学号：21122873

**《机器学习基础》**

**课程论文**

**基于梯度直方图的特征提取在MNIST数据集训练与测试数据分布存在差异下的学习优化实践**

|  |  |
| --- | --- |
| **姓 名** | 郑力铖 |
| **学 号** | 21122873 |
| **论文评分** |  |

**2024年 1月9日**

目 录

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

2、小样本检测优化算法设计..........................................2

2.1支持向量机 ..................................................2

2.2卷积神经网络.................................................2

2.3梯度直方图...................................................3

2.4数据增强.....................................................3

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

基于梯度直方图的特征提取在MNIST数据集训练与测试数据分布存在差异下的学习优化实践

[摘要] 在训练数据相对测试数据较少的情况下，直接使用SVM和CNN方法的训练效果相对较差。因此，本文提出了一种基于梯度直方图的特征提取方法，旨在通过对梯度信息的敏感度分析，增强模型对训练数据的适应性。通过实验验证，我们展示了该方法在MNIST数据集中相较于普通方法的优越性，为解决实际应用中针对训练数据较少的分类问题提供了一种有效的解决途径。本研究对于提升模型泛化性能和小样本学习的实际应用具有一定意义。

[关键字]支持向量机 卷积神经网络 梯度直方图 小样本学习

**Experiment of Differences in Training and Testing Data Distribution with Feature Extraction Based on Gradient Histograms in MNIST Dataset**

**Abstract:** When the training data is relatively small compared to the test data, the training effect of directly using SVM and CNN methods is relatively poor. Therefore, this article proposes a feature extraction method based on gradient histograms, aiming to enhance the model's adaptability to training data by analyzing the sensitivity of gradient information. Through experimental verification, we have demonstrated the superiority of this method over ordinary methods in the MNIST dataset, providing an effective solution for the classification problem with limited training data in practical applications. This study has certain significance for improving model generalization performance and practical applications of small sample learning.

**Key words:** Support Vector Machine; Convolutional Neural Network; Gradient Histogram; Small Sample LearningS

**1、引言**

天气预测是根据气象观测资料，应用天气学、动力气象学、统计学等原理和方法，参考某一区域的气候背景和天气演变规律的基础上，对该区域未来一定时段的天气状况做出定性或定量的预测[1]。天气预测本身无论对于国家还是个人都有其特殊的意义 , 它与军事 、农业 、航空航天以及人民生产生活都息息相关。

近年来，出现了很多种基于神经网络的算法对天气方面进行研究，如金龙等用神经网络方法开展了不同区域范围的暴雨预报研究，认为神经网络方法确实可以通过对网络的学习训练[2]。段文广等基于时序分析技术，建立适合于BP神经网络的输入样本模型，将其用于未来24h的精细化温度预报[3]。Vamsidhar等构建了一个反向传播神经网络模型，对印度某地的湿度、露点温度和气压数据预测降水量进行预测[4]。然而神经网络的训练通常是一个复杂的大规模优化问题，随着训练样本集和网络初始权值的变化，网络的训练结果随机性较大，存在容易陷入局部极值等缺点，这给实际建模带来了一定的困难。

本文结合模糊数学理论和神经网络天气预测研究的最新成果，将改进的粒子群优化算法（PSO）与模糊优选人工神经网络（FNN）进行融合，建立了基于PSO优化算法的模糊神经网络多模型天气预测模型。通过引入实例，将PSO-FNN模型的预测结果与BP神经网络模型预报结果进行比较，对模型的有效性进行评价。

**2、基于改进粒子群算法的模糊神经网络**

**2.1粒子群算法**

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。

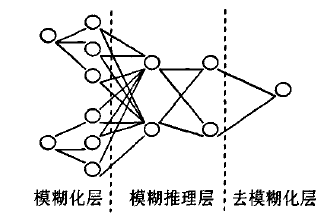
自适应权值调整（AWIF）的粒子群算法就是讲粒子的权值调整与适应度函数的输出关联起来。杨尚东等人提出的AWIF中粒子权值的改变以适应函数平均值favg(vi)为界进行分类[8]。虽然可以改善了粒子的全局寻解能力，但是由于权值改变基于简单的适应度均值进行分界，在每个时刻，都有较多的粒子保持在最大权值状态，不利于算法的稳定和收敛。本文提出改进的自适应权值调整（I-AWIF）的粒子群算法，兼顾权值的自适应特性，不进行基于均值的分类，将权值调整与适应函数的输出相关联，在权值的自适应调整算法中，粒子权值的调整更平滑。粒子权值的具体计算公式如下：

