文章编号:1671-4598(2020)06-0052-04 DOI:10.16526/j.cnki.11—4762/tp.2020.06.011 中图分类号:TP2 文献标识码:A

基于随机权神经网络集成模型的实时遥测数据处理

贾海艳
(中国人民解放军92941部队,辽宁 葫芦岛 125000)

摘要:飞行任务中的遥测数据是快速产生的时间序列数据流,其受测量设备和空间环境等因素影响易产生数据的漂移,由于过程进化属性,其数据分布属性也会发生变化,传统单一数据预测模型无法反应数据自身特征属性的这一变化;因此,提出一种联合具有随机权重的神经网络和装袋算法的集成方法实现对遥测数据的在线回归预测,设计的算法能根据数据特征属性变化而进行自主更新;利用基模型的多样性和低训练复杂度,同时满足数据处理的精度和实时性要求;通过实验仿真,结果表明该方法能明显抑制遥测数据的漂移现象,数据的预测精度提高近10 m。

关键词:遥测数据;集成;神经网络;随机权值;优化

An Ensemble Based on Neural Networks with Random Weights for Real—timeTelemetry Data Processing

Jia Haiyan
(Unit 92941 Element of the PLA, Huludao 125000, China)

Abstract: Telemetry data of flight task are time—series data streams sequentially and rapidly. Data drift Influenced by measurement devices and spatial environment is produced, the evolving nature of processes may often cause changes of data distribution, which is difficult to detect and causes loss of accuracy in data prediction accuracy. a single forecast models can't adapt the change of data feature attribute. As a consequence, an ensemble model was put off that combine neural networks with random weights algorithms and bagging algorithm, it is able to update actively according to possible changes in the data distribution. By use of the diversity of base model, low training complexity and dynamic updating mechanisms, it is a accurate algorithm that can operate in a computational time. Through experimental simulation, the results show that the method can obviously restrain the drift of telemetry data, and the prediction accuracy of the data is improved by nearly 10 m.

Keywords: telemetry data; ensemble; neural networks; random weights; optimization

0 引言

遥测数据主要用于获取飞行器状态和轨迹信息以对其进行安全控制,高精度的遥测数据处理结果是评估的重要依据。遥测数据处理具有数据量大、数据结构复杂、高维度及高实时性等特征,同时遥测数据易受测量设备和空间环境等影响产生数据漂移变化,其表现可能是数据的突变形式,也有可能是缓慢的线性变化等,变化趋势不易判断,因此建立一种高精度的在线遥测数据预测模型十分必要[1]。

目前机器学习快速发展,并广泛应用于各类复杂场景中的数据建模,通常的有监督学习方法假设数据的概率分布在训练集和实际数据集间不会发生变化,而实际应用场景中的数据的分布由于过程进化特性常常是不稳定的,会随时间而变化,进而造成随时间推移模型的预测精度下降[2]。因此要建立的遥测数据预测模型不仅要求预测结果具有较高的精度,而且能对数据分布变化敏感,可以快速适应数据漂移的变化,同时算法要满足在飞行规定的时间内完成数据处理,

考虑到以往使用单一预测模型存在的不稳定性,而集成算法模型能有效解决模型的泛化性和可信度[3],因此提出一种基于随机权神经网络(NNRW)的装袋(bagging)集成方法[4],用于解决遥测数据流的在线回归预测问题。集成模型的关键是基模型的选择[5],选择随机权神经网络作为基模型,其输入层和隐层间权值随机初始化,在优化过程中保持不变,而对隐藏层和输出层间的权值进行优化,同传统神经网络训练算法相比,这样可极大降低训练复杂度[6]。选择装袋集成方法,通过随机有放回的取样方式保证训练出的基模型间的独立性,保证所训练基模型的多样性,通过基模型定点更新策略保证集成模型对数据漂移的适应。仿真实验表明基于具有高精度和高效性[7]随机权神经网络和装袋集成学习机制可实现基模型的并行训练,该集成方法满足对数据处理的实时性和精度要求,通过基模型更新机制,减小数据漂移对模型预测精度的影响。

