I was ten years old,...,I speak Chinese ”，如果要预测最后一个单词 Chinese，那么我们得搜索较长距离，才能获取到有用的信息 China。但令人失望的是，当需预测信息和相关信息距离较远时，原始 RNN 结构的传输的效率并不让人满意。虽然有学者证明了，我们可以通过精心设计参数来达到预测较远处信息的目的，但是这样无疑是成本很高的，实现起来也很困难，也就失去了实践意义。

从上面分析可以看出，原始 RNN 中存在的长期依赖问题本质上还是梯度消失问题。

图 3：处理相关信息较远时候的 RNN

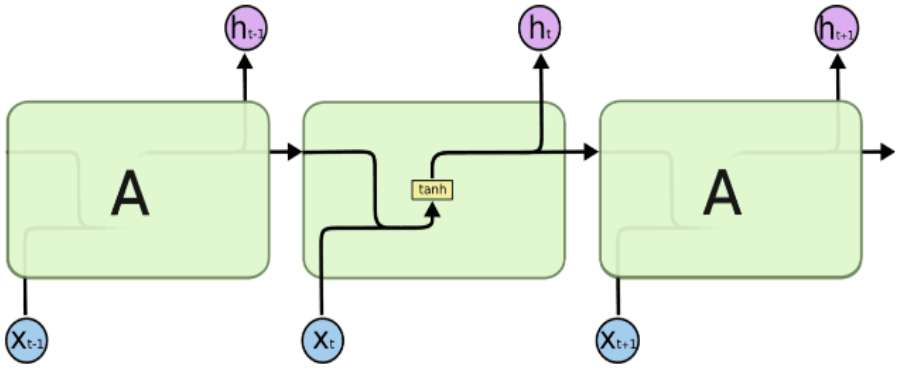


资料来源：国信证券经济研究所整理

长短期记忆网络（LSTM）

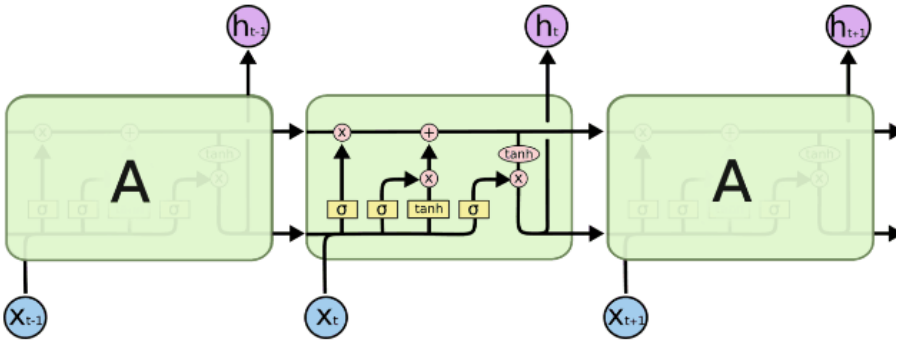
LSTM （long-short term memory），长短期记忆网络，就是为了解决上面的长期依赖问题而生的。LSTM 是一种经过精心巧妙设计的 RNN 网络，尽管 LSTM 和原始 RNN 总的来看都会三大层，即输入层、隐含层、输出层。但是 LSTM 和原始 RNN 在隐含层设计上有较大的差异，主要是 LSTM 是在隐含层具备特殊的 cell 结构。我们用下面两个对比图来进行较好的说明。

图 4：原始 RNN 的隐含层设计



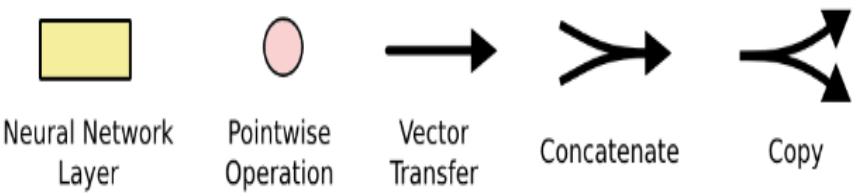
资料来源：国信证券经济研究所整理

图 5：LSTM 的隐含层设计



资料来源：国信证券经济研究所整理

图 6：LSTM 的隐含层设计中图标解释



资料来源：国信证券经济研究所整理

每一条黑线传输着一整个向量，从一个节点的输出到其他节点的输入。粉色的圈代表 pointwise 的操作，诸如向量的和，积等运算，而黄色的矩阵就是学习

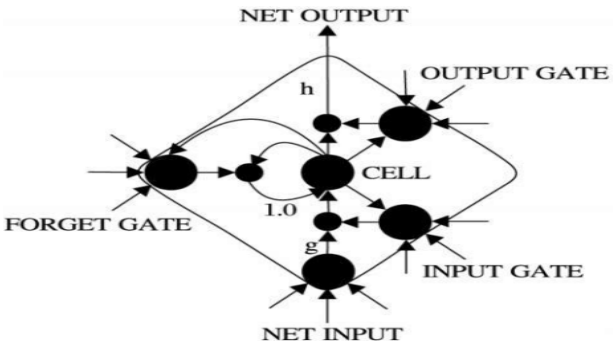
到的神经网络层。合在一起的线表示向量的连接，分开的线表示内容被复制，然后分发到不同的位置。

LSTM 结构设计与思想

LSTM，长短期记忆网络，从上面的图中也可以看出，LSTM 是将一个简单型的激活改成几部分的线性组合的储存单元 cell 去激活。相当于每次都可以控制下一步的输出信息，如是否要包含前面的信息，包含多少的问题等。类似于进行下一步操作前，根据情况提醒你需要注意的信息。好记性不如烂笔头，就是这个道理。

