用遗传算法改进的BP神经网络剪枝算法来优化决策树模型

武 彤 程 辉
(贵州大学计算机科学与信息学院 贵阳 550025)

摘 要 决策树是一种有效的分类方法,但在构建决策树模型的过程中,常常会出现模型过度拟合的现象。利用基于BP神经网络的决策树剪枝算法(BP-Pruning)进行软剪枝处理,然后根据BP-Pruning的一些不足,提出一种改进算法,简称GBP-Pruning算法。该算法通过引入遗传算法来训练BP-Pruning算法模型中的权值和阈值,从而克服了BP-Pruning算法上的不足,最后验证了GBP-Pruning算法的可行性。

关键词 数据挖掘,决策树,BP神经网络,遗传算法,剪枝算法
中图法分类号 TP39 **文献标识码** A

BP Neural Network Pruning Algorithm Improved on Base of Genetic Algorithm to Optimize Decision Tree Model

WU Tong CHENG Hui
(School of Computer Science and Information, Guizhou University, Guiyang 550025, China)

Abstract Decision is an effective classification method. But during the building process of decision tree, there usually appear over-fitting phenomena of models. This paper discussed soft pruning processing by using BP pruning which is based on BP neural network. Then, according to the shortages of BP pruning, this paper proposed a revised algorithm, named GBP-Pruning. This algorithm is able to train weight and threshold value of BP-Pruning model by bringing in genetic algorithm, so that it can overcome the shortages of BP-Pruning. It also proved the feasibility of GBP-Pruning.

Keywords Data mining, Decision-tree, BP neural network, Genetic algorithm, Pruning algorithm

决策树是一种有效的分类方法,对于多峰分布之类的问题尤为方便。利用决策树采用分级的形式,可以把一个复杂的多类别分类问题转化为若干个简单的分类问题来解决[1]。但是在建立决策树模型时会出现如下问题。

给定一个空间 H,一个假设 $h\in H$,如果存在其他的假设 $h'\in H$,使得在训练样例上 h 的错误率比 h' 小,但在整个实例分布上 h' 的错误率比 h 小,那么就称假设 h 为过度拟合训练数据[2]。其现象表现为一个假设在训练数据上能够获得比其他假设更好的拟合,但是在训练数据外的数据集上却不能很好地拟合数据。如图1所示,在决策树创建过程中,横轴表示结点总数,纵轴表示决策树做出的预测的精度。实线表示决策树算法在训练集上的精度,虚线表示决策树算法在测试集上的精度。从图中可以看出,随结点总数的增加,在训练样例上的精度上升,但是在测试集上的精度下降。当结点总是小于16时,决策树在训练集与测试集的精度都上升。当树的结点总数大于16时,其在训练集上的精度上升而测试集上的精度会下降,从而产生预测不准确的情况。

在决策树构建过程中出现上述现象的主要原因是决策树生成算法中,结点的分割标准是一直到该结点中所包含的全部实例属于同一类别时才停止划分。从而使得有些结点可能被分割为单一实例的叶节点。这种划分虽然分类比较完全,但会产生过多的冗余无用的规则,不利于预测。再加上训练集数据中存在噪音或者训练数据样例太少,以致于不能产生目标函数具有代表性的采用。

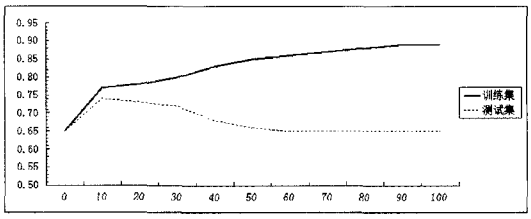

图1 决策树过度拟合现象示意图

为了防止决策树模型过度拟合现象的出现,本文研究采用基于遗传算法改进的BP神经网络剪枝算法对决策树模型进行优化处理。

1 基于BP神经网络的决策树剪枝算法

决策树剪枝操作是优化决策树模型的有效手段。它可以防止决策树模型出现过度拟合现象。传统的决策树剪枝算法是在其树结构的基础上根据某些标准裁剪一些节点,从而减小模型规模,防止过度拟合现象,最终生成剪枝后的决策树规

