目　录

一　博士学位期间拟开展研究课题论证

1.1　拟开展研究的课题名称

时间序列数据驱动下的可解释学习模型研究

1.2　拟开展研究课题的选题意义

在大数据的时代背景下，时间序列数据同样具备规模巨大、模态多样、关联复杂和真伪难辨等大数据的性质，并呈现出传统数据挖掘方法下感知度量难、特征融合难和模式挖掘难等问题。作为一种动态机制不确定的数据，如何建立一种可解释的学习模型来克服大规模时间序列数据非平稳状态下的动态性逼近或分类问题，对于学界和业界都具有重要的理论价值和实际意义。

1.3　拟开展研究课题的国内外研究现状

过去的十几年中，神经网络因其优秀的数据感知和学习能力被成功地应用在了各类场景中[1,2]。近年来，伴随着硬件计算能力的飞速提升，以深度神经网络为代表的学习模型在围棋博弈[3]、谈判推演[4] 和艺术风格迁移[5] 等诸多领域取得了突破。相对于深度学习的工程快速应用，其理论解释则一直滞后。在实际研究中，神经网络的层链结构与激活函数等参数更多的采用多次调试来予以确定，对于如何选取合适的神经网络结构或参数以取得更优的学习或生成效果至今仍是挑战。也因此，学界主流观点认为现在深度学习模型普遍是欠解释或不可解释的，如何建立一种具有可解释性的学习模型亦是NIPS2017 的关注焦点。

为探索神经网络的可解释性，一种可尝试的方法是从小规模的神经网络开始，逐渐增加神经网络的隐藏层单元直到满足预定的终止条件，以此观测神经网络中的神经元作用[6]。众所周知的是，迭代寻找最优隐藏层单元数和权重与偏置值所需要的计算复杂度和开销对数据集的大小是极其敏感的。因此在处理大规模数据集时，这种方法很难得以应用。

面对大规模数据集，随机算法具有其独特的优势[7]。在神经网络计算中，随机算法及其思想已证明了其在建立快速学习模型和算法的同时能够大幅度的减少计算开销[8,9]。同时，基于随机学习算法的神经网络其神经元权重或偏置服从给定概率密度函数下的随

机分布。这种神经元权重或偏置分布确定的神经网络相比纯训练优化神经元权重与偏置的神经网络无疑更具有解释性。

随机学习神经网络一般遵循着两个公认且基本的训练范式，即随机输入神经网络隐藏层单元的权重与偏置值和给定标准来筛选权重。在这种范式下，Igelnik 和 Pao 提出了一种随机向量函数链接神经网络（RVFL，Random Version of the Functional-Link Net）[10]。这种通过从给定范围均分分布中确定参数的神经网络在连续函数上具有良好的普适逼近效果。Tyukin 和 Prokhorov 表明 RVFL 神经网络需要监督机制来更好逼近目标函数[11]。并且，Tyukin 和 Prokhorov 的实验显示 RVFL 网络在随机算法参数设置不合适的情况下会有很高概率无法逼近目标函数。这一现象在随后的研究中被 Gorban 等人以通过数学推导的方式加以确认[12]。Gorban 等人认为 RVFL 神经网络的设计需要看考虑两个至关重要的参数，即隐藏层单元数量与随机参数范围。其中，隐藏层单元的数量与模型准确率直接相关，因此隐藏层单元数量需要足够的多。此外，随机参数的选取也直接影响到模型的逼近效果。在实际应用中，面对不同规模的数据集，如何确定合适的 RVFL 网络隐藏层单元数量和随机参数范围成为了 RVFL 网络需要解决的首要问题。针对此问题，Li 和 Wang 希望通过在给定随机参数分布函数的条件下递增 RVFL 隐藏层单元来确定数据驱动式的 RVFL 网络[13]。Li 和 Wang 发现给定随机参数范围和学习模型收敛速度条件的情况下 RVFL 网络无法保证其普适逼近能力。因此，研究保证普适逼近能力前提下神经网络根据数据自适应调整神经网络结构和隐藏层神经元的随机参数范围是十分重要和必要的。