每个存储单元由三大构件组成，输入门，输出门和跨越时间干扰的传入自身的内部状态。

图 7：LSTM 的单元结构



资料来源：国信证券经济研究所整理

输入门（input gate）：控制当前输入X_t和前一步输出h_{t-1},他们能进入新的 cell 单元的信息量。

忘记门（forget gate）：为了更有效传输，需要对信息进行过滤，决定哪些信息可以遗忘。

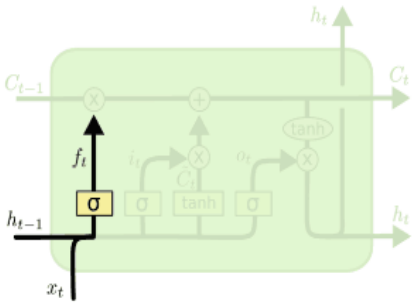
输出门：cell 的新状态下信息更新。

LSTM 详细实现步骤图解

为了更好地说明，我们下面在进行每一步图解时候，都结合语义识别功能进行说明，这样有更直观的认识。

第一步，先由忘记门决定丢弃哪些信息。

图 8：LSTM 的单元结构之忘记门



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

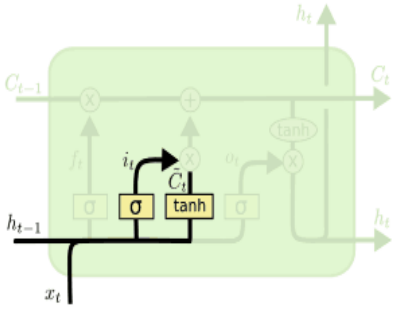
资料来源：国信证券经济研究所整理

即结合当前输入和前一步输出，经激活函数，得到一个概率变量，再与原 cell 结构C_{t-1}做运算得到遗忘后的信息。比如 f=0 时，表示C_{t-1}的所有信息都会被遗忘，f=1 时表示C_{t-1}的信息都会被保存。

让我们回头看看语义预测的例子中来基于已经看到的词去预测下一个词。在这个问题中，细胞状态可能包含当前主语的性别，因此正确的代词可以被选择出来。当我们看到新的主语时要想匹配对应的代词，则我们希望忘记旧的主语和代词。

第二步，由输入层决定什么样的信息会被存储到细胞中。

图 9：LSTM 的单元结构之输入门



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] \; + \; b_C\right)$$

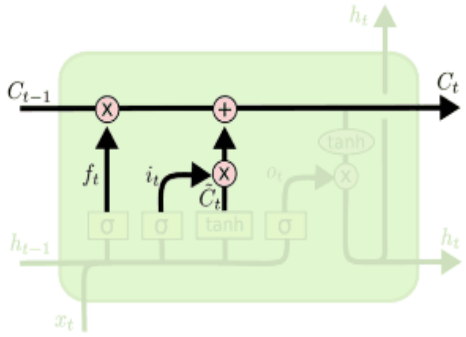
资料来源：国信证券经济研究所整理

这一步这里包含两个部分。第一，sigmoid 层决定什么值我们将要更新。然后，一个 tanh 层创建一个新的候选值向量$\tilde{C}_t$，会被加入到状态中。

在我们语义预测的例子中，我们希望增加新的主语的性别或者别的信息添加到细胞状态中，来替代旧的主语并完善新的主语的信息。

下一步，我们会讲这两个信息来产生对状态的更新。

图 10：LSTM 的单元结构之 cell 更新



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

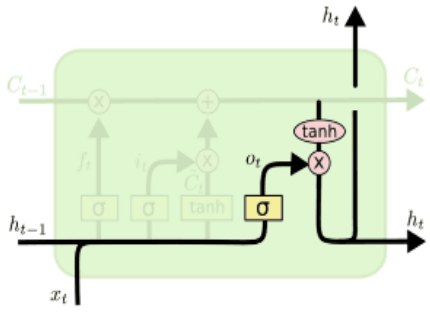
资料来源：国信证券经济研究所整理

即我们 cell 的更新是由经忘记门剩下的信息和需要更新的信息的结合，在语义预测中就是，我们忘记了旧的主语，我们在换成新的主语的时候可以由输入层决定需要更新的信息，比如性别、年龄等。这些作为整体保存在新的 cell 中。

再接着，就是输出信息。这个输出将会基于我们的细胞状态，但是也是一个过滤后的版本。首先，我们运行一个 sigmoid 激活函数来确定细胞状态的哪个部分将输出出去。接着，我们把细胞状态通过 tanh 进行处理（得到一个在 -1 到 1 之间的值）并将它和 sigmoid 的输出相乘，最终我们仅仅会输出我们确定输出的那部分。

在语义预测的例子中，当我们看到了一个代词，可能需要输出与一个动词相关的信息。例如，由 sigmoid 决定可能输出是否代词是单数还是负数，这样如果经过 tanh 函数的细胞状态是动词的话，sigmoid 和 tanh 结合我们也就知道了动词需要进行的词形变化。具体如下图所示：

图 11：LSTM 的单元结构之输出



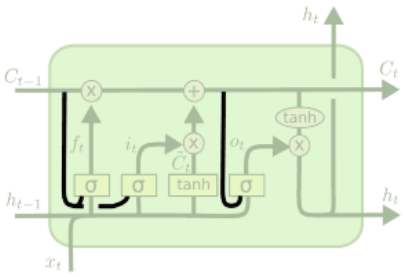
$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

资料来源：国信证券经济研究所整理

LSTM 的发展

上面我们已经把标准的 LSTM 解释清楚了，但是为了满足更复杂的需求，LSTM 出现很多变形。其中最流行的是由 Gers & Schmidhuber (2000) 提出的，增加了 “peephole connection”。是说，我们让“门”也要接受细胞状态的输入。