 (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。

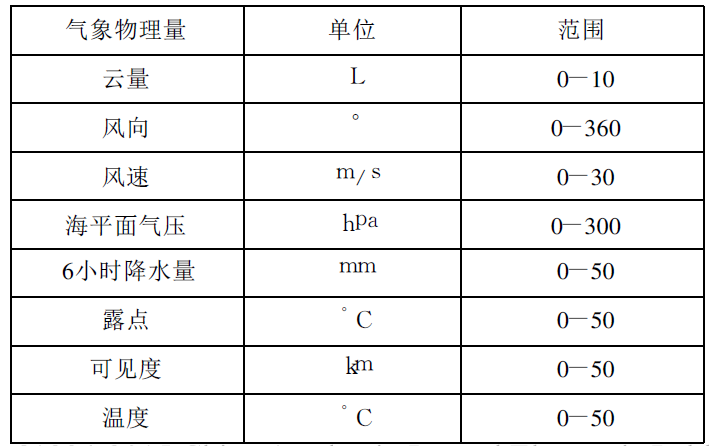
**3.1.2结构变量的选择原则**

（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。

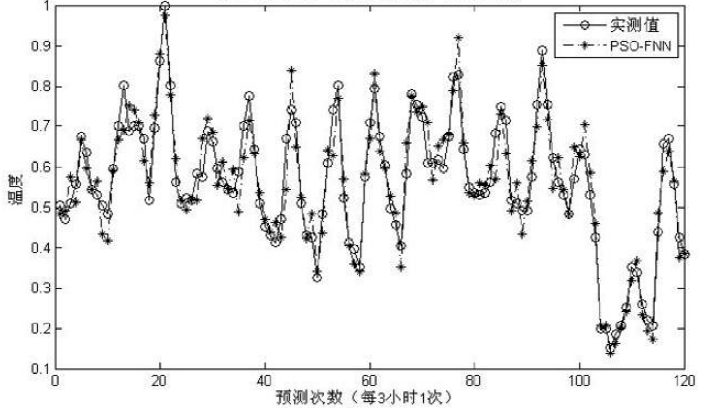
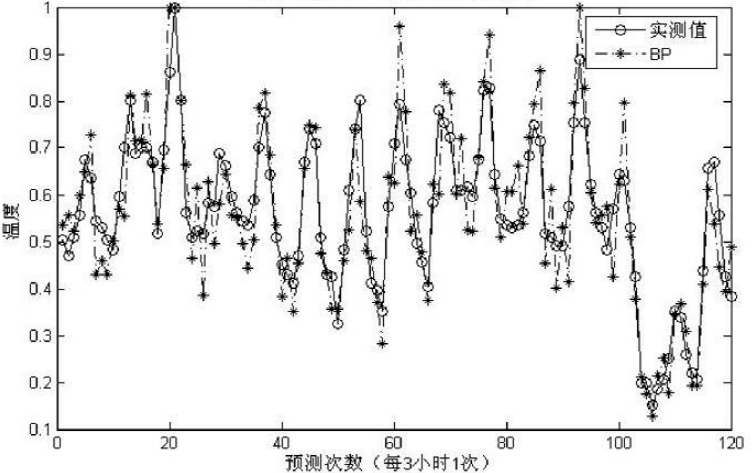
由于原始数据值的大小和量纲的不统一，因此在进行网络训练之前，必须对原始数据进行归一化处理，避免其过大造成网络麻痹。本文使用以下的线性关系作为映射函数：

 (9)

