文章编号：2095-2163(2019)05-0138-04 中图分类号：TP391.41 文献标志码：A

基于决策树的 BP 神经网络权值初始化方法及其应用研究

艾 迪，董海峰
（西安石油大学 计算机学院，西安 710065）

摘 要： 为解决传统 BP 神经网络中随机初始化参数方法的缺陷，提出一种基于决策树信息增益算法的权值初始化方法。本文介绍了 C4.5 决策树算法和 BP 神经网络算法的主要特点，以及阐述了如何利用决策树算法中信息增益初始化 BP 神经网络权值参数，以避免传统随机初始化方法所造成的缺点。并以油气层敏感性评价的实例进行验证。实验表明，该初始化方法提高了 BP 神经网络的学习效率和准确度。

关键词： BP 神经网络；决策树；信息增益；权值初始化

BP neural network weight initialization method based on decision tree and its application

AI di, DONG Haifeng
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 In order to solve the defects of the traditional random initialization parameter method, a BP neural network weight initialization method based on decision tree information gain algorithm is proposed. This paper introduces the main features of the C4.5 decision tree algorithm and the BP neural network algorithm, and explains how to use the information gain in the decision tree algorithm to initialize the BP neural network weight parameters, which could avoid the large amount of trial and error caused by the traditional random initialization method. It is verified by an example of oil and gas layer sensitivity evaluation. Experiments show that the initialization method improves the learning efficiency and accuracy of BP neural network.

【Key words】 BP neural network; decision tree; information gain; weight initialization

0 引 言

BP 神经网络是人工神经网络中最常用的算法之一，主要通过误差反向传播来达到学习模式的目的。其具有较高的模式分类能力和多维度映射能力，在模式分类、模式识别、计算机视觉等领域应用较广。

但在实际情况中，易收敛到局部极小值与收敛速度慢等缺点使得 BP 神经网络往往无法达到期望的学习效率，而造成这些问题的一个主要原因是权值参数不合理的初始化[1]。若神经网络初始权值选择不当，有可能导致神经网络初始误差较大，收敛的方向较差，从而导致上述问题的出现。为此众多科研工作者进行了大量研究，关于 BP 神经网络权值的初始化方法主要有：利用正态分布函数生成的随机数的 BP 神经网络权值初始化方法[2]、利用遗传算法全局搜索最优初始权值的方法[3]、基于粒子群算法等人工智能算法优化初始权值的方法[4]等等。但由于这些初始化方法都在某种程度上具有随机性，所以在实际情况中依然存在一些局限。

本文提出基于 C4.5 决策树算法的 BP 神经网络权值初始化方法。根据 BP 神经网络与决策树的分类模式具有等价性[5]，以及决策树算法能够计算特征的对于样本的划分能力，可以在权值初始化时对划分能力较大的特征赋予更高的权值，从而避免初始权值不当造成的初始误差较大等问题。实验结果表明，与传统的随机初始化方法相较，该方法有效地减少了初始误差，提升了神经网络的训练精度。

1 C4.5 决策树算法概述

采用决策树算法来初始化 BP 神经网络权值的方法，主要是利用信息增益能够描述特征划分能力的特点，从而预估出各个神经元节点的权值大小[6]。且由于 ID3 决策树算法不支持连续型特征，以及对多值属性的偏向等缺陷，所以本文选用 C4.5 决策树算法作为权值初始化方法。针对连续数据的

作者简介：艾 迪(1992-)，男，硕士研究生，主要研究方向：管理信息系统；董海峰(1965-)，男，博士，副教授，硕士生导师，主要研究方向：信息管理系统、办公自动化、数据库开发等。