图 12：LSTM 的变形 1-peephole connection

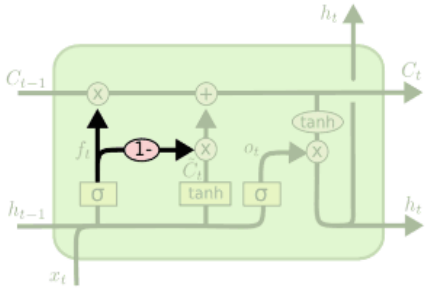


$$f_t = \sigma\left(W_f \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [\boldsymbol{C_t}, h_{t-1}, x_t] \; + \; b_o\right)$$

资料来源：国信证券经济研究所整理

另一个变体是通过使用复合忘记和输入门。不同于之前是分开确定什么忘记和需要添加什么新的信息，这里是一同做出决定。

图 13：LSTM 的变形 2-复合忘记门和输入门

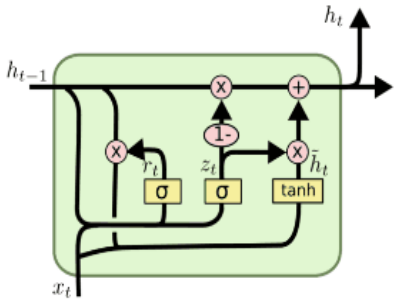


$$C_t = f_t * C_{t-1} + (1 - \boldsymbol{f_t}) * \tilde{C}_t$$

资料来源：国信证券经济研究所整理

还有比较流行的是改动较大的变体是 Gated Recurrent Unit (GRU)，这是由 Cho, et al. (2014) 提出。它将忘记门和输入门合成了一个单一的更新门。同样还混合了细胞状态和隐藏状态，和其他一些改动。最终的模型比标准的 LSTM 模型要简单

图 14：LSTM 的变形 3-GRU



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

资料来源：国信证券经济研究所整理

多因子建模

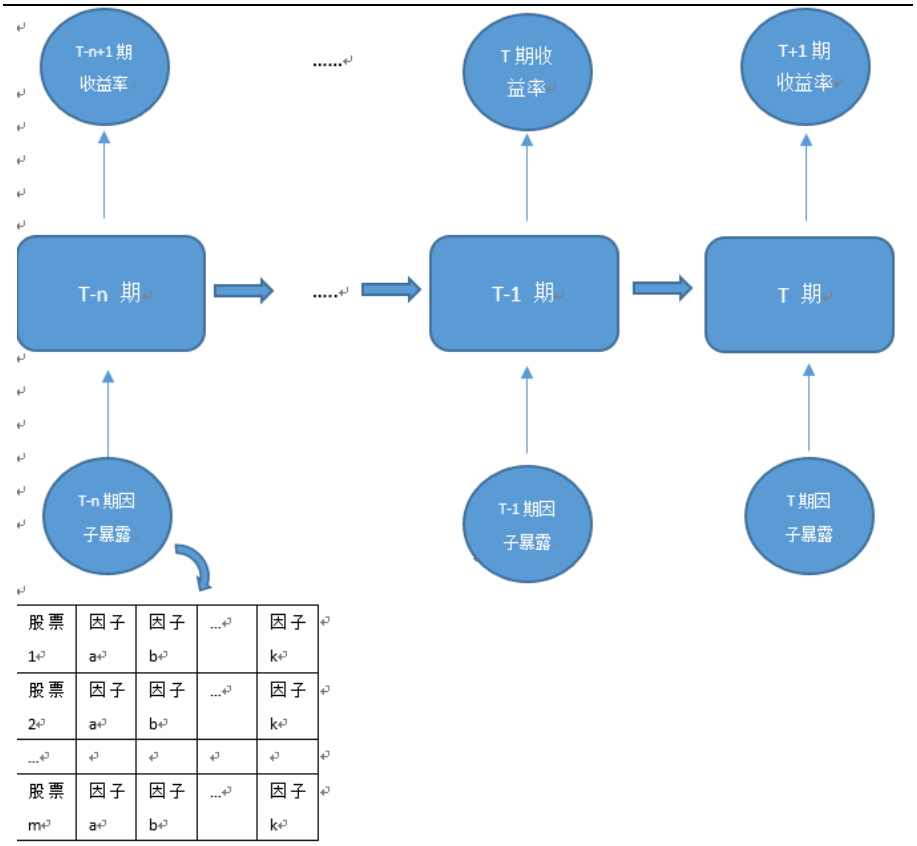
数据结构

多因子模型处理的数据结构是标准的面板数据，包括三个维度：个股、时间、因子，对应的应变量是 T+1 期的收益率。

应用于 RNN 网络结构中时，与传统的多因子模型有一定的区别：

T+1 期的收益率仍然是训练的标签(label)，因子对应的是样本的特征(feature)，个股对应的是一个样本，但是，时间维度，在 RNN 中，是一个循环的过程，将过去 T-n 期的因子数据都要纳入 T+1 期收益率的预测之中：

图 15：多因子的 RNN 数据结构



资料来源：国信证券经济研究所整理

我们先设定具体的参数，再进一步理解 RNN 在多因子训练中的具体过程。

参数设定

回测时间：2007 年 5 月 1 日-2016 年 4 月 30 日，该时间区间下月度数据训练样本数超过 18w（每一个股票每一个月底代表一个样本）

策略时间：2016 年 5 月 1 日-2017 年 4 月 30 日

RNN 时间长度（steps）：24 个月，即每一个训练样本包含过去 24 个月的因子数据，依次从第一个月输入神经网络，并将返回值与下一个月因子同时循环输入神经网络，以此类推，直到得到第 24 个月的预测值。

因子数：由于放入神经网络中训练，我们在期初并不评价因子的有效性，同时也不对因子进行合并，全部输入模型之中。（剔除部分相关性过高，且属于同类的因子，该过程可以降低模型训练过拟合的可能）最终入选 48 个小因子，属于 10 类常见的风格因子。（详见后文统计与国信多因子系列报告）

