文章编号：2095-2163(2023)05-0161-05 中图分类号：TP183 文献标志码：A

基于模型组合的网络迭代识别法

丁霖泽，葸娟霞
（广东东软学院 信息管理与工程学院，广东 佛山 528200）

摘 要：针对当前常用的多层神经网络优化算法只能获得局部极值，算法性能与初始值有关，以及网络获得的参数与实际任务的关联性非常模糊等问题，提出了一种基于模型组合的网络迭代法（LDKB），以较小速度损失，实现模型间的有效联合，强化变量间的关联性，提升整体精度。LDKB 引用 PDCA 循环工作法的概念，以完成模型间的自动迭代。同时，在训练模型中数据流正向传播过程中，用输出信息比较剔除方法，剔除非重要 BP 神经网络的学习信息，以提高循环过程中剪枝容错性。结果表明，相比单个决策树模型，LDKB 的影响识别精度持续提高；相对单个神经网络模型，LDKB 模型的识别性保持最佳。

关键词：特征分箱；WOE 编码；模型组合；数据流循环

Network iterative identification method based on model combination

DING Linze, XI Juanxia
(Information Management and Engineering College, Guangdong Neusoft University, Foshan Guangdong 528200, China)

【Abstract】 The current commonly used multilayer neural network optimization algorithm can only obtain the local extreme value, and the algorithm performance is related to the initial value.At the same time, the correlation between the parameters obtained by the network and the actual task is very fuzzy. A network iteration method (LDKB) based on model combination is proposed, so as to realize the relationship combination between models at a small speed, and realize the improvement of model accuracy and correlation output enhancement. The LDKB method references the concept of the PDCA cycle working method to complete the automatic iteration between the models.At the same time, in the forward propagation process of data flow in the training model, the output information comparison and elimination method is used to eliminate the learning information of non-important BP neural network, so as to improve the pruning fault tolerance in the cycle process.The results show that compared with a single decision tree model, LDKB continuously improves the recognition accuracy of LDKB model compared with the single neural network model.

【Key words】 feature bin; WOE coding; model combination; data flow cycle

0 引 言

当前神经网络的应用已涉及到各个领域[1]，在智能控制 、模式识别 、计算机识别[2]等方面取得了长足的发展[3]BP 算法具有非线性转移函数的三层前馈网络，体现了人工神经系统的最精华部分。

神经网络在实际问题应用中，受自身梯度下降、步长规则等因素影响，在处理时会遇到 4 种常见问题[4]：

（1）由于实际问题规模往往很大，因此理论上需要神经网络与其相匹配，而网络过大，将极大降低网络的推广能力，不能发现其合理规则，从问题中选取典型实例组合是困难的。

（2）神经网络在执行梯度下降时，其所需的最小化目标过于复杂，因此必然会出现“锯齿形现象”[4]。

（3）运行中往往可能陷入“局部最优陷阱”从而无法达到学习目的。

（4）在 BP 网络用于识别时，自身算法往往存在收敛速度缓慢、网络性能较差、误差平方和函数可能有局部极小点出现的可能性，以及学习率不稳定的问题[5]。

综上所述，如采用自适应学习速率等不依赖梯度信息[6]，但是收敛精度可能不高[7]；采用进化算法，进行优化计算来确定，对大规模 FNN，工作量大，耗时过长且无法保证效果；采用删除冗余样本信息的特征样本[8]，验证样本误差的下降趋势决定何时结束训练，一旦出现部分数据与目标关系较小，将会导致完全删除。

针对单模型容易丢弃或无视数据潜在的影响因素，本文提出基于模型组合的网络迭代法（LDKB），

作者简介：丁霖泽（2001-），男，本科生，主要研究方向：机器学习；葸娟霞（1992-），女，硕士，助教，主要研究方向：机器学习、智能信息处理。

收稿日期：2022-06-09

该方法基于多模型建立数据流循环模型,在保证主题结构完整的前提下,有效减少单模型的局限性,实现数据精确化展现。LDKB 法通过数据流循环过程,迭代模型组合,使其在现有机器学习模型下,通过模型优化数据,获得较高精度结果,多次迭代后模型组合策略为总体精度提升带来积极影响。

1 LDKB 算法实现

1.1 模型组合流程参数计算

LDKB 算法的组合流程如图 1 所示,其实现步骤如下:

(1)首次循环进入决策树部分(剪枝详见 1.4 节)分别使用 Enter、Forward、Remove、Backward、Stepwise 5 种方法筛选变量优化回归模型,并利用 Chi-Square、-2 Log likelihood、BIC、AIC 作为衡量标准确立回归部分的最优参数[9]。

(2)以 0.1 为基础,引用函数"cv"在每一次迭代中使用交叉验证,并返回理想的树数量,利用 bagging 算法降低泛化误差,计算基尼系数分割父子节点,获取子节点的计算反馈数据模型迭代前最优化树。

