学号：18120255

**《机器学习基础》**

**课程论文**

**基于高斯过程的时序数据回归预测**

|  |  |
| --- | --- |
| **姓 名** | 姚施越 |
| **学 号** | 18120255 |
| **论文评分** |  |

**2021年 ？月？日**

目 录

1、引言............................................................1

2、基于改进粒子群算法的模糊神经网络................................2

2.1粒子群算法 ..................................................2

2.2改进的粒子群算法.............................................2

2.3模糊神经网络.................................................3

2.4基于改进粒子群算法的模糊神经网络.............................5

3、数据实验 .......................................................5

3.1实验设置...................................................5

3.1.1样本区间的选择原则.....................................5

3.1.2结构变量的选择原则.....................................5

3.2数据处理..................................................6

3.3对比实验..................................................6

4、结论与展望......................................................7

基于高斯过程的时序数据回归预测

[摘要]针对粒子群算法容易造成群体次优信息被忽略和神经网络存在容易陷入局部极值等问题，本文将粒子群算法与模糊人工神经网络进行融合，提出一种基于改进粒子群算法的模糊神经网络的方法，充分发挥粒子群算法全局寻优的优势。以某地天气预报作为实例，建立了基于改进粒子群算法的多模型模糊神经网络预报模型，试验结果表明该方法对于天气预报具有较好的准确度。

[关键字]天气预报，改进粒子群算法，模糊神经网络

**An Improved PSO-Based Fuzzy Neural Network**

**and Its Application in Weather Forecast**

**Abstract:** In this paper, particle swarm optimization (PSO) and fuzzy neural network (FNN) are proposed to solve the problem. Because PSO is easy to cause the sub-optimal information of the group to be neglected and the existence of neural network is easy to fall into the local extremum. Neural network method gives full play to the advantages of global optimization of particle swarm optimization algorithm. Based on the weather forecast, a multi-model fuzzy neural network forecasting model based on improved particle swarm optimization is established. The experimental result shows that the method has good accuracy for weather forecasting.

**Key words:** Weather Forecast; Improved PSO; Fuzzy Neural Network

**1、引言**

时间序列，是将某种统计指标的数值，按时间先后顺序排到所形成的数列。时间序列预测就是通过编制和分析时间序列，根据时间序列所反映出来的发展过程、方向和趋势，进行类推或延伸，借以预测下一段时间或以后若干年内可能达到的水平【1】，主要包括趋势分析，相似性搜索，与时间有关数据的序列模式挖掘和周期模式挖掘【2】。需要注意的是，时间序列可以分为平稳序列，即存在某种周期，季节性及趋势的方差和均值不随时间变化的序列，及非平稳序列。

传统预测方法主要包括两种，一种起源于统计学领域，如移动平均法，指数平均法等；另一种是目前仍广泛使用的，包括AR,MA,ARMA,ARIMA以及灰色模型和回归模型等，这些方法比较适用于小规模，单变量的预测。李志超等针对上海市月度居民消费价格指数，分布建立ARIMA模型，灰色模型和一元n阶多项式回归模型，对三者的预测精度和适用场景做了比较[2]。

现代预测方法则是将机器学习和深度学习领域的算法引入，具体选用何种方法取决于应用场景。如杨璐等将神经网络作为传统的时序线性模型的非线性推广进行了分析，提出了一种可作为非线性时序模型的内反馈神经网络[9]。张玉瑞等采用基于OLS算法的RBF神经网络，利用某公司国际航班1949-1959年每月的乘客数预测1960年的乘客数，证明了该方法在很大程度上解决了过拟合的问题，提高了网络的泛化能力【】。Md. Rafiul Hassan等则将隐马尔可夫模型，人工神经网络和遗传算法融合用于预测金融市场行为【】。

针对时间序列缺失值处理，目前主流方法有三大类。

第一类是直接删除法，该方法可能会舍弃数据中的一些重要信息； 第二类是基于统计学的填充方法，如均值填充，中值填充，常用值填充，但这类方法忽略了数据的时序信息； 第三类是基于机器学习的填充方法，常见的方法有基于KNN（K-Nearest Neighbor），RNN（Recurrent Neural Networks） ，EM（Expectation-Maximization）和矩阵分解（Matrix Factorization）的缺失值填充算法。然而，这类方法也很少考虑到两个相邻数据间的时序信息。