分类数：为了验证预测的准确性，同时排除样本中的部分噪声，我们将样本的收益率类型分为三类：上涨（月收益率大于 3%）、下跌（月收益率小于-3%）、中性（月收益率处于-3%与 3%之间）

batch size：1000，该参数属于 RNN 神经网络的系统参数，是 BP 算法中用来计算梯度的参数，即每次的训练，随机抽取 18w 训练样本中的 1000 个样本作为训练样本。

隐层神经元个数：400，2 层，该参数同样属于 RNN 神经网络的系统参数，是输入样本与隐层细胞连接的“神经”个数，受限于电脑的性能，只能设定为三位数，隐层个数也仅为 2 层。

学习速率：0001，RNN 神经网络的系统参数，是模型训练时梯度下降的速度，过高容易导致梯度消失，过低则训练过慢。

交叉检验比例：10%，为了防止模型过拟合，选择 18w 样本中的 90%作为训练集，用以训练模型参数，而剩余 10%不参与训练，只作为测试集进行检验，如果训练集准确率与测试集准确率同时上升，则说明模型过拟合的可能较小。

需要说明的是，通过训练，我们发现，**最后的 4 个 RNN 系统参数，在本次报告中并不敏感，我们只设定为常见的参数值，就可以得到较为理想的准确率**。

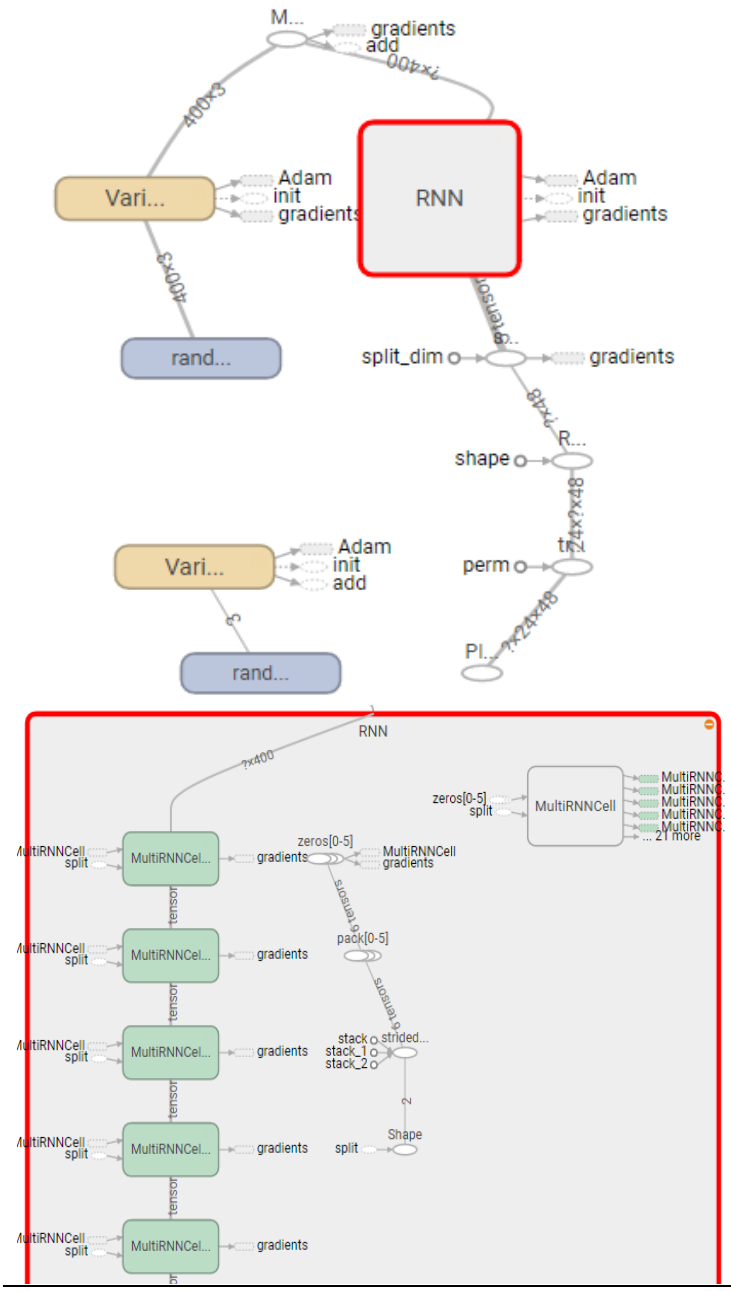
训练结果

数据预处理：仿照多因子的流程，对截面因子进行去极值、标准化的处理，同时，为了剔除行业的效果，截面单因子对行业矩阵回归，取残差作为最终输入的因子数据。

样本内训练

经过 100 次迭代，已经能够观察到训练收敛的结果。