本文受贵州省2010年工业攻关项目:生产线质量控制决策支持系统的开发研究(黔科合GY字[2010]3061)资助。
武 彤(1964—),女,硕士,教授,CCF会员,主要研究方向为数据仓库技术、OLAP、数据挖掘。

则，以提高模型的预测精度。但是这些方法在提升模型预测精度方面还存在着一些不足。因为某些节点可能属于修剪算法应当裁剪掉的范围，但从全局角度考虑，这些节点又存在着重要的分类预测意义，对模型预测起到一些关键的作用。最终随着节点被裁剪掉，模型的预测分类精度会受到影响。

本文以“基于决策树算法的电视机生产故障维修模型”为研究的实验模型。采用后剪枝算法-BP 神经网络的决策树剪枝算法对其实验模型进行修剪，即允许决策树过度拟合数据。

预测电视机故障模型中，我们通过实验得出 If-Then 规则，然后将其编码，利用基于 BP 神经网络算法的决策树剪枝算法对其规则进一步修剪，得出更加有效的、更加利于决策的规则。反向传播神经网络剪枝算法克服了传统剪枝算法的不足。该算法通过反向传播神经网络算法对其结点进行训练，得出其结点的权值；再将这些权值赋给这些结点，这些权值就代表着该结点在其决策树模型中的权重值；再结合结点的权重值参数来对其模型进行剪枝操作；这将大幅度地提高模型预测准确度[3]。

BP 神经网络的决策树剪枝算法包括 3 个步骤：

步骤 1　利用决策树算法 C4.5 构建决策树模型。每一条 If-Then 规则都是来自决策树模型根结点到一个叶结点的路径。从根结点到该路径上的最后一个非叶结点称为该规则的前驱条件，放在 If 之后 Then 之前。该路径上的叶结点作为规则结果放在 Then 之后。其规则集合如表 1 所列。

步骤 2　根据表 1 所列的 If-Then 规则表，构建三层神经网络。将前驱条件作为输入层，与其隐藏层相联系。其隐藏层的神经元的个数 $N2$ 与输入层神经元的个数 $N1$ 存在 $N2=N1*2+1$ 的关系。输出层神经元的个数与规则集合 If-Then 中的规则结果的个数相同。在此模型中输出层神经元个数为 13，其中中间层的全部神经元与每个输出层神经元相联系。部分结构如图 2 所示。

步骤 3　训练网络结构模型。训练阶段，首先随机初始化神经网络中的权值和阈值，并在训练过程中通过反馈神经网络模型反馈过来的参数，调整权值大小。使用单极性 Sigmoid 函数作为转移函数。在训练过程中每个输出神经元都参与训练。该模型通过使用训练集合重复地进行训练直到收敛或者训练迭代次数超过预先设定的临界值。