1.4 考生开展本课题研究的主要思路、基本内容和重要观点

在大数据的背景下，面对非平稳状态下动态机制不确定和高复杂度的大规模时间序列数据，本课题认为建立一种具有普适逼近能力的、基于随机分布统计视角的和可解释可自适应调整的神经网络学习模型能够有效地学习时间序列数据的状态与规律。基于此，本课题具有重要的理论与实际意义。

对于如图 1-1所示的时间序列数据示例，其公式为 $f(x) = 0.2e^{-(10x-4)^2}+0.5e^{-(80x-40)^2}+0.3e^{-(80x-20)^2}$ 。这类时间序列数据在金融市场和故障监测等实际情景中极为常见，其在某一区域内剧烈波动的状态难以被传统的非线性回归方法逼近。神经网络作为多重感知机，可以凭借其不同的基函数与非线性激活函数的组合了理论上拟合出这类不平稳且局部高复杂度的函数。但在实际应用确定神经网络模型去逼近此类函数时，亦有明显的

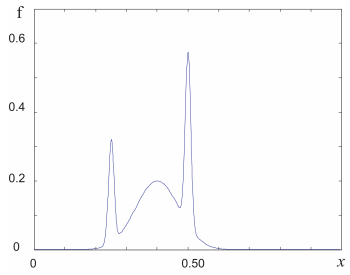

图 1-1　时间数据数据示例

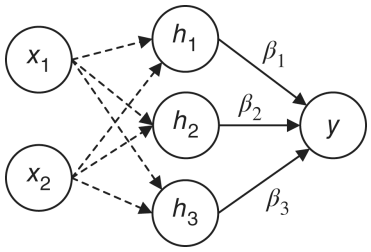

图 1-2　神经网络示例

缺陷和问题。首先确定神经网络模型需要经过多次的实验比较来调整神经网络的层数、各层隐藏层神经元数量和激活函数类型等参数,这一过程非常耗费研究人员或工作人员的精力同时对于不同的时间序列数据需要重复调参来优化神经网络的逼近效果;另一方面,在采用梯度下降算法进行逼近收敛时,无法保证收敛速度;且在此过程中,无法解释神经网络中神经单元作用,即无法得知在网络优化中各层各神经单元对整体网络优化的影响。这些缺陷和问题制约了时间序列数据的逼近研究,基于此,本课题研究一个具有普适能力,神经网络权重分布可知和可根据不同时间序列数据自适应调整网络结构的神经网络学习模型,并给出了理论证明。

对于如图 1-2所示的神经网络示例,X_i 表示输入层的时间序列数据,该数据具有 d 维特征;h_q 表示隐藏层单元,该隐藏层神经元有 m 个,每个神经元可表示为 $g(w^{\mathrm{T}}x+b)$;β 表示隐藏层单元的输出矩阵;Y 表示输出层的时间序列数据。

在进行时间序列数据驱动下的可解释学习模型证明前,首先对模型中的函数条件、范式和内积进行约束和定义。在时间序列数据的范畴下,模型中的函数均为勒贝格测度函数(Lebesgue measurable functions)且均为实值函数。在此,以 $\Gamma:=\{g_1,g_2,g_3...\}$ 表示一组实值函数,span(Γ) 表示由 Γ 张成的函数空间;$L_2(D)$ 表示包含所有勒贝格测度函数 $f=[f_1,f_2,\ldots,f_m]:\mathbb{R}^d\to\mathbb{R}^m$, $D\subset\mathbb{R}^d$ 的函数空间;L_2 范式被定义为:

$$\|f\|:=\left(\sum_{q=1}^{m}\int_D|f_q(x)|^2dx\right)^{1/2}<\infty \tag{1.1}$$

勒贝格测度函数的内积被定义为:

$$\langle f,\theta\rangle:=\sum_{q=1}^{m}\langle f_q,\theta_q\rangle=\sum_{q=1}^{m}\int_D f_q(x)\theta_q(x)dx \tag{1.2}$$

