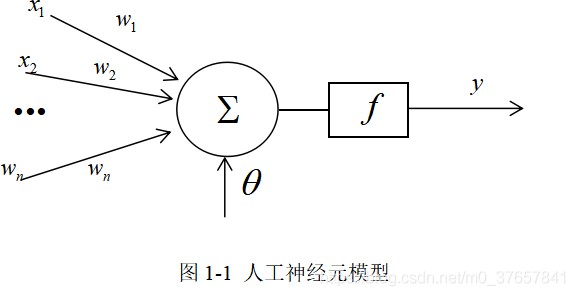
BP神经网络训练算法的比较

# 摘 要

BP神经网络可以有效地对非线性系统进行逼近，但是传统的最速下降搜索方法存在收敛速度慢的问题。本文通过对常用的BP神经网络训练算法进行比较，说明了不同训练算法的适用范围，为不同场景下BP神经网络训练算法的选择提供了实验依据。

**关键词**：BP神经网络；训练算法；适用范围一、BP神经网络的原理

# 一、人工神经网络概述

人工神经网络(Artificial Neural Network，简称ANN)由大量神经元广泛互联组成，是人功能的某种简化、抽象及模拟。在神经网络中，通过大量神经元之间的 相互联系存储知识与信息，神经元之间的相互作用实现知识信息的处理。一个经过训练的神经网络即具备类似人的模式识别、联想记忆的功能。 神经元的结构模 型如图1-1所示。

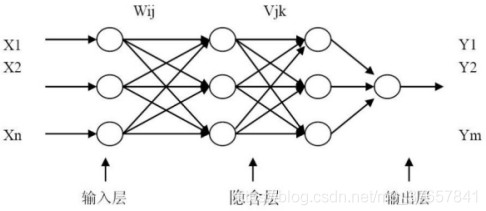
其中为神经元的输入信号。

为突触强度或联结权。

 为神经元的阈值或称为偏差用 表示。

是激励函数（或者叫作用函数）。

 是神经元的输出。

基本BP算法包括信号的前向传播和误差的反向传播两个过程。即计算误差输出时按从输入到输出的方向进行，而调整权值和阈值则从输出到输入的方向进行。正向传播时，输入信号通过隐含层作用于输出节点，经过非线性变换，产生输出信号，若实际输出与期望输出不相符，则转入误差的反向传播过程。误差反传是将输出误差通过隐含层向输入层逐层反传，并将误差分摊给各层所有单元，以从各层获得的误差信号作为调整各单元权值的依据。通过调整输入节点与隐层节点的联接强度和隐层节点与输出节点的联接强度以及阈值，使误差沿梯度方向下降，经过反复学习训练，确定与最小误差相对应的网络参数(权值和阈值)，训练即告停止。此时经过训练的神经网络即能对类似样本的输入信息，自行处理输出误差最小的经过非线形转换的信息。BP神经网络的一般结构如图2-2所示。

神经元结点的作用函数一般选用Sigmoid函数，即:

误差函数一般选用：



神经网络的训练过程就是是最小误差准则来不断调整网络结点的权值, 直至误差小于。可以看到，BP神经网络实际可转化为求解一个最优化问题，它的目标函数就是：

# 二、BP神经网络的训练算法

梯度下降法及其优化

梯度下降法是利用一阶的梯度信息找到函数局部最优解的一种方法，也是机器学习里面最简单最常用的一种优化方法。梯度下降法是直线搜索方法中的一种，主要迭代公式如下：



其中，是第 k 次迭代我们选择移动的方向，在梯度下降法中，移动的方向设定为梯度的负方向，是第 k 次迭代用直线搜索方法选择移动的距离，每次移动的距离系数可以相同，也可以不同，有时候我们也叫学习率（learning rate）。

在数学上，移动的距离可以通过直线搜索令导数为零找到该方向上的最小值，但是在实际编程的过程中，这样计算的代价太大，我们一般可以将它设定为一个常量。

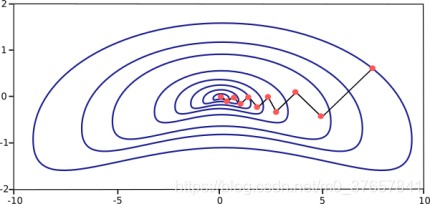
梯度下降方法有一个严重的弊端，若函数的梯度变化如图所示呈现出细长的结构时，该方法需要进行很多次迭代运算。而且，尽管梯度下降的方向就是损失函数值减小最快的方向，但是这并不一定是收敛最快的路径