收稿日期：2019-05-13

样本集，C4.5 决策树算法训练步骤具体如下：

Step 1　对训练数据集中的连续数据离散化，缺失值补全。

Step 2　分别按照已处理的数据集的各个特征计算相应的信息增益、信息增益率。

Step 3　按照 C4.5 决策树算法，先找出信息增益高于平均水平的特征，再从中选择增益率最高的特征将当前数据集划分为不同的子集，建立相应的决策树。

Step 4　递归调用 Step 2、Step 3，直至所有特征都参与决策划分，并根据结果来建立完整决策树。

C4.5 决策树算法整体流程如图 1 所示。

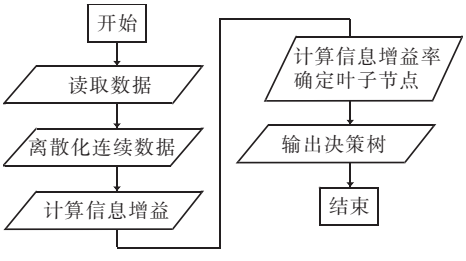

图 1　C4.5 决策树算法流程图

Fig. 1　Flow chart of C4.5 decision tree algorithm

2　BP 神经网络概述

理论上，单隐藏层的 BP 神经网络可以逼近任何有理函数，而随意增加隐藏层的数目可能导致网络结构更加复杂，进而增加神经网络的训练时间，降低训练效率[7]，因此，本文重点讨论单层隐藏层 BP 神经网络。该神经网络由输入层、隐藏层和输出层组成，主要学习过程可分为正向传播和反向传播两部分。其中，在正向传播时，样本数据传入输入层，经过正向传播至隐藏层，以同样方式再传播至输出层，并在输出端输出实际结果，神经网络的初始权值在正向传播的过程中不产生变化，若实际结果与期望值的误差没有达到标准，则误差由输出层开始进入反向传播过程。各层各个节点的权值通过梯度下降算法进行更新，直至将所有节点的权值更新完毕，再次进行正向传播。通过反复的正向传播和反向传播使得各个节点的权值得到不断的修正，这 2 个过程一直循环交替进行，直到误差减少至设定的范围内，或者达到迭代次数为止。神经网络结构如图 2 所示。

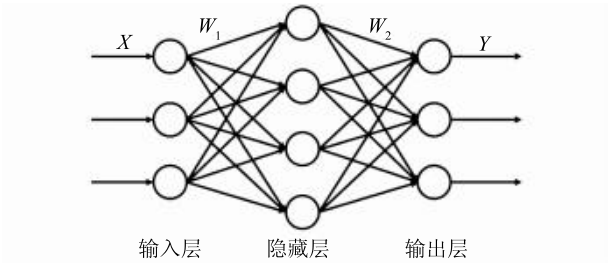

图 2　3 层 BP 神经网络结构图

Fig. 2　Structure of three-layer BP neural network

3　基于决策树算法信息增益的权值初始化方法

基于决策树算法信息增益的权值初始化方法，其基本思想是在初始化输入层与隐藏层的连接权值时，通过 C4.5 决策树算法对样本集进行分析，利用特征对输出特征的划分能力来确定输入初始权值的大小，其中特征的划分能力将主要体现在特征的信息增益上。基于决策树的权值初始化方法建立的算法步骤详见如下。

Step 1　假定当前有样本集 D，样本个数为 $|D|$，由决策特征集 $S=\{S_1, S_2, \cdots, S_n\}$ 将样本分为 n 个子集。其中 S_j 表示第 j 个子集，假设该集合中元素个数为 $|S_j|$。则样本集 D 的信息熵为：

$$Ent(D) = -\sum_{j=1}^{n} p_j \log_2 p_j \ , \qquad (1)$$

其中，P_j 为分类 j 出现的概率，用 $P_j = |S_j| / |D|$ 计算。

假设特征 A 可将样本集划分为 T 个分支节点，其中第 t 个分支节点包含了 D 中所有在特征 A 上取值为 a_t 的样本，记为 D^t，根据公式(1)，计算 D^t 的信息熵 $Ent(D^t)$。在此基础上，计算特征 A 对样本集 D 的信息增益，用来表示特征对样本的划分能力。其数学计算公式可表示为：

$$Gain(D,A) = Ent(D) - \sum_{t=1}^{T} \frac{|D^t|}{|D|} Ent(D^t) \ , \qquad (2)$$

由于直接使用信息增益划分数据集对可取值数目较多的特征有所偏向，进一步引入增益率来选择最优划分特征，并对无效的多值属性进行清洗。采用通过 4.5 决策树算法构建的决策树模型如图 3 所示。

Step 2　根据样本集 D，假设 BP 神经网络的输入层为 m 个节点，对应决策树中特征子集，隐藏层为 n 个节点，输出层为 k 个节点，对应决策树中决策特征。其中，w_{ij} 表示第 i 个节点到隐藏层第 j 个节点的权重 。BP 神经网络部分结构如图 4 所示。