在 $m=1$ 的特殊情况下,对于一个实值函数 $\psi:\mathbb{R}^d\to\mathbb{R}$, $D\subset\mathbb{R}^d$, 其 L_2 范式为 $\|\psi\|:=(\int_D|\psi(x)|^2dx)^{1/2}$,其内积为 $\langle\psi_1,\psi_2\rangle=\int_D\psi_1(x)\psi_2(x)dx$。

在时间序列数据驱动下的可解释学习模型中, 对于一个目标函数 $f:\mathbb{R}^d\to\mathbb{R}^m$, 假设已经产生了 $L-1$ 层隐藏层单元, 即 $f_{L-1}(x)=\sum_{j=1}^{L-1}\beta_j g_j(w_j^{\mathrm{T}}x+b_j)$ $(L=1,2,\ldots,$ $f_0=0)$, 其中 $\beta_j=[\beta_{j,1},\ldots,\beta_{j,m}]^{\mathrm{T}}$。学习模型的残差可表示为 $e_{L-1}=f-f_{L-1}=[e_{L-1,1},\ldots,e_{L-1,m}]$。若 $\|e_{L-1}\|$ 未达到预定的误差范围, 此时该学习模型的学习能力不够, 需要继续生成一个隐藏层单元即随机基函数 g_L (w_L and b_L) 并计算或优化其输出权重 β_L, 通过这种方式学习模型的残差 $f_L=f_{L-1}+\beta_L g_L$ 将进一步缩小。在本课题研究中, 学习模型的随机权重将受到约束并保证普适逼近能力。

对此, 有基准的学习模型 I。假设 span(Γ) 在 L_2 空间中是密集的, 且 $\forall g\in\Gamma$, $0<\|g\|<b_g, b_g\in\mathbb{R}^+$。有 [0-1] 区间的学习速率 $r, 0<r<1$ 和非负实数序列 $\{\mu_L\}$ 且 $\lim_{L\to+\infty}\mu_L=0, \mu_L\le(1-r)$。对于隐藏层神经元 $L=1,2,\ldots,$ 有:

$$\delta_L=\sum_{q=1}^{m}\delta_{L,q}, \delta_{L,q}=(1-r-\mu_L)\|e_{L-1,q}\|^2, q=1,2,...,m \tag{1.3}$$

若随机基函数 g_L 在满足以下不等式约束的前提下生成:

$$\langle e_{L-1,q},g_L\rangle^2\ge b_g^2\delta_{L,q}, q=1,2,...,m \tag{1.4}$$

且输出权重由以下公式计算得出:

$$\beta_{L,q}=\frac{\langle e_{L-1,q},g_L\rangle}{\|g_L\|^2}, q=1,2,\ldots,m \tag{1.5}$$

则可得出此学习模型具有普适逼近能力, 即 $\lim_{L\to+\infty}\|f-f_L\|=0$, 其中 $f_L=\sum_{j=1}^{L}\beta_j g_j, \beta_j=[\beta_{j,1},\ldots,\beta_{j,m}]^{\mathrm{T}}$。

该普适逼近能力可被证明为: 根据公式 1.5, 可以验证其残差 $\{\|e_L^2\|\}$ 是单调递减的。因此, $\{\|e_L\|\}$ 在 $L\to+\infty$ 时收敛。根据公式 1.3、公式 1.4和公式 1.5 , 有:

$$\begin{aligned}
&\|e_L\|^2-(r+\mu_L)\|e_{L-1}\|^2\\
=&\sum_{q=1}^{m}\left(\langle e_{L-1,q}-\beta_{L,q}g_L, e_{L-1,q}-\beta_{L,q}g_L\rangle-(r+\mu_L)\langle e_{L-1,q},e_{L-1,q}\rangle\right)\\
=&\sum_{q=1}^{m}\left((1-r-\mu_L)\langle e_{L-1,q},e_{L-1,q}\rangle-2\langle e_{L-1,q},\beta_{L,q}g_L\rangle+\langle\beta_{L,q}g_L,\beta_{L,q}g_L\rangle\right)\\
=&(1-r-\mu_L)\|e_{L-1}\|^2-\frac{\sum_{q=1}^{m}\langle e_{L-1,q},g_L\rangle^2}{\|g_L\|^2}
\end{aligned}$$