表 1　决策树模型生成的规则集合

(1) If Fault Appearance=‘无光’ Then FaultType=CRT 不良
(2)IF Producttype=‘平板产品’ and Fault Appearance=‘无遥控’ Then FaultType= SMT 不良
(3)IF Producttype=‘平板产品’ and Fault Appearance=‘AV 无输出’ Then FaultType= SMT 不良
(4)IF Producttype=‘普通产品’ and Fault Appearance=‘行扭’ Then FaultType=插件不良
(5)IF Producttype=‘平板产品’ and Fault Appearance=‘灯不亮’ Then FaultType=插件不良
(6)IF Producttype=‘平板产品’ and Fault Appearance=‘节目键无作用’ Then FaultType=插件不良
(7)IF Producttype=‘平板产品’ and Fault Appearance=‘节能无用’ Then FaultType=插件不良
(8)IF Producttype=‘平板产品’ and Fault Appearance=‘频率低’ Then FaultType=调试不良
(9)If Producttype=‘普通产品’ and Fault Appearance=‘左喇叭无伴音’ Then FaultType= 焊接不良
(10)IF Producttype=‘平板产品’ and Fault Appearance=‘部份按键无作用’ Then FaultType= 焊接不良
(11)IF Producttype=‘平板产品’ and Fault Appearance=‘节能无作用’ Then FaultType= 焊接不良
(12)IF Producttype=‘平板产品’ and Fault Appearance=‘图暗’ Then FaultType= 焊接不良
(13)IF Producttype=‘平板产品’ and Fault Appearance=‘VGA 图异’ Then FaultType= 焊接不良
(14)IF Producttype=‘平板产品’and Fault Appearance=‘待机键无作用’ Then FaultType= 焊接不良
(15)IF Producttype=‘外观’ and Fault Appearance=‘机震’ Then FaultType=结构不良
(16)IF Producttype=‘平板产品’ and Fault Appearance=‘菜单键无作用’ Then FaultType=结构不良
(17)IF Producttype=‘平板产品’ and Fault Appearance=‘机震’ Then FaultType=结构不良
(18)IF Producttype=‘平板产品’ and Fault Appearance=‘黑屏’ Then FaultType=屏幕不良
(19)IF Producttype=‘平板产品’ and Fault Appearance=‘屏异常’ Then FaultType=屏幕不良
(20)IF Producttype=‘普通产品’ and Fault Appearance=‘接触’ Then FaultType=设计不良
(21)IF Producttype=‘平板产品’ and Fault Appearance=‘图异’ Then FaultType=设计不良
(22)IF Producttype=‘新品’ and Fault Appearance=‘图异’ Then FaultType=设计不良
(23)IF Producttype=‘平板产品’ and Fault Appearance=‘单伴音’ Then FaultType=设计不良
(24)IF Producttype=‘平板产品’ and Fault Appearance=‘按键无作用’ Then FaultType=元器件不良
(25)IF Producttype=‘平板产品’ ’and Fault Appearance=‘AV 伴音异常’ Then FaultType=元器件不良
(26)IF Producttype=‘平板产品’ and Fault Appearance=‘无伴音’ Then FaultType=元器件不良
(27)IF Producttype=‘平板产品’ and Fault Appearance=‘伴音异常’ Then FaultType=元器件不良
(28)IF Producttype=‘平板产品’ and Fault Appearance=‘左无伴音’ Then FaultType=元器件不良
(29)IF Producttype=‘平板产品’ and Fault Appearance=‘按键不良’ Then FaultType=元器件不良
(30)IF Producttype=‘平板产品’ and Fault Appearance=‘按键不良’ Then FaultType=元器件不良
(31)IF Producttype=‘平板产品’ and Fault Appearance=‘强信号’ Then FaultType=元器件不良
(32)IF Producttype=‘新品’ and Fault Appearance=‘遥控无作用’ Then FaultType=装配不良
(33)IF Producttype=‘平板产品’ and Fault Appearance=‘遥控无作用’ Then FaultType=装配不良
(34)IF Producttype=‘平板产品’ and Fault Appearance=‘图闪’ Then FaultType=装配不良
(35)IF Producttype=‘平板产品’ and Fault Appearance=‘无信号’ Then FaultType=主键装配不良