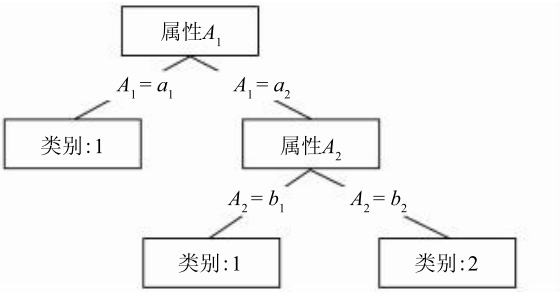

图3 决策树分类结果

Fig. 3 Decision tree classification results

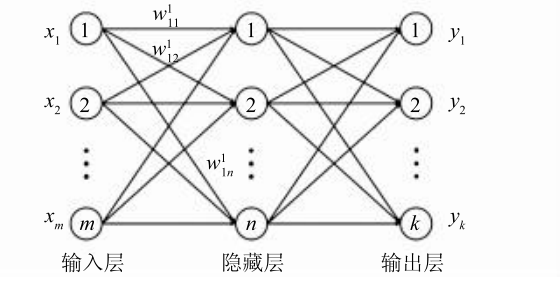

图4 BP神经网络部分结构

Fig. 4 Part structrue of BP neural network

Step 3 样本集经过特征A的划分,成为2个子集n_1和$(n-n_1)$,n_1为样本数量最大的子集元素个数,则在BP神经网络中输入层初始权值为:

$$w_i = Gain(D,a) * \frac{n_1}{n}, \tag{3}$$

其中,$Gain(D,A)$为特征A的信息增益值,初始化w_{ij}的计算方式为:

$$w_{ij} = w_i + \frac{1}{\sqrt{2\pi}} e^{-\frac{(random(-\infty,+\infty))^2}{2}}. \tag{4}$$

Step 4 进行循环训练,更新权值。

Step 5 误差达到标准,训练结束,输出模型。

4 油气层敏感性评价实例分析

本文原始数据来源于某油田岩石实验,其中油气层水敏性主要影响因素为:泥质含量、石英含量、蒙脱石含量、伊-蒙含量、胶结物总量、粒度中值、分选系数、孔隙度、渗透率、总矿化度[8]。针对油气层敏感性中水敏指数进行评价,部分训练样本见表1。

表1 水敏感性评价源数据

Tab. 1 Source data of water sensitivity evaluation

序号	泥质含量/%	石英含量/%	蒙脱石含量/%	伊-蒙含量/%	胶结物总量/%	粒度均值/mm	分选系数	孔隙度/%	渗透率/md	总矿化度/(g/L)
1	8.06	34.11	0.20	6.09	8.58	0.12	12.68	12.82	43.15	54.15
2	8.14	40.41	0.30	7.09	7.84	0.10	11.68	10.72	30.18	90.15
3	8.11	33.16	0.31	7.09	8.98	0.18	11.69	10.62	30.19	19.09
4	10.55	40.36	0.12	7.09	8.90	0.10	10.68	10.56	30.92	51.69
5	8.12	36.39	0.20	7.09	8.93	0.21	11.65	10.86	41.09	32.15
6	8.31	40.32	0.30	7.09	8.95	0.10	11.72	11.00	22.35	32.19
7	9.01	40.28	0.12	7.09	8.26	0.10	12.70	11.01	18.38	84.10
8	8.66	42.56	0.12	7.09	8.86	0.10	11.73	13.87	17.97	35.60
9	6.18	44.41	0.12	7.75	9.32	0.10	12.70	11.84	23.25	16.67
10	5.11	38.94	0.30	7.09	8.32	0.10	11.72	10.81	6.31	41.76

对训练样本集进行数据离散化处理,使用随机化算法打乱样本集顺序后,选取处理后样本数量的75%为训练样本,剩余25%为测试样本。计算训练样本中特征的信息增益与信息增益率,构建C4.5决策树模型。再将训练样本进行数据归一化作为BP神经网络的输入层,利用信息增益对随机初始化的权值进行加权处理,阈值初始化为0。将处理后的权值导入神经网络进行训练,误差达到标准后,输出评价模型,使用测试样本检测评价模型的准确率。采用完全相同的网络结构和初始阈值构建BP神经网络,使用基于随机初始化方法初始化权值。以相同的数据集进行迭代训练,并对比各实验结果。平均实验结果见表2。