$$
\begin{aligned}
&= \delta_L - \frac{\sum_{q=1}^{m}\langle e_{L-1,q}, g_L\rangle^2}{\|g_L\|^2} \\
&\leq \delta_L - \frac{\sum_{q=1}^{m}\langle e_{L-1,q}, g_L\rangle^2}{b_g^2} \leq 0
\end{aligned} \tag{1.6}
$$

学习模型 I 提供了一个基于随机分布的可解释和可自适应调整网络结构的神经网络学习模型。不同于梯度下降算法优化整个神经网络的权重，学习模型 I 通过公式 1.4的不等式约束为新增的隐藏层神经元搜索合适的随机权重 w_L 与偏置 b_L。这类时间序列数据在金融市场和故障监测等实际情景中极为常见，同时由于 $\Psi(w,b) = \sum_{q=1}^{m}\langle e_{L-1,q}, g_L\rangle^2/\|g_L\|^2$ 在参数空间中为连续函数，满足公式 1.4约束的随机权重 w_L 与偏置 b_L 能够很快的被搜索出来。在公式 1.4的约束下，学习模型 I 不能能够随机的产生隐藏层权重，并且能够在根据时间序列数据自适应调整网络结构的同时满足模型对时间序列数据的普适逼近能力。

在学习模型 I 中，隐藏层单元的输出权重 $\beta_L = [\beta_{L,1}, \ldots, \beta_{L,m}]^{\mathrm{T}}$ 由 $\beta_{L,q} = \langle e_{L-1,q}, g_L\rangle/\|g_L\|^2$ 由计算得出，并在之后的神经网络调整中保持不变。考虑到这种确定方法可能会导致整个学习模型的收敛速度较慢，因此在学习模型 I 的基础上加以对输出权重计算方法可加以改进。具体为，在每次生成新的隐藏层单元即随机基函数 $g_j(j = 1, 2, \ldots, L)$ 后，对整个学习模型中隐藏层单元输出权重 $\beta_1, \beta_2, \ldots, \beta_L$ 采用最小二乘法进行优化，以此加快学习模型的逼近收敛速度。这种方法同样可以保证学习模型对时间序列数据的普适逼近能力。

对此，有改进的学习模型 II。已优化后的输出权重可表示为 $[\beta_1^*, \beta_2^*, \ldots, \beta_L^*] = \arg\min_\beta \|f - \sum_{j=1}^{L}\beta_j g_j\|$，$e_L^* = f - \sum_{j=1}^{L}\beta_j^* g_j = [e_{L,1}^*, \ldots, e_{L,m}^*]$，定义未优化输出权重为 $\tilde{\beta}_{L,q} = \langle e_{L-1,q}^*, g_L\rangle/\|g_L\|^2, q = 1, \ldots, m$，$\tilde{e}_L = e_{L-1}^* - \tilde{\beta}_L g_L$，其中 $\tilde{\beta}_L = [\tilde{\beta}_{L,1}, \ldots, \tilde{\beta}_{L,m}]^{\mathrm{T}}$ 且 $e_0^* = f$。

假设 span(Γ) 在 L_2 空间中是密集的，且 $\forall g \in \Gamma, 0 < \|g\| < b_g, b_g \in \mathbb{R}^+$。有 [0-1] 区间的学习速率 $r, 0 < r < 1$ 和非负实数序列 $\{\mu_L\}$ 且 $\lim_{L\to+\infty}\mu_L = 0, \mu_L \leq (1-r)$。对于隐藏层神经元 $L = 1, 2, \ldots$，有：

$$
\delta_L^* = \sum_{q=1}^{m}\delta_{L,q}^*, \delta_{L,q}^* = (1 - r - \mu_L)\|e_{L-1,q}^*\|^2, q = 1, 2, ..., m \tag{1.7}
$$