传统的缺失值处理方法包括直接删除法、直接填充统计数据法以及基于机器学习算法的填充法。然而上述缺失值处理方法均没有考虑到时序数据中的时间先后信息,故难以取得准确的填充效果。

Luo, Yonghong等结合生成对抗网络技术，提出了一种基于生成对抗网络的端到端生成模型，在插补精度中优于baseline，训练中也拥有更好的时间效率【】。

本文着重讨论如何对存在缺失值的时序数据进行回归预测，以某制药公司研制的对菌群抑制的药物实验过程中，菌群分泌物的含量变化的时序数据为主体展开分析。首先将高斯过程相关概念引入，之后利用其回归模型进行缺失值的插补。其次，建立以XGBoost为首的决策树模型，对插补处理之后的数据进行分类，并与原标签进行比对，作为缺失值插补效果的检验。最后，使用高斯过程回归，根据数据现有变化规律对后续变化进行预测。此外，选取目前该领域主流的回归模型，设计相关实验，将本模型与之比较，对模型的有效性和抗干扰性进行评估。

**2、基于高斯过程的回归预测**

**2.1高斯过程**

高斯过程是观测值出现在一个连续域（例如时间或空间）的随机过程，它是概率论和统计学领域的重要模型。其定义中

在高斯过程中，连续输入空间中每个点都是与一个正态分布的随机变量相关联。此外，这些随机变量的每个有限集合都有一个多元正态分布，换句话说他们的任意有限线性组合是一个正态分布。高斯过程的分布是所有那些（无限多个）随机变量的联合分布，正因如此，它是连续域（例如时间或空间）上函数的分布。[wiki]

粒子群优化算法（PSO）是由Kennedy和Eberhart于1995 年提出的一种新的进化算法[5]，该算法来源于对鸟群体觅食行为的研究，其理论依据是演化计算理论。鸟群在随机搜索食物过程中，如果某区域里只有一块食物，那么找到食物的最简单有效的策略是搜索目前离食物最近的鸟的邻近区域。PSO就是从这种模型中得到启发而产生的，粒子群算法通过个体之间的协作来寻找最优解[6]。粒子群算法可用于解决大量非线性 、不可微和多峰值的复杂问题的优化，已经被广泛应用于如函数优化、神经网络、模式分类等科学和工程领域。

PSO算法初始时随机产生目标函数的一组解，该解空间内的每个解被称作粒子。每个粒子通过跟随当前最优粒子的方法在求解问题的空间内搜索最优解。 按照式（1）和（2）更新每个粒子的位置和速度：

 (1)

 (2)

式（1）中的pbest 表示该粒子从开始到现在搜索产生的最优解，gbest 表示粒子群目前找到的最优解，即运行到当前代为止所找到的全局最优。，被称作惯性因子，用于根据上一次迭代得到的速度调节本次选代过程中粒子的运动速度。C1 和C2 是加速系数，分别调节向全局最好粒子和个体最好位子方向飞行的最大步长，若太小，则粒子可能远离目标区域，若太大则可能会导致飞过目标区域。r1 和r2 是（0,1）之间的随机散。根据目标函数计算的适应度值用来衡量搜索空间内哪个位置更好。适应度值驱使各位于朝着搜索空间内个体最优和到目前为止发现的全局最优方向移动。

**2.2改进的粒子群算法**

与遗传算法相比，粒子群算法也是一种基于群体的迭代算法。通过群体性能在解空间进行优化计算，因此在某些方面二者有些类似。相比而言，粒子群算法的操作比较简单，需要调整的参数也不多。而遗传算法因为模拟遗传进化过程，需要事先进行编码，然后在迭代过程中进行交叉和变异操作以避免群体退化。由于遗传算法基于一定的概率操作，因此收敛的速度不及粒子群算法快，但是比较稳定。粒子群算法中群体最优gbest部分作为主要的收敛动力，虽然加速了收敛，但造成了其他群体次优信息被忽略。为了弥补彼此的不足，扬长避短，产生了二者的结合[7]，在每次速度改变的调整中引入遗传算法的思想，进行交叉变异操作，从而改善整体算法的寻优能力。