图2-1 梯度下降法的迭代过程

标准的梯度梯度下降算法是一种贪心算法，虽然可以很有效地求解优化问题，但是可能会陷入局部极小值。为了避免寻找优化解的过程中陷入局部极小值，D.E.Rumelhart提出了一种方法[2]，该方法是在修改规则中增加一个动量项。该项会考虑之前时刻梯度的贡献，其权值迭代公式如下所示：

(2-2)

第一项是常规算法的学习因子项，第二项是动量项，a为动量因子。由于加入了之前时刻梯度的贡献，这相当于给迭代过程添加了一个低通滤波器，可以使网络忽略误差曲面上的细节特征，从而避免了陷入周部极小值的问题。

另外，在基本网络算法中，一般学习因子项n取为固定值，为了更好的控制网络的收敛性和学习速度，可以根据需要选择自适应的可变学习因子，则上述方法称之为自适应学习算法。

牛顿法和拟牛顿法

在梯度下降法中，我们看到，该方法主要利用的是目标函数的局部性质，具有一定的“盲目性”。牛顿法则是利用局部的一阶和二阶偏导信息，推测整个目标函数的形状，进而可以求得出近似函数的全局最小值，然后将当前的最小值设定近似函数的最小值。相比梯度下降法，牛顿法带有一定对全局的预测性，收敛性质也更优良。牛顿法的主要迭代公式如下：

(2-3)

对于高维函数，其二阶导数就变为了一个海塞矩阵，记为，那么迭代公式就变成了：

(2-4)

我们可以看到，当 为正定的时候，可以保证牛顿法的搜索方向是向下搜索的。

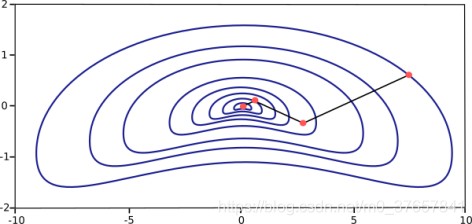
牛顿法的性能如图2-2所示。

图2-2

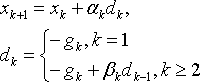
从相同的初始值开始寻找损失函数的最小值，它比梯度下降方法需要更少的步骤。然而，牛顿法的难点在于准确计算海森矩阵和其逆矩阵需要大量的计算资源。

由于牛顿法需要计算海森矩阵和逆矩阵，需要较多的计算资源，因此出现了一个变种算法，称为拟牛顿法，可以弥补计算量大的缺陷。此方法不是直接计算海森矩阵及其逆矩阵，而是在每一次迭代估计计算海森矩阵的逆矩阵，只需要用到损失函数的一阶偏导数。

海森矩阵是由损失函数的二阶偏导数组成。准牛顿法的主要思想是用另一个矩阵来估计海森矩阵的逆矩阵，只需要损失函数的一阶偏导数。

海森矩阵逆矩阵的估计有多种不同类型。根据其类型拟牛顿法的类型也不同，常用的拟牛顿法有DFP算法和BFGS算法等。

## 共轭梯度法

共轭梯度法是介于梯度下降法与牛顿法之间的一个方法，它仅需利用一阶导数信息，但克服了梯度下降法收敛慢的缺点，又避免了牛顿法需要存储和计算海塞矩阵并求逆的缺点，共轭梯度法不仅是解决大型线性方程组最有用的方法之一，也是解大型非线性最优化最有效的算法之一。在各种优化算法中，共轭梯度法是非常重要的一种。其优点是所需存储量小，具有步收敛性，稳定性高，而且不需要任何外来参数。共轭梯度法的迭代公式为:

(2-5)