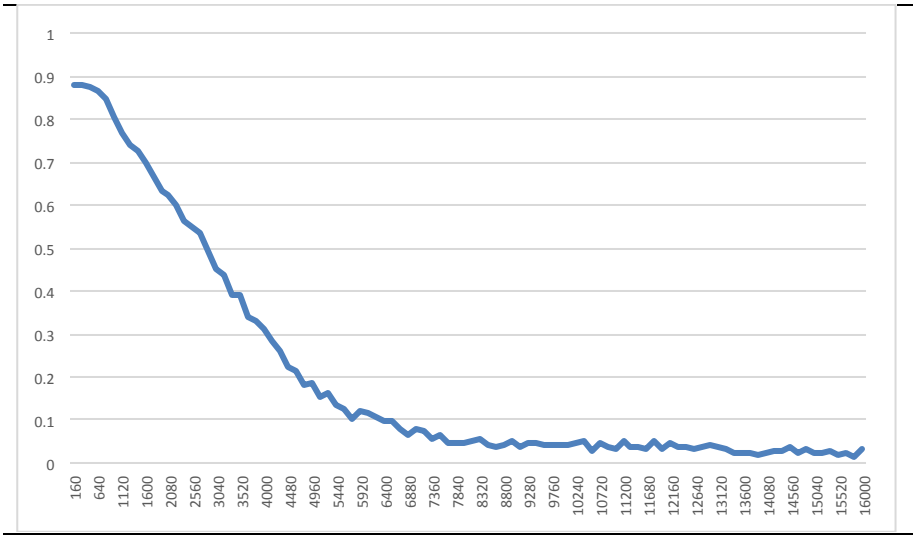
图 16： RNN 可视化结构



资料来源：国信证券经济研究所整理

基于上图的基本的两层 RNN 网络结构，得到的损失率如下图：

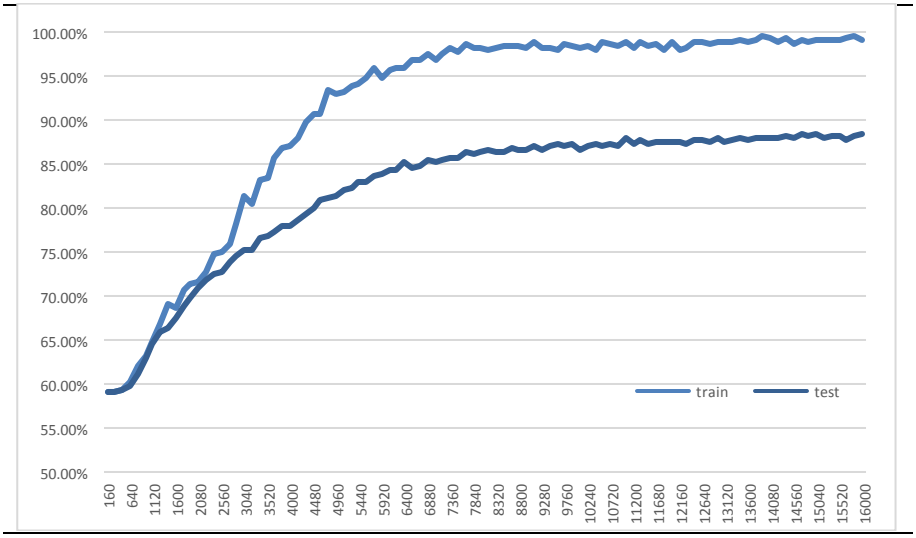
图 17： Basic_LSTM 损失率



资料来源：wind，国信证券经济研究所整理

转换为模型的 3 类收益率预测值与真实值的对比准确率：

图 18： Basic_LSTM 交叉检验准确率



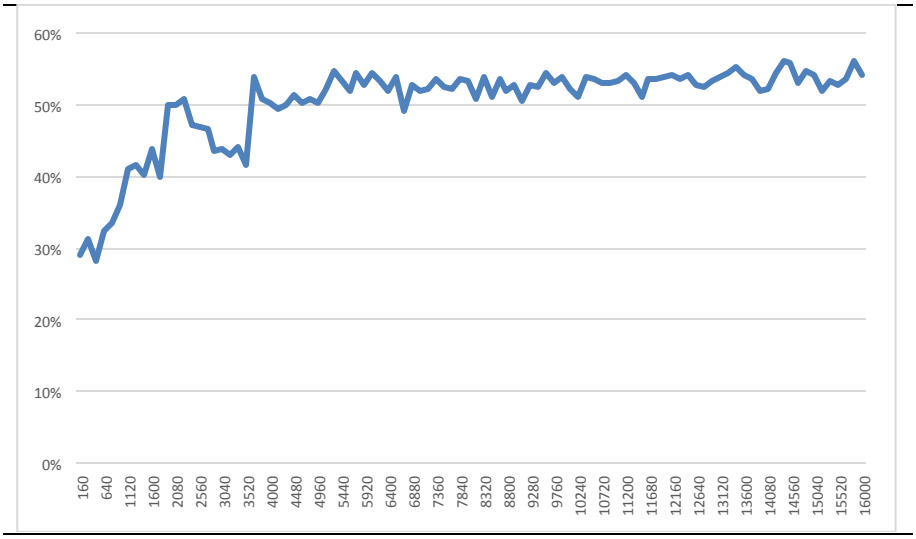
资料来源：wind，国信证券经济研究所整理

从曲线中可以看到，检验集的准确率最终收敛于 85%-90%之间，这个准确率水平在机器学习的大多数模型中并不足够高，但考虑到我们使用的是基本的 RNN 结构，同时是存在市场博弈的股票市场，我们认为这一收敛水平能够反映出 LSTM 神经网络对多因子数据进行了有效的训练与特征抓取。

样本外检验

通过训练的最终结果，我们将样本外数据 2016-2017 输入，得到模型对未来 12 个月的股票收益率的估计。其准确率的走势如下图：

图 19： Basic_LSTM 样本外选股准确率

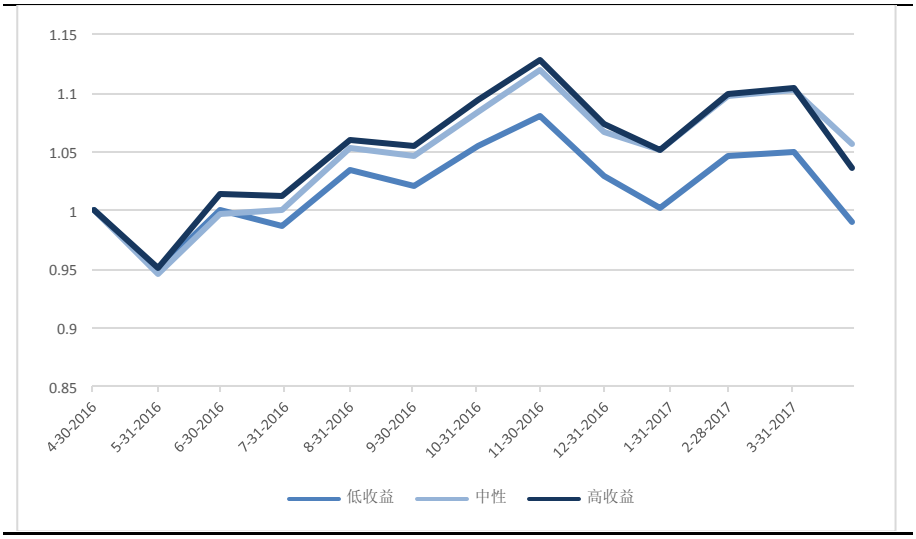


资料来源：wind，国信证券经济研究所整理

样本外的准确率最终收敛水平仅高于 50%，但是需要区分这一水平所能够反映的真实预测程度。为了直观的检验 LSTM 模型样本外的选股效果，我们选择模型给出的每个月个股的预测结果作为选股标准。