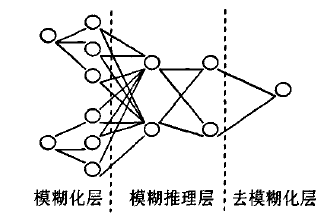
自适应权值调整（AWIF）的粒子群算法就是讲粒子的权值调整与适应度函数的输出关联起来。杨尚东等人提出的AWIF中粒子权值的改变以适应函数平均值favg(vi)为界进行分类[8]。虽然可以改善了粒子的全局寻解能力，但是由于权值改变基于简单的适应度均值进行分界，在每个时刻，都有较多的粒子保持在最大权值状态，不利于算法的稳定和收敛。本文提出改进的自适应权值调整（I-AWIF）的粒子群算法，兼顾权值的自适应特性，不进行基于均值的分类，将权值调整与适应函数的输出相关联，在权值的自适应调整算法中，粒子权值的调整更平滑。粒子权值的具体计算公式如下：

 (3)

其中粒子f(vt)为当前粒子位置对应的适应函数值，fgbest为群体最优对应的适应函数值，fworst为最差的适应函数值。在每次迭代计算中，每个粒子权值的改变有其适应函数自适应决定，当粒子接近群体最优是，权值降低，粒子趋于稳定，变现出较好的收敛型，从而减少震荡的产生[9]。当粒子距离群体最优较远时，权值增大，粒子具有较大的速度，从而保证较好的空间搜索能力。通过改进的自适应权值调整（I-AWIF）的粒子群算法，既维护了粒子群本身的收敛能力，也改善了传统粒子群LPSO算法的空间搜索能力的不足。

**2.3模糊神经网络**

模糊神经网络，即在传统神经网络分类器基础上引入模糊逻辑，构造模糊神经网络分类器。其网络结构一般可采用多层前馈网络，一般分为三大部分，如图1所示。



**图1.** 模糊神经网络结构

第一部分为模糊化层，它实现输入变量的模糊化，完成一个隶属函数的计算，计算出变量相对于每个模糊集合的隶属度。

第二部分是模糊推理层。模糊推理层是网络结构中的重要部分，它联系着模糊推理的前提和结论，实现网络的模糊映射。模糊推理层的结构是多样化的，本文将采用的是BP网络。

第三部分是去模糊化层，去模糊的常用方法有最大隶属原则和模糊质心法。本文暂不考虑。

1. 设有n个训练的样本集合，每个样本有M项预报因子特征值，则有实测预报因子特征值矩阵：

 (4)

（4）式中，i=1，2，...，m；j=1，2，...，n；xij为待训练样本j的预报因子特征值i的实测值。

1. 应用隶属度公式对样本进行归一化，即：

 (5)

（5）式中，，分别为样本的j个预报因子的最大值、最小值。

1. N个预测对象组成样本集合，其特征量为：

 (6)

1. 应用隶属度公式，对预测对象进行归一化，即：

 (7)

（7）式中，，分别为预测对象的最大、最小值。

**2.4基于改进粒子群算法的模糊神经网络**

考虑到PSO算法的全局寻优能力，以及反向传播算法善于局部细致搜索和发展较为成熟的特点，在对模糊神经网络训练时，采用二者相结合的方法。粒子群优化算法中，适应度函数采用基于神经网络输出与期望值的误差平方和，在Matlab中表示为：

 (8)

算法具体步骤如下：

（1）根据初始约束条件，初始化粒子群；

（2）根据当前粒子位置，将粒子群中每个粒子的坐标作为神经网络对应的权值与阈值；

（3）根据预先定义的训练集合P，得到网络仿真输出，并根据公式（8）与实际真值集合*t*，计算粒子群的适应度；

（4）根据粒子适应度，计算Xpbest与Xgbest，并调整每个粒子的速度矢量Vt；

（5）根据Vt生成*t*+1时刻的新群，判断结束条件，不满足则返回步（2）；

（6）在PSO全局寻优的基础上，运行小步长反向传播算法，进行局部细致搜索，达到要求的收敛精度，则结束网络训练。

**3、数据实验**

**3.1实验设置**

**3.1.1样本区间的选择原则**

（1）区间的选择要参考天气变化本身的特点；

（2）样本容量尽可能的大，使其满足模型预测的无偏性和有效性的要求；

（3）尽量避免无关样本的干扰，以免影响模型的预测能力。

综合以上的原则，在初步分析数据的基础上发现，每年不同时期的天气特点不同，而不同年份相同时期的特点相似。本文决定选择适当多的年份数据作为样本，并将分十二个月，对每个月份分别建立一个模型，进行训练、预测。