# 三、实验与分析

为了比较上述算法的性能，使用matlab进行仿真实验。

## 实验数据集

Iris数据集是常用的分类实验数据集，由Fisher收集整理。Iris也称鸢尾花卉数据集，是一类多重变量分析的数据集。数据集包含150个数据集，分为3类，每类50 个数据，每个数据包含4个属性[4]。可通过花萼长度，花萼宽度，花瓣长度，花瓣宽度4个属性预测鸢尾花卉属于三个种类 中的哪一类。

本实验中将Iris数据集分为2组，每组各75个样本，每组中每种花各有25个样本。其中一组作为训练样本，另外一组作为检验样本。为了方便训练，将3类花分别编号为1，2，3。

## 创建BP神经网络。

使用matlab的神经网络工具箱建立BP神经网络模型。本实验采用如图2-2所示的BP神经网络通用模型，其中，输入层神经元数目为4，分别读取Iris数据集的4个属性，即花萼长度、花萼宽度、花瓣长度和花瓣宽度。共设置两个隐含层，隐含层节点数目分别为10和3。输出层神经元数目为3，分别对应样本属于某一品种的可能性的大小。

实验流程

使用matlab编写代码实现神经网络的训练过程，具体做法如下：

## 读取训练数据

第一节中已经把Iris数据集分成了训练样本和检验样本，首先读取训练样本。

特征值归一化

把读取到的训练样本数据都归一化到[0,1]之间。

## 创建神经网络

编写matlab代码实现第二节中的神经网络。

设置训练参数

对BP神经网络的训练参数进行设置，本实验中取训练精度为0.01，最大训练次数为5000次，学习率设为固定值0.01，其余参数设为默认即可。

## 开始训练

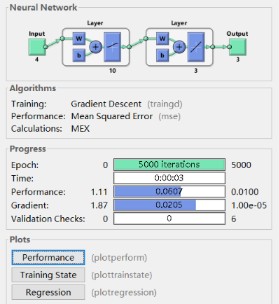
对神经网络进行训练，matlab神经网络工具箱可对训练过程进行实时跟踪，如图3-1所示 