每月末，将样本外数据输入模型，并根据模型输出的对个股收益的三类（-3%、3%）预测，将全 A 股股票分为三个组合——高收益预测、低收益预测、中性预测。

图 20： 全 A 股预测组合净值



资料来源：wind，国信证券经济研究所整理

可以看到，模型在最近一年，对高、低收益的预测胜率较高，但对于居中的中性组合预测效果较差。

图 21: 全 A 股多空组合累计净值

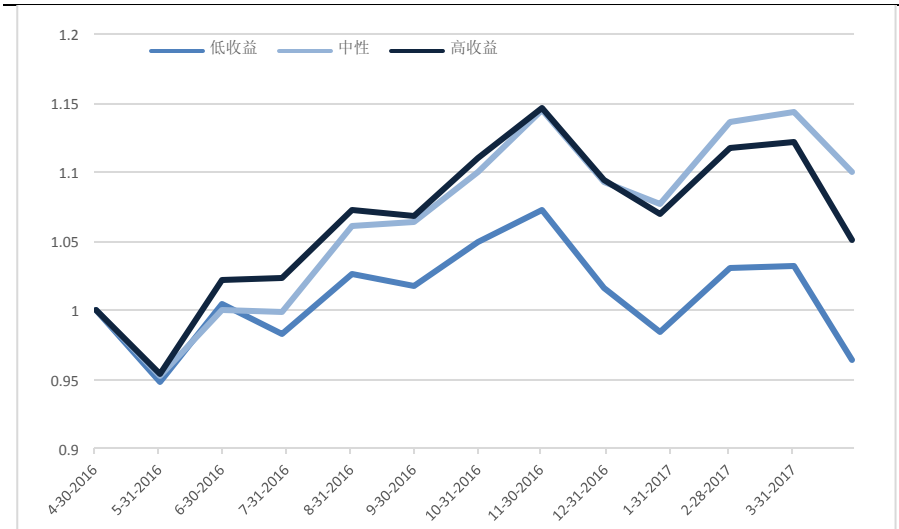
资料来源：wind，国信证券经济研究所整理

多空超额收益在最近 12 个月的胜率为 75%。从多空累计净值上看，多空超额收益最近 12 个月在 4.5%。

为了进一步验证模型对于股票预测的准确性，我们把选股的标准从模型输出的预测变为模型最终预测前的**激活值**。由于我们将预测目标分为了三类（高、中、低），神经网络会选择激活值最大的类别，作为预测类别。因此，激活值实际上反映了模型对个股未来收益的预测概率。

基于此，我们重新构建三类股票组合，每一期，选择激活值最大的 30%的股票最为对应组合：

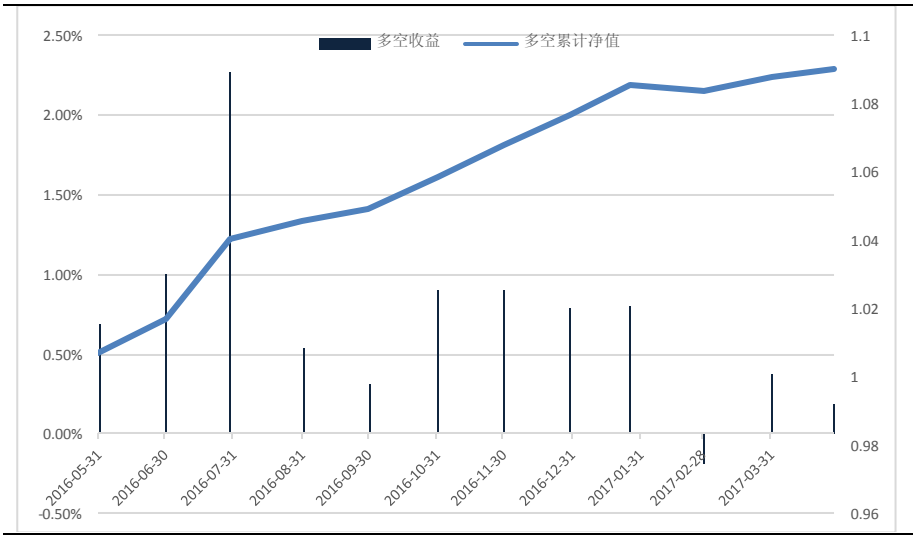
图 22: 30%多空组合净值



资料来源：wind，国信证券经济研究所整理

可以发现，模型对于中性收益的预测效果仍然没有改进，但是多空收益的预测效果比全 A 股更加准确。

图 23： 30%多空组合累计净值



资料来源： wind，国信证券经济研究所整理

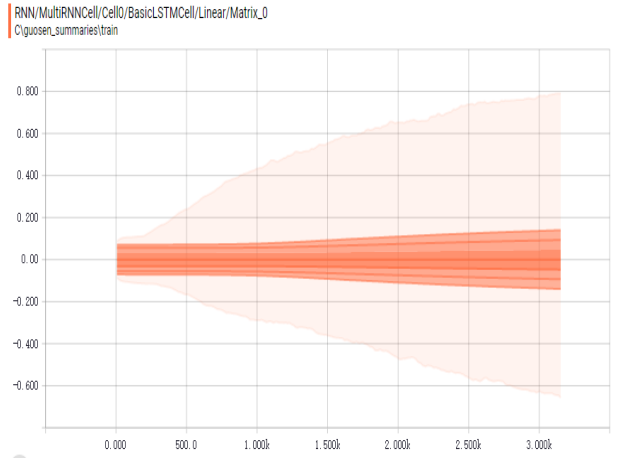
多空组合的超额收益超过 9%，而最近 12 个月的月度胜率超过 90%。通过样本外数据的回测，我们发现，通过 LSTM 的 RNN 网络学习，对股票的收益率预测实际上是较为准确的，同时，模型对于不同收益类型的预测概率能够更进一步的反映出股票上涨与下跌的概率大小。