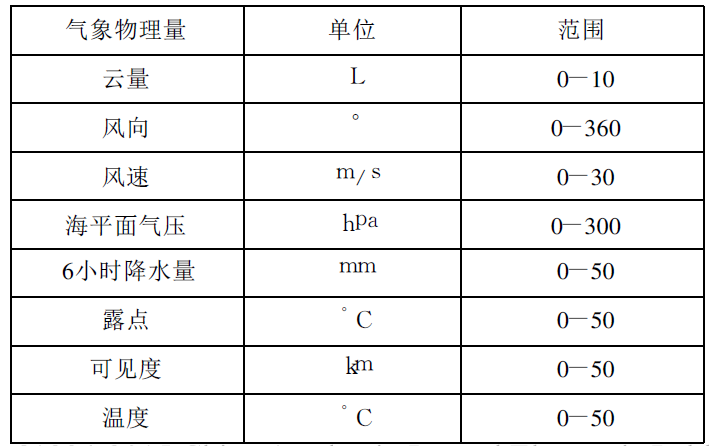
**3.1.2结构变量的选择原则**

（1）尽量包含较多的输入信息，使其容错性更强；

（2）尽可能排除输入变量之间的相关性，提高其网络学习时间。

在以上原则的前提下，本文选择的结构变量如表1所示。

**表1**. 选定的结构变量



**3.2数据处理**

本文实验中所用的数据是某地雷达观测的每三小时一次的实测数据。早数据获取过程中，难免会出现一些明显的错误数据，在一开始就尽量将他们剔除，以便剔除对神经网络训练的无谓扰动。在数据分析时，若遇到某一时刻数据的遗失或者被剔除时，这时对数据进行插值。

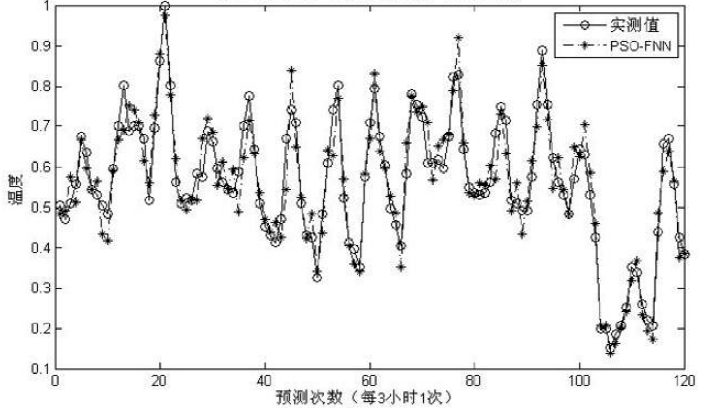
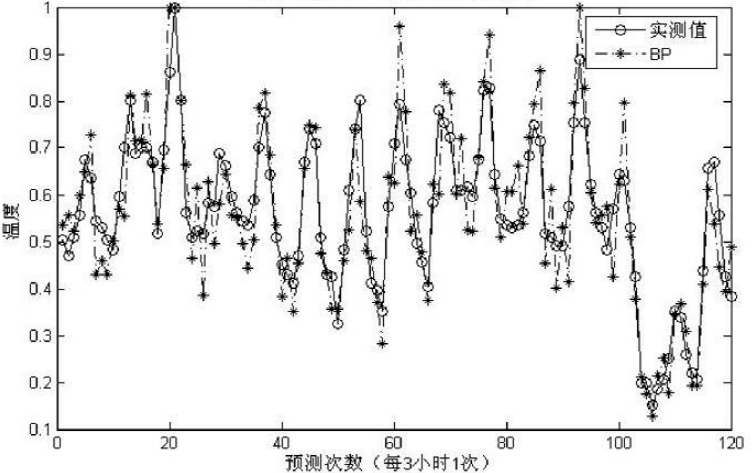
由于原始数据值的大小和量纲的不统一，因此在进行网络训练之前，必须对原始数据进行归一化处理，避免其过大造成网络麻痹。本文使用以下的线性关系作为映射函数：