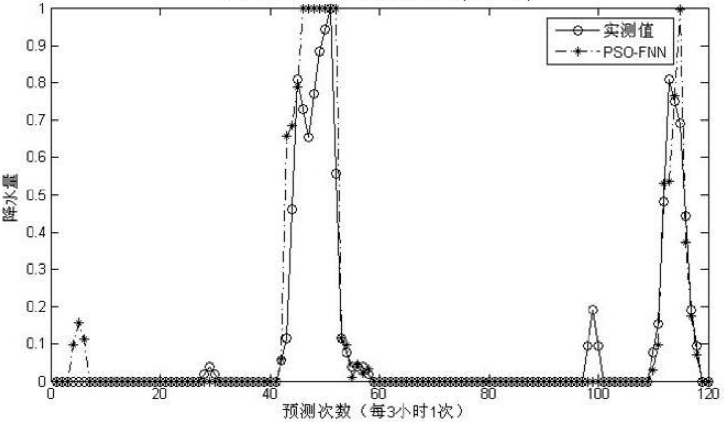
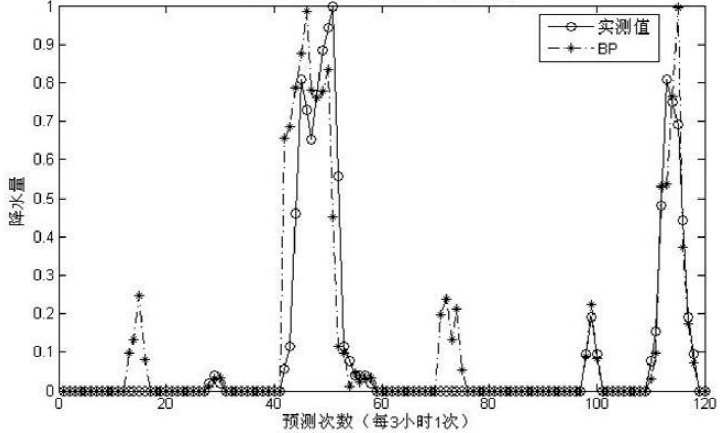
经过剔除有害数据、插值和归一化等预处理过程，最后生成训练样数据和检验数据。

**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

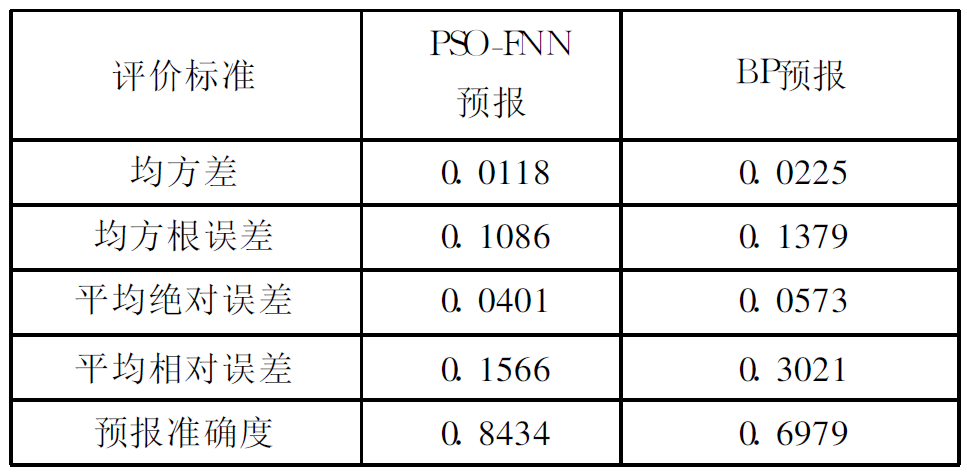
**图4.** PSO-FNN降水预报 **图5.** BP降水预报

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。

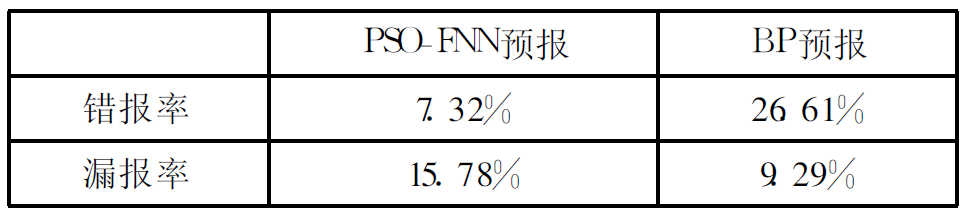
**表2**. 基于PSO-FNN与BP神经网络温度预报结果比较



**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] Cortes, C., & Vapnik, V. (1995). "Support-vector networks. " Machine Learning, 20(3), 273-297.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

[3]