结果分析

回顾 RNN 神经网络的结构，在基本的 LSTM 结构中，每一期的输入样本，其与隐层、隐层与输出层的连接权重是共享的，也就是说，在我们具体的模型里，每一期 48 个因子的输入，都对应有 400 个权重连接到隐层 400 个神经元上，每一期的循环都会对这 400 个权重进行更新。

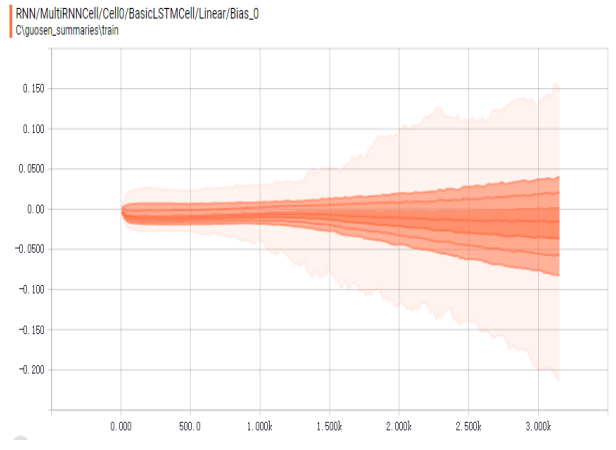
下图是输入层的权重分布的更新过程：

图 1： 输入层权重 w 分布变化



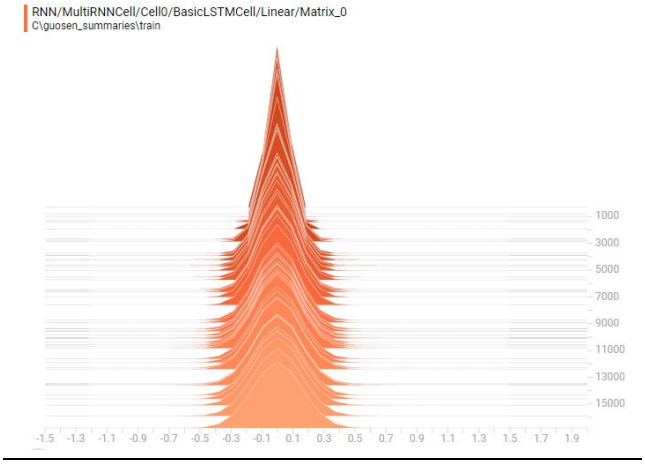
数据来源： 国信证券经济研究所整理

图 2： 输入层 bias 分布变化



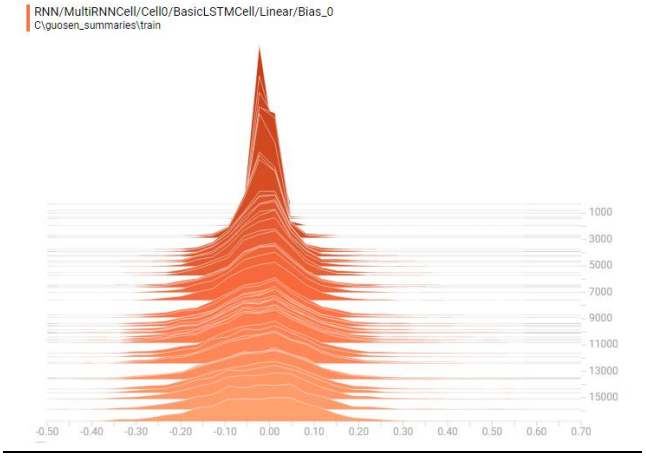
数据来源： 国信证券经济研究所整理

图 1：输入层权重 w 分布



数据来源：国信证券经济研究所整理

图 2：输入层 bias 分布

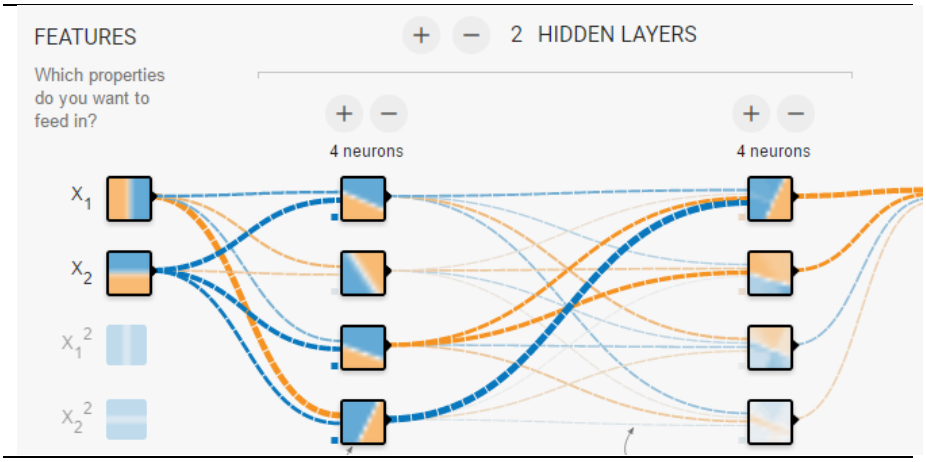


数据来源：国信证券经济研究所整理

与我们观察到的模型训练的损失率收敛特征一致，神经网络在较早的迭代次数时已经逐渐收敛，参数权重趋于稳定。

既然知道了输入层的权重，我们考虑观察训练结束时，模型得到的因子与隐层的连接权重，作为因子在系统中的贡献程度。

图 24：参数权重变化示意图



资料来源：google，国信证券经济研究所整理

从示意图能够直观的理解因子的权重。X1、X2 代表实际模型中的因子，连接输入样本 X1、X2 与隐层神经元的曲线即为训练的权重，随着样本迭代，权重朝着最优解的方向变化，权重越大，则示意图中的曲线越粗。