图3-1 matlab神经网络工具箱

6）读取测试数据并进行测试

对第一节中Iris数据集的检验样本进行读取，然后输入神经网络中进行测试。

7）对实验结果进行分析

使用matlab神经网络工具箱对神经网络的训练过程进行可视化分析，并统计神经网络的样本识别正确率。

## 实验结果分析

## 梯度下降法及其改进算法的实验

首先对梯度下降法及其改进算法进行实验，分别使用标准梯度下降法，带有动量项的梯度下降法，以及带有动量项的自适应学习算法进行对比实验，实验结果如图3-2至3-4所示

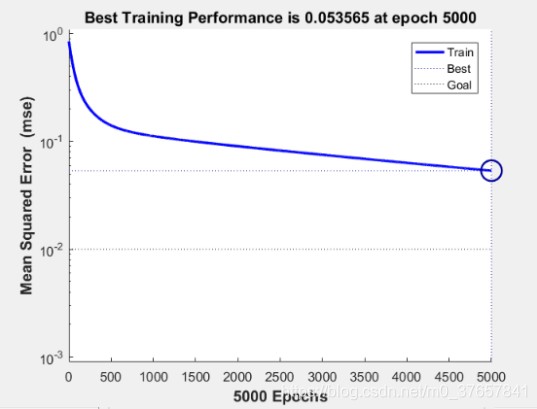


图3-2 标准梯度下降法训练性能，识别率为86.667%

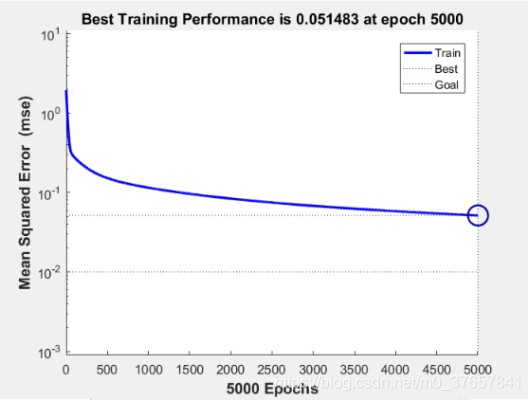


图3-3 带有动量项的梯度下降法训练性能，识别率为94.667%

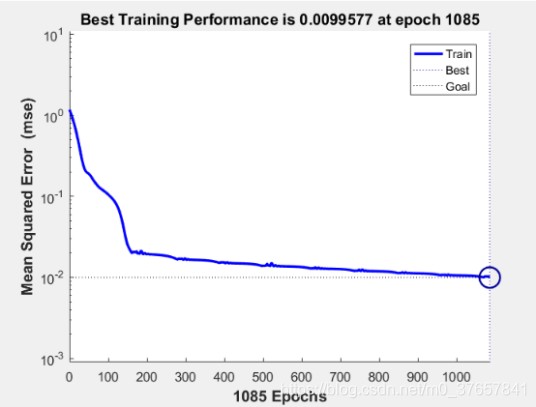


图3-4 带有动量项的自适应学习算法训练性能，识别率为97.333%

从上述实验结果中可以很明显的看到，标准梯度下降法的训练效率最低，即使经过5000次训练之后仍然达不到精度要求，并且其网络的识别率仅为86.667%。 相比较而言，带有动量项的梯度下降法的识别率更高，而且其初始下降速度明显比标准梯度下降法要快，但是在接近最优解时的下降速度也会逐渐减慢，最后仍然达不到精度要求。这三种方法中，带有动量项的自适应学习算法仅需1085次迭代即可达到精度要求，并且其识别率明显高于其他两种算法，达到了

97.333%。

## 牛顿法和共轭梯度法的实验

本部分主要对牛顿法和共轭梯度法中存储量要求最小的算法进行对比实验，实验环境和梯度下降法中的实验环境相同，实验结果如图3-5和3-6所示。

c

图3-5 牛顿法训练性能，识别率为97.333%

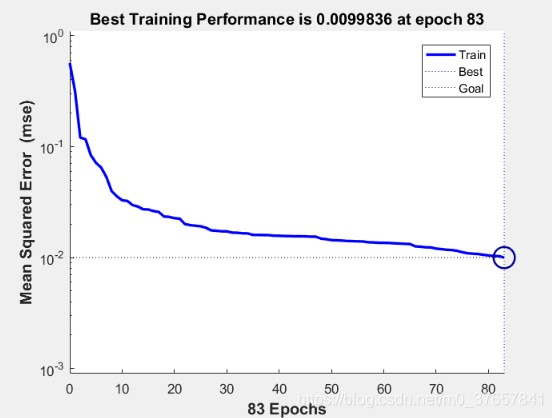


图3-6 共轭梯度法的训练性能，识别率为97.333%

从实验结果中可以看出，这两种方法的收敛速度均远远高于梯度下降法的收敛速度，仅需100步以内就能达到精度要求，其中，共轭梯度法的收敛速度虽然略慢于牛顿法，但其不需要计算海塞矩阵，故其训练速度远快于牛顿法，这种速度在大型网络中极为明显。两种方法最后达到的识别率均为97.333%，说明其训练的神经网络的泛化能力基本相同。

Levenberg-Marquardt算法的实验

本部分只对LM算法进行实验，实验环境和前文相同，实验结果如图3-7所示。

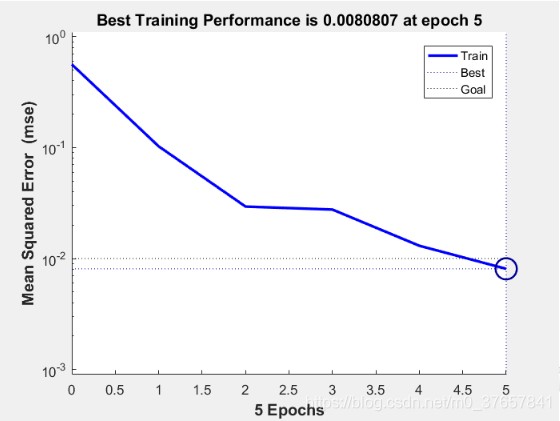


图3-7 LM算法的训练性能，识别率为97.333%

从实验结果中可以看出，LM算法在这几种方法中的收敛速度最快，仅仅只需5步即可达到精度要求，并且其识别率也可达到97.333%。但本实验中所采用的神经网络规模较小，因此并不能明显对比出各个算法内存占用的差异，而LM算法正是适用于中小型网络，对于大型网络，其内存开销将会非常庞大。