 (9)

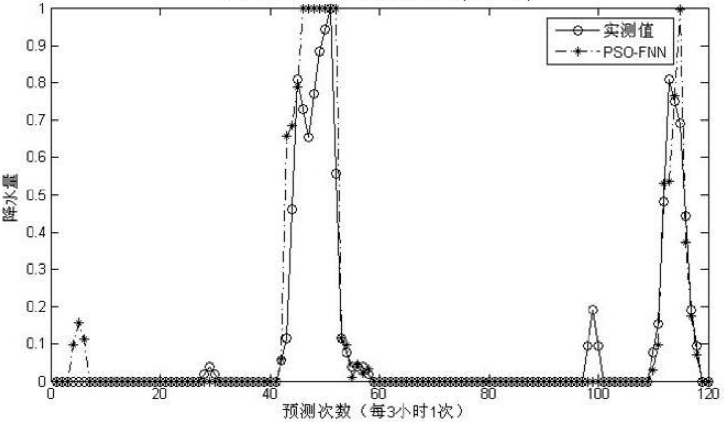
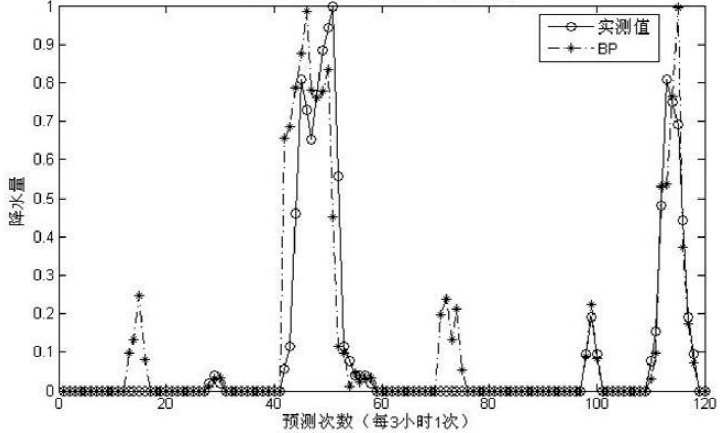
经过剔除有害数据、插值和归一化等预处理过程，最后生成训练样数据和检验数据。

**3.3 对比实验**

本文的建模是在Matlab下实现的，实验中粒子的数量选取为100，粒子初始化的下届为-1，上届为1，粒子最大速度Vmax=0.5，权值最大值Wmax=0.95，权值最小值Wmin=0.3，迭代次数为100次。本文对某地2009年6月的温度和降水进行了预报，基于改进PSO算法的模糊神经网络和BP神经网络的预报结果分别如图2-图5所示[9]。