由于我们观察到本报告中的参数分布较为稳定，因此我们认为，可以大致将输入层因子的权重总和作为该因子在模型中的贡献度。虽然因子真实的贡献度也会受到隐层权重的影响，但输入层的最终结果仍然具有一定的参考意义。

图 25: 输入层因子权重绝对值之和

ETOP	44.01	T_CFOG	55.90	EBITDAvsEV	47.60
CETOP	50.97	C_ROEG	42.88	HILO	67.36
EXTE	48.76	C_ROAG	41.71	BTSG	48.45
VFLO	54.24	BLEV	49.77	LPRI	48.13
VERN	48.74	DTOA	44.89	CMRA	50.20
AGRO	50.72	DTOAS	45.87	VOLBT	48.05
EGRO	48.23	STO_1M	47.94	BETA	41.72
SGRO	46.55	STO_6M	41.44	SIGMA	50.51
DELE	50.20	RSTR_1M	68.29	HALPHA	54.58
S_SalseG	45.89	RSTR_3M	54.78	S_GPM	44.45
C_SalseG	43.07	RSTR_6M	44.95	S_NPM	43.72
T_SalseG	46.03	RSTR_12M	46.41	S_CTP	52.39
S_ProfitG	50.66	LNCAP	50.11	C_CTP	44.30
T_ProfitG	51.99	BTOP	43.73	T_CTP	48.04
S_CFOG	55.52	STOP	42.20	S_ROE	43.66
C_CFOG	51.51	CFTOP	50.66	S_ROA	43.77

资料来源：国信证券经济研究所整理

结论

多因子模型的发展趋于成熟，因子的 alpha 收益出现了下降的趋势。如果维持多因子模型的收益是量化领域的一个核心问题。根据我们以往的报告，我们认为扩展的方向包括新的因子挖掘、股票池的区分，以及非线性因子特征的挖掘。而机器学习，正式非线性问题的一个有效解决途径。**具体到本篇报告涉及的深度神经网络 RNN 来说，即是通过时间维度的扩展，以及空间深度的扩展，将目前的因子空间，扩散到更高维度的空间中去，并在其中找寻有效的路径，实现对因子模型的预测**。

在严格区分了训练集、测试集、样本外数据集之后，我们通过训练能够得到较高准确度的收敛结果，并且在样本外数据回测中，得到显著的超额收益。交叉检验的准确度接近 90%，样本外多空收益最近 12 个月的胜率则超过 90%。

这些结果的意外之处在于，利用基本的 LSTM 结构，能够在参数未优化之前得到如此高的准确率与显著水平，对于模型的进一步改进和优化令人有所期待。同时，这些结果的意料之中在于，当我们不再将机器学习、神经网络当做复杂的“黑箱”，其强大的数据处理能力必将在投资领域展露出来，也同样令人期待。

国信证券投资评级

类别	级别	定义
股票投资评级	买入	预计 6 个月内，股价表现优于市场指数 20%以上
	增持	预计 6 个月内，股价表现优于市场指数 10%-20%之间
	中性	预计 6 个月内，股价表现介于市场指数 ±10%之间
	卖出	预计 6 个月内，股价表现弱于市场指数 10%以上
行业投资评级	超配	预计 6 个月内，行业指数表现优于市场指数 10%以上
	中性	预计 6 个月内，行业指数表现介于市场指数 ±10%之间
	低配	预计 6 个月内，行业指数表现弱于市场指数 10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道，分析逻辑基于本人的职业理解，通过合理判断并得出结论，力求客观、公正，结论不受任何第三方的授意、影响，特此声明。

风险提示

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询业务是指取得监管部门颁发的相关资格的机构及其咨询人员为证券投资者或客户提供证券投资的相关信息、分析、预测或建议，并直接或间接收取服务费用的活动。

证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券机构销售团队

华北区（机构销售一部）	华东区（机构销售二部）	华南区（机构销售三部）	海外销售交易部
李文英 010-88005334 13910793700 liwying@guosen.com.cn	汤静文 021-60875164 13636399097 tangjingwen@guosen.com.cn	邵燕芳 0755-82133148 13480668226 shaoyf@guosen.com.cn	赵冰童 0755-82134282 13693633573 zhaobt@guosen.com.cn
王　玮 13726685252	吴　国 15800476582	赵晓曦 0755-82134356 15999667170 zhaoxxi@guosen.com.cn	梁　佳 0755-25472670 13602596740 liangjia@guosen.com.cn
许　婧 18600319171	梁轶聪 021-60873149 18601679992 liangyc@guosen.com.cn	颜小燕 0755-82133147 13590436977 yanxy@guosen.com.cn	程可欣 886-0975503529(台湾) chengkx@guosen.com.cn
边祎维 13726685252	倪　婧 18616741177	刘紫微 13828854899	夏　雪 18682071096
王艺汀 13726685252	林　若 13726685252	简　洁 13726685252	吴翰文 13726685252
詹　云 13901062999	张南威 13726685252	欧子炜 18150530525	
陈雪庆 18150530525	周　鑫 13726685252		
杨云崧 18150530525	张欣慰 13726685252		
赵海英 010-66025249 13810917275 zhaohy@guosen.com.cn			