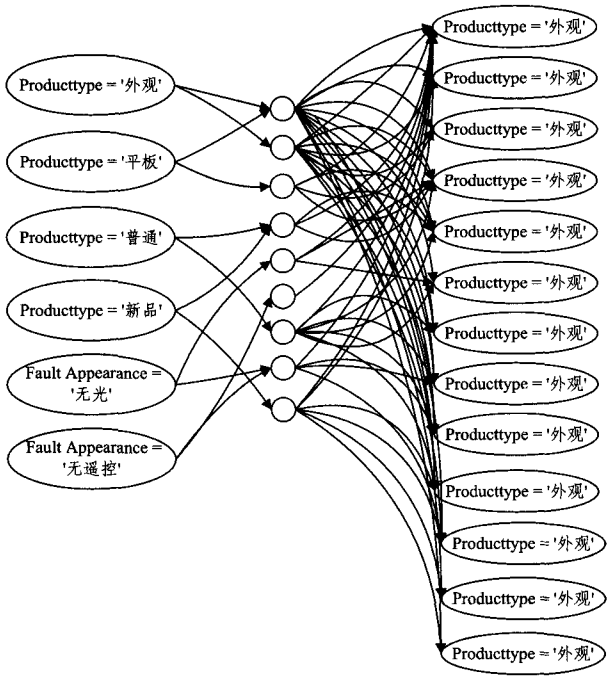

图 2　部分 BP 神经网络模型示意图

2　基于遗传算法改进的 BP 神经网络剪枝算法

根据上节所述，我们使用 BP-pruning 算法取代传统剪枝算法对其实验决策树模型进行裁剪，以提高决策树模型预测精度。但该算法本身也存在着一些不足，比如学习收敛速度太慢、不能保证收敛到全局最小点。另外初始连接权值和阈值是随机选定，但权值和阈值对其网络训练的影响很大，可能会影响最后模型的实验结果。由于遗传算法具有全局搜索能力，因此采用遗传算法对 BP 神经网络剪枝算法进行优化处理。改进模型体系结构如图 3 所示。

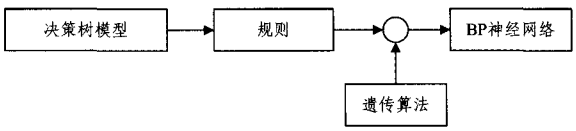

图 3　改进模型体系结构图

2.1　算法思路及其流程

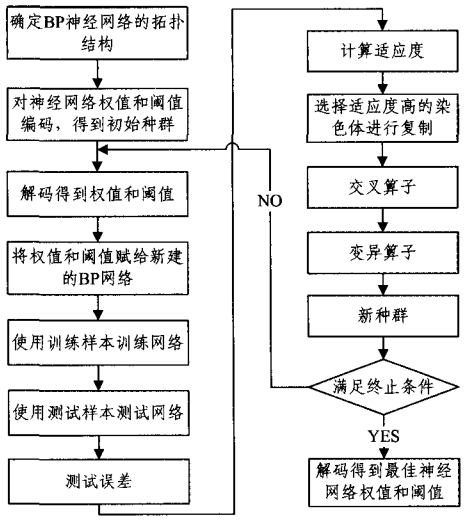

图 4　算法流程图

基于遗传算法优化的 BP 神经网络剪枝算法主要包括 BP 神经网络结构确定、遗传算法优化权值和阈值、BP 神经网络训练及预测 3 个部分。其中 BP 神经网络的拓扑结构根据上节所述的步骤进行构造，当网络模型确定以后，其输入/输出参数个数也确定完成。这样就可以确定遗传算法优化参数的个数，从而确定种群个数的编码长度。引入遗传算法就是为了优化最佳的初始权值和阈值[4]。其算法流程如图 4 所示。

2.2　算法实现

基于遗传算法优化的 BP 神经网络剪枝算法（简称 GBP-Pruning 算法）的实现主要包括两大部分：BP-Pruning 算法的实现以及遗传算法的实现。

首先，对于 BP-Pruning 算法根据上节所述的构建步骤，确定 BP 神经网络结构为 8-17-13。即输入层有 8 个节点，隐含层有 17 个节点，输出层有 13 个节点，共有 $8\times17+17\times13=357$ 个权值，$17+13=30$ 个阈值，遗传算法优化参数的个数为 $357+30=387$。将测试样本的测试误差的范数作为衡量 BP 网络的一个泛化能力，再通过误差范数计算个体的适应度值，个体的误差范数越小，适应度值越大，该个体越优。在神经网络的传递函数的选择方面，隐含层神经元和输出层神经元的传递函数采用 S 型对数函数 logsig()，这是由于输入输出编码模式为 0/1 所确定，正好满足网络的输入输出要求。