# 四、结论

图4-1绘制了本文讨论的几种算法的计算速度和内存需求。如图所示，梯度下降法往往是最慢的训练方法，它所需要的内存也往往最少。相反，速度最快的算法一般是Levenberg-Marquardt算法，但需要的内存也更多。拟牛顿法则较好地平衡了两者。

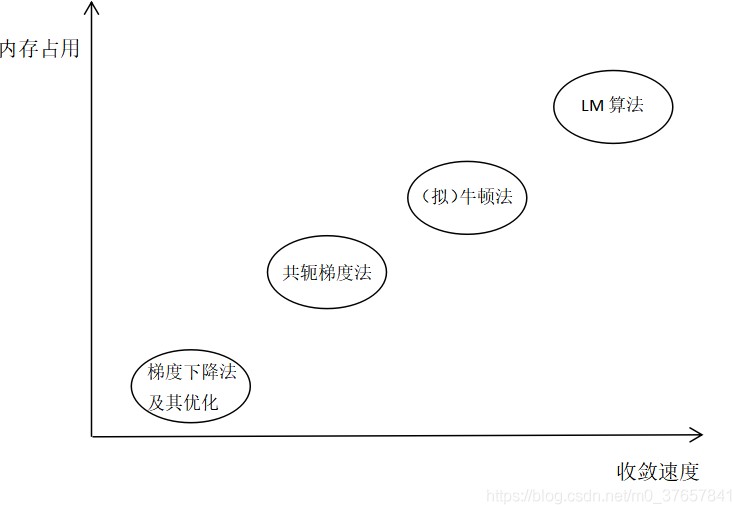


图4-1

## 几种训练方法的性能比较

如果我们的神经网络模型有上千个参数，则可以用节省存储的梯度下降法和共轭梯度法。如果我们需要训练很多网络模型，每个模型只有几千个训练数据和几百个参数，则Levenberg-Marquardt可能会是一个好选择。其余情况下，拟牛顿法的效果都不错。

# 参考文献

[1]张德丰.神经网络应用设计[M].北京:机械工业出版社，2009

[2] Rumelhart D.E.,Hinton G.E.,Williams R,J…Parall Distributed Processing-Explorations in the Microstructre of Congnition[J].MIT press,1986,page:318-362.

[3] 杨益君,古春生,朱广萍,蒋红芬,陈丹.LM神经网络在环境质量评估中的应用.江南大学学报.2010(4),9(2):213-216

[4] Iris Data Set.Famous Databose for Pattern Recognition from Fisher[OL].[http://archive.ics.uci](http://archive.ics.uci/). edu/ml/datasets/lris,2011.3.20.

[5] <https://baike.baidu.com/item/%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/16600562?fr=aladdin>

[6]龙坡,陈浩,何晶 & 何堃.().基于深度学习的心律模拟算法及其应用. *计算机应用研究*(), doi:10.19734/j.issn.1001-3695.2021.10.0420.

[7]高彤,陈鸿,王晋祺,李建鑫 & 柴世豪.(2021).基于LM算法的接触式轮廓扫描系统参数标定. *电子测量技术*(01),65-69. doi:10.19651/j.cnki.emt.2005334.

[8]潘富圣 & 黄世震.(2020).基于改进LM算法的可见光定位研究. *电子产品世界*(11),75-78. doi:CNKI:SUN:DZCS.0.2020-11-029.

[9]苏赫. 基于改进BP神经网络的空气质量预测研究[D].南昌大学,2020.DOI:10.27232/d.cnki.gnchu.2020.001491、

[10]杨嘉明. 基于LSTM-BP神经网络的列控车载设备故障诊断方法[D].北京交通大学,2018.

[11]姜雪杰. 基于改进BP神经网络驼峰减速器故障诊断研究[D].兰州交通大学,2017.

[12]王立政. BP神经网络在城市轨道交通客流短时预测中的应用研究[D].苏州科技大学,2017.