(3)选取目标中的一个样本点作为第一个聚类中心,计算每个样本点与当前已有聚类中心的最短距离,即

$$D(x^i) = \min[dist(x^i,\mu_1),\cdots,dist(x^i,\mu_n)] \quad (1)$$

则样本点被选为下一个簇中心的概率为

$$\frac{D(x^i)^2}{\sum D(x^j)^2} \quad (2)$$

(4)在神经网络阶段,确定隐藏层中采用 tanh 函数作为激活函数,计算隐藏层神经元个数为

$$h = \frac{s}{c(n+m)} \in [2,10] \quad (3)$$

数据中共有 11 个字段,分别计算其权重 w 与偏差 b 相对于损失的梯度,所有字段的权重 w 总和与偏差 b 的总和为:

$$w_{i+1} = w_i - \alpha * \frac{dL}{dw_i} \quad (4)$$

$$b_{i+1} = b_i - \alpha * \frac{dL}{db_i} \quad (5)$$

(5)更新系统数据,以融合得到新的类别与数据。

(6)重复执行步骤(2)~步骤(5)直至循环的第 N 次输出模型效果小于 $N-1$,保存最佳输出结果。

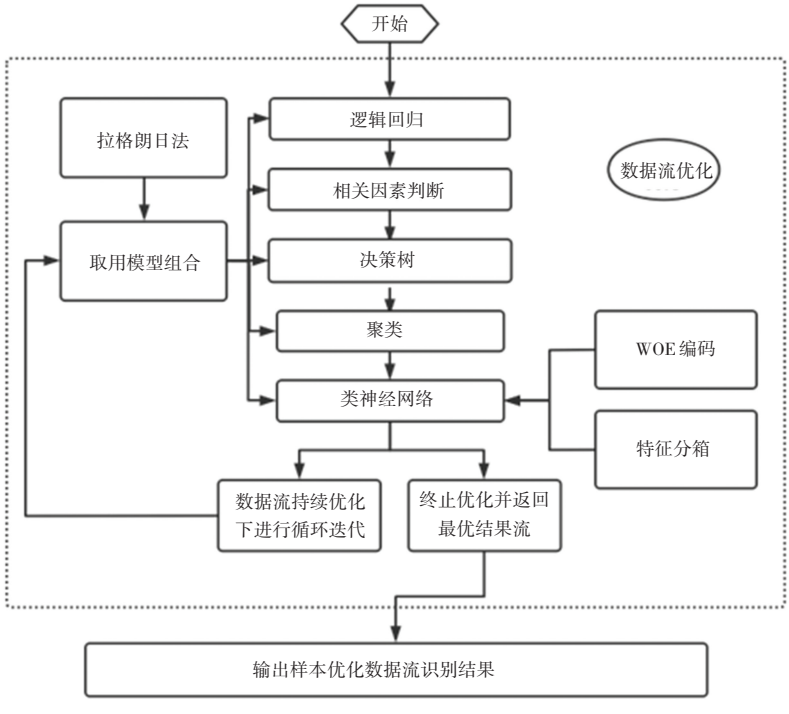


图 1 LDKB 算法流程

Fig. 1 LDKB flow path

1.2 组循环判断实现

(1)原理:取最优组合权重向量,利用精度法判断最优模型组合。

(2)实现:当 c 为真实值,b 为使用 LDKB 法输

出其中一节得到的预测值数，a 为使用 LDKB 法另一节得到的预测值，那么利用精度法有：

$$e_a = \mathrm{sum}((c-b)^2)$$
$$e_b = \mathrm{sum}((c-a)^2) \tag{6}$$

$$w_1 = \frac{1/e_a}{(1/e_a)+(1/e_b)} \tag{7}$$
$$w_2 = \frac{1/e_b}{(1/e_a)+(1/e_b)}$$

$$x = w_1 * a + w_2 * b \tag{8}$$

即可由公式(7)、(8)得到输出预测值 x。

(3)利用结果构造损失函数，用来估量模型的预测值 $f(x)$ 与真实值 Y 的不一致程度：

$$j = \sum_{i=1}^{n}\sum_{j=1}^{n}\left[w_i w_j\left(\sum_{t=1}^{N} e_{it}e_{jt}\right)\right] \tag{9}$$

(4)利用拉格朗日乘数法得最优组合权重向量，以预测误差平方最小为最优构架优化模型，获取最优组合：

$$\min j = W^T E_n W$$
$$\text{s.t. } R_n^T W = 1 \tag{10}$$

$$W = \frac{E_n^{-1}R_n}{R_n^T E_n^{-1} R_n} \tag{11}$$

$$j = \frac{1}{R_n^T E_n^{-1} R_n} \tag{12}$$

由式(10)~式(12)可得组合预测平方和最小值 j。

1.3　BP 模型损失函数

在 LDKB 法的 BP 网络部分进行识别类时，使用 softmax 模型利用权重参数 w，偏差 b（即上文提到的预测输出值个数），将输出值作为对该类别的置信度，通过 softmax 模型将其转化为正且为 1 的概率分布[10]。对于训练集样本，构造向量 $\boldsymbol{y} \in R^q$，其属于哪个类别，就将那个类别的值置为 1。使用交叉熵函数训练：

$$H(p,q) = -\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \tag{13}$$

训练 BP 网络时，由于输入数据标签已经确定（分布率 $P(x)$ 已经确定），因此信息熵为常量[11]。KL 散度等于交叉熵-信息熵，因此需要最小化 KL 散度，所以选用交叉熵损失函数计算 loss 即可。