若随机基函数 g_L 在满足以下不等式约束的前提下生成：

$$
\langle e_{L-1,q}^*, g_L\rangle^2 \geq b_g^2\delta_{L,q}^*, q = 1, 2, ..., m \tag{1.8}
$$

且输出权重由以下公式计算得出：

$$[\beta_1^*, \beta_2^*, \ldots, \beta_L^*] = \arg\min_{\beta} \|f - \sum_{j=1}^{L} \beta_j g_j\| \tag{1.9}$$

则同样可得出此学习模型具有普适逼近能力，即 $\lim_{L\to+\infty} \|f - f_L^*\| = 0,$，其中 $f_L^* = \sum_{j=1}^{L} \beta_j^* g_j, \beta_j^* = [\beta_{j,1}^*, \ldots, \beta_{j,m}^*]^{\mathrm{T}}$。

该普适逼近能力可被证明为：对于 $\|e_L^*\|^2 \leq \|\tilde{e}_L\|^2 = \|e_{L-1}^* - \tilde{\beta}_L g_L\|^2 \leq \|e_{L-1}^*\|^2 \leq \|\tilde{e}_{L-1}\|^2, L = 1, 2, \ldots$，可验证其残差 $\{\|e_L^*\|^2\}$ 是单调递减的。因此，有：

$$\begin{aligned}
& \|e_L^*\|^2 - (r+\mu_L)\|e_{L-1}^*\|^2 \\
\leq\ & \|\tilde{e}_L\|^2 - (r+\mu_L)\|e_{L-1}^*\|^2 \\
=\ & \sum_{q=1}^{m} \Big(\langle e_{L-1,q}^* - \tilde{\beta}_{L,q} g_L, e_{L-1,q}^* - \tilde{\beta}_{L,q} g_L \rangle - (r+\mu_L) \langle e_{L-1,q}^*, e_{L-1,q}^* \rangle \Big) \\
=\ & \sum_{q=1}^{m} \Big((1-r-\mu_L) \langle e_{L-1,q}^*, e_{L-1,q}^* \rangle - 2 \langle e_{L-1,q}^*, \tilde{\beta}_{L,q} g_L \rangle + \langle \tilde{\beta}_{L,q} g_L, \tilde{\beta}_{L,q} g_L \rangle \Big) \\
=\ & (1-r-\mu_L)\|e_{L-1}^*\|^2 - \frac{\sum_{q=1}^{m} \langle e_{L-1,q}^*, g_L \rangle^2}{\|g_L\|^2} \\
=\ & \delta_L^* - \frac{\sum_{q=1}^{m} \langle e_{L-1,q}^*, g_L \rangle^2}{\|g_L\|^2} \\
\leq\ & \delta_L^* - \frac{\sum_{q=1}^{m} \langle e_{L-1,q}^*, g_L \rangle^2}{b_g^2} \leq 0
\end{aligned} \tag{1.10}$$

学习模型 II 中的输出权重通过摩尔-彭诺斯广义逆(Moore-Penrose Inverse)和最小二乘求解得出。在面对大规模数据集时，这种全局最小二乘方法亦会大大增加学习模型的计算复杂度。为了在学习模型的逼近收敛程度和计算复杂度之间进行权衡，可考虑加入滑动窗口机制。即在隐藏层层数大于窗口数时，仅对窗口内的隐藏层单元采用最小二乘法更新其输出权重，以此加快模型对大规模时间序列数据的逼近收敛速度。