BP 神经网络训练和测试阶段是一个不断修正权值和阈值的过程，通过训练使得网络的输出误差越来越小。训练函数采用 Levenberg-mar quardt 算法对网络进行训练。

最后，遗传算法实现阶段是利用遗传算法来优化 BP 神经网络的初始权值和阈值，使优化后的 BP 神经网络剪枝算法具有更好的剪枝精确度。利用遗传算法来优化 BP 神经网络剪枝算法的要素包括初始化编码、适应度函数选择、选择算子、交叉算子和变异算子的选择。下面分别对其简要介绍。

(1)初始化编码阶段。该阶段包括前期神经网络输入输出段编码和优化阶段种群初始化编码。对于 BP 神经网络输入端编码，结合表 1 所列的规则集合，将其规则集合中每条规则首先进行数值化，即将离散量转化为数值量。利用二进制编码方法进行编码。例如：Producttype 属性值为“平板”编码成为 000，属性 FaultAppearance 值为“无光故障”编码为 00001。即编码 00000001 表示规则前驱条件为：Producttype =“平板”∧FaultAppearance=“无光故障”。综合全部规则前驱条件可知使用 8 位二进制进行编码。输出端由于规则的结论有 13 种类别，故采用 13 位二进制编码进行编码。每一位表示一种类别，值为“1”表示属于此类别，“0”表示不属于此类别。BP-Pruning 算法的模型为 8-17-13，针对种群初始化编码阶段，也采用二进制编码，每个个体均为一个二进制串，由输入层与隐含层连接权值、隐含层阈值、隐含层与输出层连接权值、输出层阈值 4 个部分组成。每个权值和阈值使用 M 位的二进制编码，将所有权值和阈值的编码连接起来即为一个个体的编码。

(2)适应度函数选择阶段。为了使 BP 网络在预测时，预测值与期望值的残差尽量地小，选择预测样本的预测值与期望值的误差矩阵的范数作为目标函数的输出。适应度函数采用排序的适应度分配函数。

(3)遗传算法选择阶段。对于选择算子的选择，采用随机

（下转第 295 页）

参 考 文 献

[1] 蓝章礼,等. 数字图像处理与图像通信[M]. 北京:清华大学出版社,2009:157-162

[2] 常娜. 图像处理中的边缘检测算法研究综述[J]. 中国科技信息,2011,4:131-149

[3] 高朝阳,等. 图像边缘检测研究进展[J]. 科学导报,2010,28(20):112-117

[4] 李杰,等. 基于数学形态学的图像边缘检测算法的研究[J]. 计算机科学,2012,6:546-548

[5] 侯宝生. 一种基于数学形态学的图像边缘检测方法[J]. 计算机应用技术,2010,8:93-96

[6] 陈恩庆,等. 采用多结构元素模板的形态学边缘检测新算法[J]. 计算机工程与应用,2012

[7] 孙继平,吴冰,刘晓阳. 基于膨胀/腐蚀运算的神经网络图像预处理方法及其应用研究[J]. 计算机学报,2005,28(6):985-990

[8] Roerdink J B T M. Adaptivity and Group Invariance in Mathematical Morphology[C]//IEEE ICIP 2009. 2009:2253-2256

[9] Bouaynaya N, Charif-Chefchaouni M, Schonfeld D. Spatially variant morphological restoration and skeleton representation[J]. IEEE Transactions on Image Processing, 2006, 15(11): 3579-3591

[10] Bouaynaya N, Charif-Chefchaouni M, Schonfeld D. Theoretical foundations of Spatially-Variant mathematical morphology-Part I: Binary images[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2008, 30(5): 823-836

[11] Bouaynaya N, Charif-Chefchaouni M, Schonfeld D. Theoretical foundations of Spatially-Variant mathematical morphology-Part II: Gray-level images[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2008, 30(5): 837-850