1 随机权神经网络的装袋集成模型

该集成学习模型采用装袋集成方法,通过随机采样方

收稿日期:2020-04-06; **修回日期**:2020-04-20。

作者简介:贾海艳(1974-),辽宁锦州人,硕士,高级工程师,主要从事数据处理方向的研究。

法对原始数据集抽取出多个训练集，随机性抽样保证了各训练集各自独立又包含有数据集的共同特征，进而训练得到具有多样性的随机权神经网络基模型池，从池中选择最优的基模型用于预测输出。该集成方法通过样本的随机性，保证训练出的随机权神经网络基模型具有多样性[8]，通过赋予高精度的基模型更大的输出权值，联合多个高精度基模型的加权输出实现对输出数据预测，保证最终数据预测结果的精度。同时为应对输入数据存在的漂移现象，当数据出现漂移或模型预测精度下降，对构成输出的基模型采用更新机制，通过新增数据训练获得新的基模型，对集成模型中输出性能较差的基模型进行替换，保证用于预测的基模型为最优。为满足遥测数据处理要求，保证该集成模型数据处理的实时性，由于随机权神经网络的结构和学习过程简单，在对新增数据在线学习过程中可将若干神经网络同时训练，极大缩减了集成模型的构造时间。

1.1　随机权神经网络（NNRW）基模型

基模型的设计要求具有较高的预测精度以及计算效率，以满足实时遥测数据处理的要求，因此选择单隐藏层前向反馈神经网络结构的随机权神经网络作为基模型，基模型原理如图 1 所示。

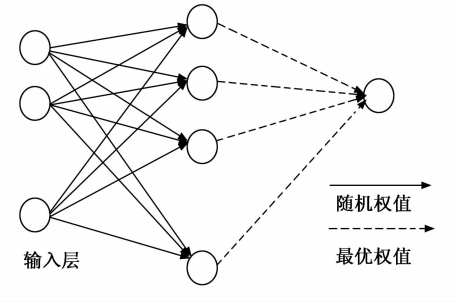

图 1　单隐层前馈神经网络基模型

图 1 中输入层和隐藏层间的权值随机选择，在训练过程中保持不变，而隐藏层与输出层间的权值通过训练获得。学习算法选择岭回归，学习函数如式（1）所示：

$$T = g(X \cdot W_H + B) \cdot W_o \tag{1}$$

式中，T 为目标向量，X 为输入训练向量，W_H 为输入层到隐藏层的权重向量，B 为偏置向量，W_o 为隐藏层到输出层权重向量。$(g)\ \cdot$ 为激活函数，如式（2）所示：

$$g(x) = \frac{1}{1+e^{-x}} \tag{2}$$

由于 W_H 和 B 为随机选择向量，且在训练过程中保持不变，训练函数变为线性系统，如式（3）所示：

$$T = H \cdot W_o \tag{3}$$

其中：H 为隐藏层输出，通过式（4）计算：

$$H = g(X \cdot W_H + B) \tag{4}$$

最优向量 W_o 满足式（5）：

$$W_o^* = (H^T \cdot H + C \cdot \boldsymbol{I})^{-1} \cdot \mathrm{H}^{\mathrm{T}} \cdot \mathrm{T} \tag{5}$$

C 为小的常数，$\boldsymbol{I}$ 为作为惩罚项的单位矩阵。

随机权神经网络通过随机选取内权与偏置值，将网络参数需要计算的问题转化成线性方程，通过不同的组合进行求解，采用广义逆求解方程组的最小二乘解作为网络外权，通过外权与内权的结合计算，避免了其他传统算法的缺点，极大地减少了训练时间，有效避免了陷入局部最小循环的问题。

1.2　随机权神经网络的装袋集成算法

由于遥测数据含有多维信息，进行该类数据处理的模型通常极其复杂，另外遥测数据具有高的数据吞吐量，因此要求数据处理具有较高的实时性，为满足模型实现在线对数据进行预测计算的要求[9]，将随机权神经网络和装袋集成方法进行联合，其模型复杂度为 $O(M(N^3))$，模型中的每个随机权神经网络基模型可单独立优化，各基模型优化可并行执行。集成模型的多样性通过装袋集成方法的自助采样（bootstrapping）算法实现，自助采样不仅产生新的训练数据集用于基模型的训练，并且能保证从训练集中抽取出特征属性，特征属性的数量用于构建每个基模型，特征属性的数量根据总特征量的百分比得出。