表2 油气层敏感性评价结果对比

Tab. 2 Comparison of reservoir sensitivity evaluation results

初始化方法	初始误差	分类精度	训练时间/s
随机初始化方法	3.06	0.848	9.56
基于决策树的初始化方法	2.36	0.866	8.90

从表2可以看出,基于决策树的初始化方法平均初始误差为2.36,低于随机初始化方法的初始误差;训练完成后,分类精度提升了1.8%,提升幅度

不大，同时训练时间有略微下降。实验过程中，对基于决策树算法的权值初始化方法与随机初始化方法的训练误差进行对比，对比绘制结果如图 5 所示。

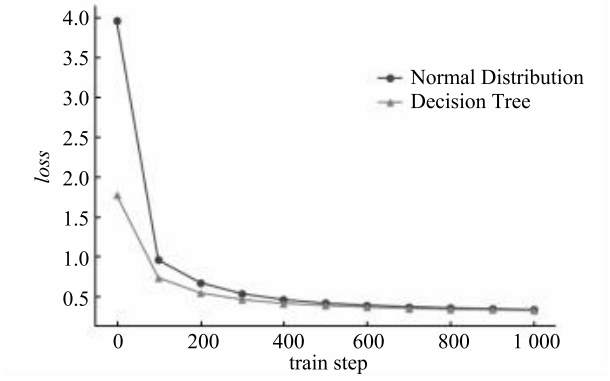

图 5　BP 网络误差曲线对比

Fig. 5　BP neural network error curve comparison

由图 5 的训练误差曲线可以看出，基于决策树算法的权值初始化方法的初始训练误差小于传统随机初始化方法，并优先到达误差最小值。

在此基础上，研究得出结论，基于决策树算法的 BP 神经网络权值初始化方法优于随机初始化方法，主要表现在初始误差相对较小，模型的分类精度得到提高。同时，研究表明该方法增强了 BP 神经网络的收敛能力，提高了学习的速度。

5　结束语

本文提出基于决策树算法的 BP 神经网络权值初始化方法，并与传统随机初始化方法在油气层敏感性评价实验中进行了对比。结果表明，以该方法初始化权值的 BP 神经网络初始误差相对较小，评价准确率更高，具有一定的实用性。同时，对于本文方法的研究优越性，则可概述为：

（1）决策树算法可以评估特征的区分能力，使得初始权值更为合理。

（2）大概率避开了局部极小值点。

参考文献

[1] 刘鲭洁，陈桂明，刘小方，等. BP 神经网络权重和阈值初始化方法研究[J]. 西南师范大学学报（自然科学版），2010，35（6）：137-141.

[2] 范业仙，叶茂枝. BP 神经网络初始化方法研究[J]. 韶关学院学报，2013，34（12）：18-21.

[3] 墨蒙，赵龙章，龚媛雯，等. 基于遗传算法优化的 BP 神经网络研究应用[J]. 现代电子技术，2018，41（9）：41-44.

[4] 冯非凡，武雪玲，牛瑞卿，等. 粒子群优化 BP 神经网络的滑坡敏感性评价[J]. 测绘科学，2017，42（10）：170-175.

[5] 李爱军，罗四维，刘蕴辉，等. 基于熵准则的神经网络设计方法（英文）[J]. 复旦学报（自然科学版），2004，43（5）：721-724，728.

[6] 苗煜飞，张霄宏. 决策树 C4.5 算法的优化与应用[J]. 计算机工程与应用，2015，51（13）：255-258，270.

[7] 焦斌，叶明星. BP 神经网络隐层单元数确定方法[J]. 上海电机学院学报，2013，16（3）：113-116，124.

[8] 樊世忠，鄂捷年，周大晨，等. 钻井液完井液及保护油气层技术[M]. 东营：石油大学出版社，1996.

（上接第 137 页）

设计教学方案，切实提高课堂教学质量，从而达到预期教学效果。

参考文献

[1] 徐琳，严淑斐，史利涛，等. 两岸高校学生课堂行为比较研究—以东北师范大学与台湾彰化师范大学为例[J]. 文教资料，2014（15）：157-159.

[2] 胡卫星，赵苗苗. 多媒体教学过程中学生学习行为的实验研究[J]. 中小学电教，2005（11）：50-51.

[3] 崔允漷. 有效教学[M]. 上海：华东师范大学出版社，2009.

[4] 周敏. 初中小班化教育背景下学生课堂问题行为的实证研究[D]. 宁波：宁波大学，2013.

[5] 陈霞. 语文课堂学生学习行为的研究[D]. 桂林：广西师范大学，2007.

[6] 赵庆红，徐锦芬. 大学英语课堂环境与学生课堂行为的关系研究[J]. 外语与外语教学，2012（4）：66-74.

[7] 王会廷，张艳平，阎慧. 大学生课堂行为的心理学研究[J]. 安徽工业大学学报（社会科学版），2012，29（3）：37-38.