(上接第 280 页)

遍历抽样的方法。交叉算子采用基本的单点交叉算子方法。对于变异算子,则使用随机方法选出发生变异基因。如果所选的基因的编码为 1,则变为 0;反之,则变为 1。

3 改进对比实验结果及分析

根据第 2 节所介绍的改进的 GBP-Pruning 剪枝模型理论,利用 sheffield 遗传算法工具箱,在 MATLAB 软件中编程,实现基于遗传算法优化的 BP 神经网络剪枝算法。通过仿真对比剪枝模型算法改进前后的误差值,证明改进的 GBP-Pruning 算法对其剪枝模型的预测精确度有所提高。实现结果如图 5、图 6 所示。

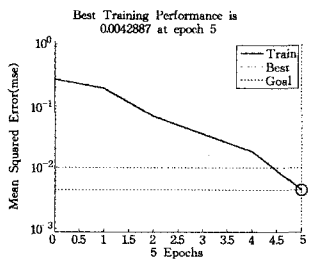

图 5 随机权值和阈值训练误差曲线

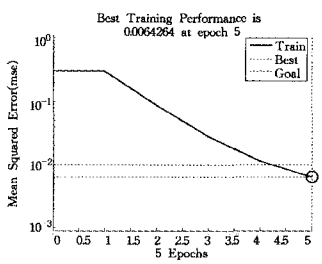

图 6 优化权值和阈值训练误差曲线

如图 5、图 6 所示,纵坐标为均方误差,横坐标为迭代次数,可以看出图 5 在迭代次数为 4.4 时达到实验最优结果,图 6 在迭代次数为 4.1 时达到实验最优结果。两次实验的模型误差值如下:

1. 使用随机权值和阈值

测试样本预测结果:

测试样本的仿真误差:0.67067

训练样本的仿真误差:0.90895

Warning:NEWFF used in an obsolete way.

〉In obs_use at 18

In newff 〉create_network at 127

In newff at 102

In callbackfun at 29

See help for NEWFF to update ca

2. 使用优化化的权值和阈值

测试样本预测结果:

测试样本的仿真误差:0.37476

训练样本的仿真误差:0.6371

通过比较可以看出,未改进的 BP 神经网络剪枝算法模型的测试数据样本的误差值为 0.67,改进后的 GBP-Pruning 算法剪枝模型的测试数据样本的误差值为 0.37。通过实验仿真结果可以看出,GBP-Pruning 算法在一定程度上比原来的 BP 神经网络剪枝算法在模型预测精确度上有所提高,即改进的算法达到了实验的预期结果,提高了模型的预测精度[5]。

结束语 本文研究的基于遗传算法改进的 BP 神经网络剪枝算法,是在已经构建的决策树模型的基础上,为了提高模型预测准确性,采用先进的称为软剪枝算法的 BP 神经网络决策树剪枝算法对其模型进行后剪枝处理,并在此算法基础上引入遗传算法对其进行改进,优化其权值,从而进一步提高挖掘模型的预测精度,防止过度拟合现象。通过实验已经证明了该算法的正确性。下一步就经过 GBP-Pruning 算法优化过的数据挖掘模型构建时间过长的问题进行进一步的研究,以使其应用于实际的工业产品生产线上。

参 考 文 献

[1] 王丽珍,周丽华,等. 数据仓库与数据挖掘原理及应用[M]. 北京:科学出版社,2005

[2] 邵峰晶,于忠清,王金龙,等. 数据挖掘原理与算法[M]. 北京:科学出版社. 2009

[3] 魏红宁. 决策树剪枝方法的比较[J]. 西南交通大学学报,2005,2(40):44-48

[4] 王小平,曹立明. 遗传算法-理论、应用于软件实现[M]. 西安:西安交通大学出版社,2002

[5] 程辉. 决策树算法在生产线质量控制决策系统的应用研究[D]. 贵阳:贵州大学,2013