集成算法的具体原理如下：

1.2.1　基模型的选择优化

集成模型中包含多个基模型，它们利用对原始训练集的采样集进行训练获得各自的最优参数，因此能保证基模型的多样性，每个基模型都可以对输入数据集进行预测，但集成模型的输出只由所选择的最优 Q 个基模型决定，给定修剪率 $p \in \mathbb{R}^{(0,1]}$，参加预测的基模型的数量 Q 满足式（6）：

$$Q = p * M \tag{6}$$

对每个数据集 C，对具有最小输出误差的 Q 个模型（$Q < M$）进行联合，作为集成模型的基模型，剩余基模型处于无效状态，但考虑到这些模型可能携带从以前样本集中学习的有用信息，因此只是暂时不参与集成输出，但随时对模型输出性能进行跟踪，在新的数据集中满足精度要求时，则激活该基模型，重新纳入集成模型。

由于输入数据存在数据漂移现象，势必造成集成模型的预测精度下降，当集成模型中某个基模型精度较差时，对该基模型暂时免于激活，而从处于失效状态而随时跟踪数据的基模型中选择预测精度较佳者进行替换。

1.2.2　最优权值确定

集成模型的输出结果为所选择最优的 Q 个基模型加权输出和。该权值直接影响集成模型性能，同时过于复杂的优化方法会增加算法的计算时间。因此选择的方法为集成模型中的每个基模型根据其在最近数据块的预测精度被分配权重。给定当前数据集 C，输出由 M 个基模型（$m=1, 2, \cdots, M$）集成，每个基模型的权值计算如式（7）所示：

$$w_m = \frac{1}{mse_m} \tag{7}$$

mse_m 为第 m 个基模型在当前数据集 C 上计算得到的均方误差。数据样本 x_n 的集成输出 y_E 如式（8）所示：

$$y_E(x_n)=\frac{\sum_{m-1}^{M}w_m*\hat{y}_m(x_n)}{\sum_{m=1}^{M}w_m} \tag{8}$$

式中，$\hat{y}_m$（x_n）为第 m 个基模型在数据样本 x_n 的输出。

由式（8）可知，当基模型的预测输出的均方误差值较大时，则在集成模型的输出中所占比例相应较小。当预测均方误差值较小，则分配较大权重。

1.2.3 基模型的更新

为解决输入数据的漂移问题，参与集成输出的基模型需要进行更新，另外当基模型数量不满足设计要求时需要使用最新的数据集中的标记数据训练出新的基模型，在输入数据集中随机划分出 70％用于训练，30％用于验证。如果新的模型精度好于已存在模型中精度最差者，则进行模型更新，这个过程一直重复进行，直到新模型的数量达到设计要求。给定替换率为 $r\in\mathbb{R}^{(0,1]}$，新模型的数量 M_{new} 计算如式（9）所示：

$$M_{new}=round(r*M) \tag{9}$$

更新机制不仅保证集成模型与最近的数据同步，而且作为一种自然选择机制，能不断剔除低性能的基模型。

1.2.4 集成算法工作流程

随机权神经网络集成模型工作原理如图 2 所示。

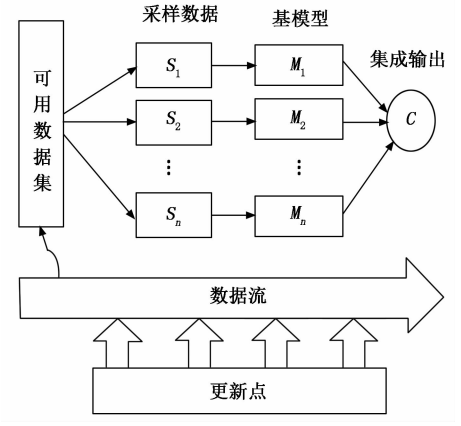

图 2 随机神经网络集成模型原理

将飞行遥测数据作为数据流，初始要求提供足够可用数据集用于构建一个原始的集成模型，该可用数据集作为原始数据集，采样数据 S 为对可用数据集进行有放回的随机采样获得，每个采样集用来训练得到一个作为基模型 M 的随机权神经网络，通过对 N 个基模型训练获得基模型池，选择最优的 Q 个基模型进行加权输出实现对输入数据的预测。初始集成模型构建完成之后开始对输入数据流进行预测，同时上述更新机制开始工作，由于在数据流中设置有更新点标志，当检测到更新点标志时，表示可用数据集为新增的数据集，一部分基模型对新增数据进行在线学习，当判断集成模型预测性能下降时对预测性能较差的基模型进行更新操作。