对此，有改进的学习模型 III。假设 span(Γ) 在 L_2 空间中是密集的，且 $\forall g \in \Gamma$, $0 < \|g\| < b_g, b_g \in \mathbb{R}^+$。有 [0-1] 区间的学习速率 $r, 0 < r < 1$ 和非负实数序列 $\{\mu_L\}$ 且 $\lim_{L\to+\infty} \mu_L = 0, \mu_L \leq (1-r)$。对于一个给定的窗口 K 和隐藏层神经元 $L = 1, 2, \ldots,$ 有：

$$\delta_L^* = \sum_{q=1}^{m} \delta_{L,q}^*, \delta_{L,q}^* = (1-r-\mu_L)\|e_{L-1,q}^*\|^2, q = 1, 2, ..., m \tag{1.11}$$

若随机基函数 g_L 在满足以下不等式约束的前提下生成：

$$\langle e^*_{L-1,q}, g_L\rangle^2 \geq b_g^2 \delta^*_{L,q}, q = 1, 2, ..., m \tag{1.12}$$

在 $L \leq K$ 时，输出权重由以下公式计算得出：

$$[\beta_1^*, \beta_2^*, \ldots, \beta_L^*] = \arg\min_{\beta} \|f - \sum_{j=1}^{L} \beta_j g_j\| \tag{1.13}$$

否则，保持输出权重 $\beta_1^*, \ldots, \beta_{L-K}^*$ 不变，$\beta_{L-K+1}, \ldots, \beta_L$ 由以下公式计算得出：

$$[\beta_{L-K+1}^*, \beta_{L-K+2}^*, \ldots, \beta_L^*] = \arg \min_{\beta_{L-K+1},\ldots,\beta_L} \|f - \sum_{j=1}^{L-K} \beta_j^* g_j - \sum_{j=L-K+1}^{L} \beta_j g_j\| \tag{1.14}$$

同样可得出此学习模型具有普适逼近能力，即 $\lim_{L\to+\infty} \|f - f_L^*\| = 0,$，其中 $f_L^* = \sum_{j=1}^{L} \beta_j^* g_j, \beta_j^* = [\beta_{j,1}^*, \ldots, \beta_{j,m}^*]^{\mathrm{T}}$。

基于此，完成了大规模时间序列数据驱动下具有普适逼近能力、可解释可自适应调整神经网络结构的学习模型。后续的课题研究工作将着手于构建随机激活函数下的可解释学习模型，以期更快的逼近时间序列数据，并进行相关的实验验证。

参考文献

[1] Cybenko G. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems (MCSS), 1992, 5(4):455–455.

[2] Chen T, Chen H. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. IEEE Transactions on Neural Networks, 1995, 6(4):911–917.

[3] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. Nature, 2017, 550(7676):354–359.

[4] Lewis M, Yarats D, Dauphin Y N, et al. Deal or no deal? end-to-end learning for negotiation dialogues. arXiv preprint arXiv:1706.05125, 2017..

[5] He K, Wang Y, Hopcroft J. A powerful generative model using random weights for the deep image representation. Advances in Neural Information Processing Systems, 2016. 631–639.

[6] Kwok T Y, Yeung D Y. Objective functions for training new hidden units in constructive neural networks. IEEE Transactions on neural networks, 1997, 8(5):1131–1148.

[7] Mahoney M W, et al. Randomized algorithms for matrices and data. Foundations and Trends® in Machine Learning, 2011, 3(2):123–224.

[8] Scardapane S, Wang D. Randomness in neural networks: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2017, 7(2).

[9] Cui C, Wang D. High dimensional data regression using lasso model and neural networks with random weights. Information Sciences, 2016, 372:505–517.

[10] Pao Y H, Takefuji Y. Functional-link net computing: theory, system architecture, and functionalities. Computer, 1992, 25(5):76–79.

[11] Tyukin I Y, Prokhorov D V. Feasibility of random basis function approximators for modeling and control. Control Applications,(CCA) & Intelligent Control,(ISIC), 2009 IEEE. IEEE, 2009. 1391–1396.

[12] Gorban A N, Tyukin I Y, Prokhorov D V, et al. Approximation with random bases: Pro et contra. Information Sciences, 2016, 364:129–145.

[13] Li M, Wang D. Insights into randomized algorithms for neural networks: Practical issues and common pitfalls. Information Sciences, 2017, 382:170–178.