**图2**. PSO-FNN温度预报 **图3**. BP温度预报

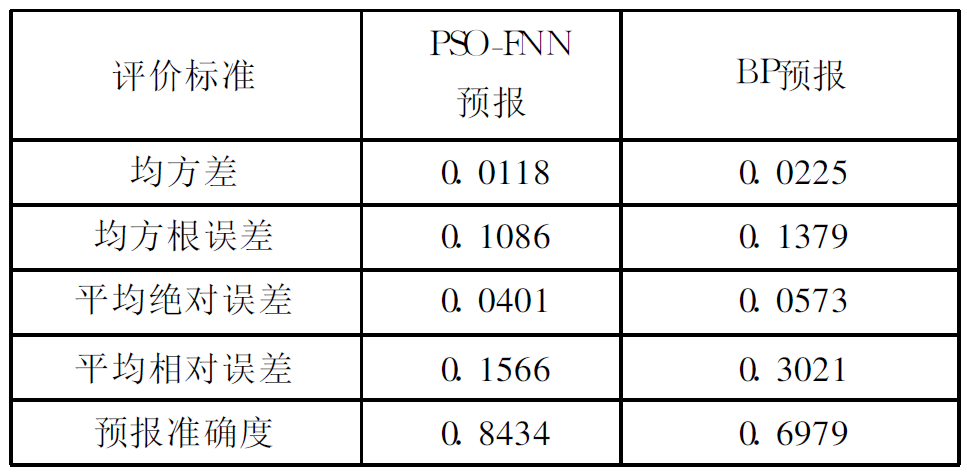
**图4.** PSO-FNN降水预报 **图5.** BP降水预报

本文统计了2009年整年的预报评价指标以及降水预报的错报率和漏报率，如表2-表4所示。

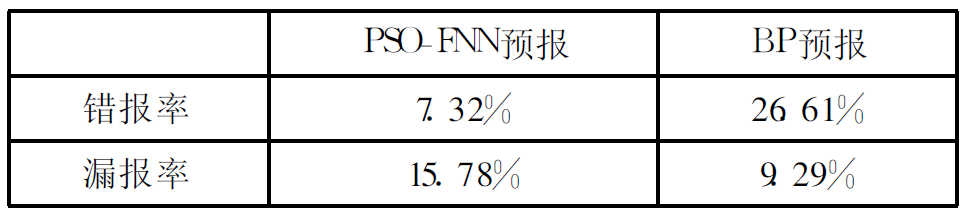
**表2**. 基于PSO-FNN与BP神经网络温度预报结果比较



**表3**. 基于PSO-FNN与BP神经网络降水预报结果比较



**表4.** 降水错报率和漏报率



通过对实验和统计结果的分析，可以看出基于改进粒子群算法的模糊神经网络在精度上有着明显的优势，尤其是温度的预报取得较为满意的结果；在降水方面的预测，基于改进粒子群算法的模糊神经网络预报的漏报率较高，但错报率位于可以接受的值空间， 而BP神经网络预报的错报率较高，在实际的天气预报中，将二者结合起来可能会取得更好的效果。

**4、结论与展望**

本文将粒子群算法与模糊人工神经网络进行融合，提出一种基于改进粒子群算法的模糊神经网络的方法，充分发挥粒子群算法全局寻优的优势。以某地天气预报作为实例，建立了基于改进粒子群算法的多模型模糊神经网络预报模型，试验结果表明该方法对于天气预报具有较好的准确度。如果气象专家的知识和经验能够用模糊推理规则表示出来，并应用到模糊神经网络来进行天气预报，相信会取得更好的效果。

**参考文献**

[1]彭昱忠，王谦，元昌安，林开平.数据挖掘技术在气象预报研究中的应用[J].干旱气象,2015,33(1):19-27.

[2]金龙,吴建生.基于遗传算法的神经网络短期预报预测模型[J].灾害学,2004,8(1): 15-16．

[3]段文广,周晓军,石永炜.数据挖掘技术在精细化温度预报中的应用[J].干旱气象，2012,30(1):130-135．

[4]Enireddy Vamsidhar，Varma K V S RP,Sankara RaoP，et al. Prediction of Rainfall Using Back propagation Neural Network Mode[J].International Journal on Computer Science and Engi- neering,2010,2(4):1119-1121．

[5]Sendonaris A, Erkip E, Aazhang B.User cooperation diversity-Part I and II [J]. IEEE Transactions on Communications, 2003 ,51(11):1927-1948.

[6]江涛，张玉芳，王银辉.一种改进的粒子群算法在BP网络中的应用研究[J].计算机科学，2006，33（9）：164-166.

[7]Peter J. Angeline Evolutionary Optimization Versus Particle Swarm Optimization, Philosophy and Performance Difference[J]. Lecture Notes In Computer Science, 1998, 1447:601-610.

[8]Yang Shangdong. Li Xiang. A New Ann Optimized By Improved PSO Algorithm Combined With Chaos And Its Application In Short-term Load Forecasting[J]. IEEE Transaction on Computational Intelligence and Security, 2006 International Conference, 2006, 2:945-948.

[9]周岩,王盛,高传善,孙慰迟.基于改进粒子群算法的模糊神经网络及其在短时天气预报中的应用[J].计算机应用与软件,2010,27(5):234-237.