2 集成模型超参数的优化

模型的超参数包括基模型的个数，每个神经网络内隐层层节点的个数，惩罚因子，随机权重值的范围等，超参数的选择优化对模型精度起着关键作用，而且超参间还具有相互作用，例如，在超参数 A 为 1 级别，算法精度随超参数 B 从 1～2 变化时提高，同时在 A 为 2 级别，算法精度随超参数 B 从 1～2 变化时降低，因此在超参数较多的情况下，增加了优化过程的复杂性。目前用于数据流预测的机器学习算法中，还缺乏系统的方法对超参数进行优化，一些方法如手动调整，网格搜索，贝叶斯超参数优化等，缺乏对某个超参数的重要性和超参数间的相互作用的考虑[10]。

鉴于使用全因子的实验设计（DOE）方法进行超参数调整能识别显著的超参数间的相互作用，因此提出采用全因子设计方法对集成模型的超参数进行调整，将初始可用数据集的前 1 000 个数据随机划分为 70％用于训练，30％用于测试。具体过程分两步进行，首先对所有超参数在预定义的 5 个级别进行穷尽组合计算，通过对每个超参数的灵敏度进行分析，以及对超参数之间的相互作用的分析识别出重要的超参数和优化方式。之后开展新实验进行超参数的微调，通过将具有低重要性的超参数保持在固定水平，缩小参数的搜索空间，进而保证了算法的实时性。模型超参数具体设置如表 1 所示。

表 1 超参数预设值表参数

因子	级别				
M	40	60	80	100	120
N	$8x$	$10x$	$12x$	$14x$	$16x$
R	0.000 1	0.001 0	0.002 0	0.010 0	0.050 0
W	[−0.5, 0.5]	[−0.75, 0.75]	[−1.0, 1.0]	[−1.25, 1.25]	[−1.5, 1.5]
A	0.6	0.7	0.8	0.9	1.0

其中，M 为构成集成模型的基模型数量；N 为基模型中隐藏层节点数量，N 为输入数量的函数，具体为 A 因子与输入数量的积；R 为惩罚因子，用于在优化过程中惩罚大的权值；W 为随机权值分布区间，该参数确定初始化随机权值均匀分布在该区间内，模型精度起关键作用；A 为特征属性数量，用于随机选择输入的一部分用于基模型进行训练。

3 仿真及分析

选取飞行器位置、速度、弹道等 10 个属性的仿真数据作为训练集，其中飞行弹道为目标输出，其他参数作为输入向量集，利用本文提出的方法建立预测模型。

为评估模型对数据漂移的响应，基于遥测数据构造 5 000采样点的数据集，将数据特征属性的变量域划分为 10 级，用前 7 级构造数据集的前 2 000 点，之后每隔 1 000 点扩展变量范围，人为形成数据的漂移现象。

3.1 超参数优化仿真及分析

每个超参数都决定模型的性能，但重要性不同，用于超参数优化的数据集为起始的 1 000 点数据，将前 1 000 点数据划分为两部分，70％用于训练，30％用于验证。每个

超参数按表 1 划分的 5 个级别，共进行 10 次优化过程处理，经统计分析对超参数重要性进行评定，从而确定超参数优化顺序。根据 F_0 显著性统计确定最重要的 3 个超参数顺序为随机权值范围 W、基模型数量 M 和神经网络中的节点数 N，获取参数最优设置如表 2 所示。

表 2　最优超参数设置

超参数	W	M	N
最优值	±1.5	100	14

通过超参数优化过程，也证明了隐藏层节点数量和随机权值范围间相互影响，如图 3 所示。

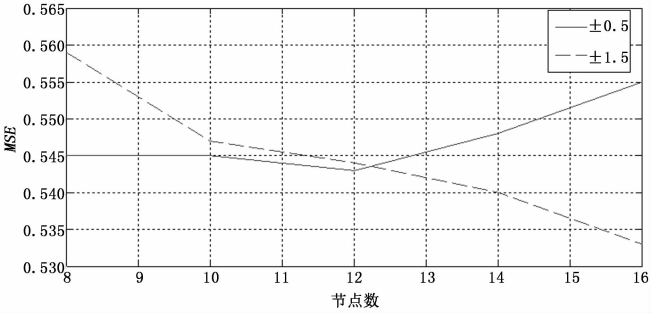

图 3　超参数间相互关系

3.2　遥测数据仿真分析

为评估集成模型性能，基于余下 4 000 点数据集进行仿真测试，测试结果的 MSE 曲线如图 4 所示。

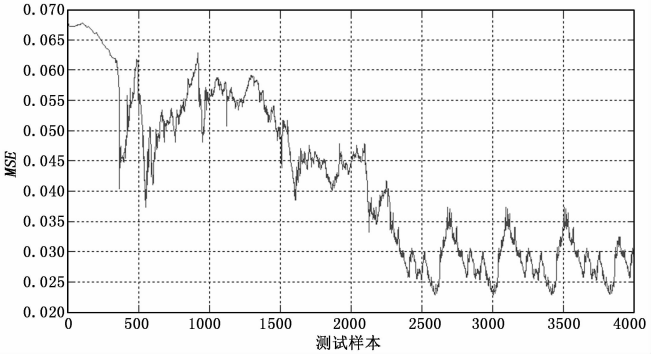

图 4　模型预测误差曲线

由图 4 可见，在漂移点出现后，模型仍具有较高的预测精度，说明该模型具有明显的抑制数据漂移的作用。

集成模型不仅提升了数据预测的精度，借助超参数最优化和基模型更新策略，也极大缩减了模型更新时间，模型更新时间可控制在 0.02 s 内完成。

4　结束语

针对遥测数据预测模型中，要求处理满足实时性和抑制数据漂移的问题，提出一种采用随机权值神经网络和 bagging 集成相联合的集成方法，通过随机的 bootstrap 采样保证训练集和随机权神经网络基模型的独立性和多样性，利用随机权神经网络的高效性，降低基模型训练的复杂性，缩短了基模型训练时间，提高了集成模型的处理速度，通过基模型的多样性和在线更新机制，提高了在数据出现漂移时模型预测的精度。通过对飞行遥测数据的仿真实验，结果表明该模型对遥测数据的预测精度得到了提升，对数据漂移现象具有抑制作用，同时该方法具有快速性等特点。

参考文献：

[1] 闫谦时，崔广立. 基于时间序列的航天器遥测数据预测算法 [J]. 计算机测量与控制，2017，25 (5)：188－191.

[2] Zhou Z H，Wu J X，Tang W. Ensembling neural networks：Many could be better thanall [J]. Artificial Intelligence：An International Journal，2002 (1/2)：239－263.

[3] Alhamdoosh M. Wang D H. Fast decorrelated neural network ensembles with random weights [J]. Information Sciences：An International Journal，2014：104－117.

[4] Monther A，Wang D H. Fast Decorrelated neural network ensembles with random weights [J]. Information Sciences，2014，264 (6)：104－117.

[5] 杨孝玉. 基于随机权神经网络的重置多分类算法研究 [D]. 杭州：浙江工商大学，2018.

[6] 姜　乐，周　平. 优化增量型随机权神经网络及应用 [J]. 化工学报，2019，70 (12)：4710－4721.

[7] 乔俊飞，李凡军，杨翠丽. 随机权神经网络研究现状与展望 [J]. 智能系统学报，2016，11 (6)：758－767.

[8] 张天伦. 多隐层前馈神经网络的随机赋权训练算法研究 [D]. 保定：河北大学，2016.

[9] Ding J，Wang H，Li C，et al. An online learningneural network ensemble with random weights for regression ofsequential data stream [J]. Soft Comput.，2017，21 (20)：5919－5937.

[10] 梁青青. 基于关键超参数选择的监督式 AutoML 性能优化 [D]. 贵阳：贵州大学，2019.

(上接第 31 页)

[9] Jasper W J，Gamier S J，PotlaPalli H. Texture characterization and defect detection using adaptive wavelets [J]. Optical Engineering，1996，35 (11)：3140－3149.

[10] Chan C H，Pang G K H. Fabric defect detection by Fourier analysis [J]. IEEE Transactions on Industry Applications，1999，36 (5)：1267－1276.

[11] Jasper W J，Gamier S J，Potla P H. Texture characterization and defect deteetion using adaptive Wavelets [J]. Optical Engineering，1996，35 (11)：3140－3149.

[12] Harinath D，Babu K R，Satyanarayana P，et al. Defect detection in fabric using wavelet transform and genetic algorithm [J]. Transactions on Machine Learning and Artificial Intelligence，2016，3 (6)：10.

[13] 赵宏威，王亦红. 基于改进 Gabor 优化选择的布匹瑕疵检测方法 [J]. 计算机工程与应用，2019，55 (24)：202－207.

[14] 辛斌杰，余序芬，吴兆平. 应用图像分析技术自动识别织物的组织结构 [J]. 东华大学学报：自然科学版，2011，37 (1)：35－41.

[15] Krizhevsky A，Sutskever I，Hinton G E. Imagenet classification with deep convolutional neural networks [C]. Advances in Neural Information Processing Systems (NIPS)，2012：1097－1105.