1.4　特征分箱与 WOE 编码

LDKB 法进入循环前，需将连续变量离散化或将多状态的离散变量组合并成少状态的变量，易于 LDKB 模型的快速迭代。根据向导变量，将现有的连续变量按照导向变量间差异最大化的原则离散化为分类变量。

在 LDKB 法的逻辑回归部分，由于数据流中数据难以判断线性相关关系，数据存在极大偶然性，因此需要 WOE 编码将回归系数“正则化”。WOE 法公式如下：

$$\mathrm{WOE}_i = \ln\left(\frac{B_i/B_T}{G_i/G_T}\right) \tag{14}$$

公式(14)利用审批中的 good 与 bad 比例作为条件，分别计算数据流分箱各项的 WOE 值。通常情况下，可以通过建立较少的分箱提高数据的平滑性，WOE 重新编码后可以很容易的建立自变量与目标变量间的单调关系。

1.5　决策剪枝

在 LDKB 法的决策树部分，最容易出现的问题是过拟合[12]。剪枝过程需针对 i 层的计算，其关键需要获得等于 $n+1$ 循环操作的 *FLOPS*，即

$$T_{FLOPS} = 2(1-z)c_{out}f + (1-z)c_{out}[(zc_{out}+1)h + zc_{out}h] \tag{15}$$

用权重因素的剪枝算法运行中，第 i 层保留的 *FLOPS* 为

$$P_{FLOPS} = 2(1-z)c_{out}f \tag{16}$$

因此，在 LDKB 算法下的 *FLOPS* 计算量差异为

$$\frac{(T_{FLOPS}-P_{FLOPS})}{P_{FLOPS}} \tag{17}$$

通过式(14)的推导，在相同的剪枝策略下，LDKB 法的剪枝不但没有增加计算，反而降低了训练时间。

2　实验结果分析

2.1　实验设置

文章采用 LDKB 法，利用银行中的个人信用评级数据作为测试数据流进行循环处理，分别记录循环中输出的模型概况与模型准确率等提升情况，比较 LDKB 法与单模型训练差距，用以验证 LDKB 法的优势。

2.1.1　简介与预处理

本文使用数据是来自银行对于客户的申请信息表、汇总信息表，消费记录表与拖欠记录表构成，数据总量 1 万条。摘除重复的字段，将缺失率超过 50%的数据字段舍去。无法简单归一化变量，利用随机森林算法进行数值填充。加入 WOE 编码将分箱后的字段分别计算其 WOE 值，将回归系数“正则化”（详见 1.4 节）。

2.1.2 回归部分

回归阶段使用部分差量较大的字段,利用 WOE 编码(详见 1.4 节)平滑数据,显示 good 与 bad 的差异。利用 Forward、Remove、Backward、Stepwise 优化回归模型,以及 Chi-Square、-2 Log likelihood、BIC、AIC 作为衡量标准,确立回归部分的最优参数。

2.1.3 决策树部分

利用拉格朗日乘数法得 LDKB 循环最优组合中含有决策树模型结果输出见表 1。LDKB 法循环中数据流会逐步通过选用模型调整与完善,因此不会受到原数据反向 Acc 的质量影响。

表 1 决策树迭代节点

Tab. 1 Decision tree iteration section

	Acc	means	AUC	Gini	count
1	21.69	DT	0.494	-0.012	6 970
2	21.85	DT	0.61	0.219	6 970
3	63.6	DT	0.656	0.399	6 970
4	71.9	DT	0.835	0.463	6 970
5	72.13	DT	0.835	0.47	6 970
6	70.01	DT	0.833	0.466	6 970

循环中的总数据流经过 LDKB 算法优化,验证对于识别目标的影响程度,仅用作数据增强后的字段筛选,输出终止 $n-1$ 次最优效果。

由表 1 可看出,决策树模型被选中循环 6 次,从 Acc、AUC、Gini 系数看出其中第 1,2 次循环数据杂冗严重且准确率低,在循环过程中数据流逐步被优化,在第 6 次调用时准确率下降循环终止,输出影响因素字段。

2.1.4 聚类部分

表 2,利用 LDKB 法循环中,聚类被选用 3 次,同时 LDKB 法报出 amount 为 2 或 3 时,quality 值相等的特殊情况。因此,LDKB 法在循环运行时首先在 BP 网络识别阶段验证首类中是否存在 A 与 B 完全分开互补影响的情况,再进行下一类迭代。

表 2 聚类迭代节点

Tab. 2 Clustering iteration node

	amount	quality	ratio
1*	2	0.9	1.47
2	3	0.9	13.59
3	4	0.7	2.60

2.1.5 类神经网络部分

本研究在 BP 网络循环节点中准确率于 2 簇网络起并于 4 簇网络开始回弱,见表 3。

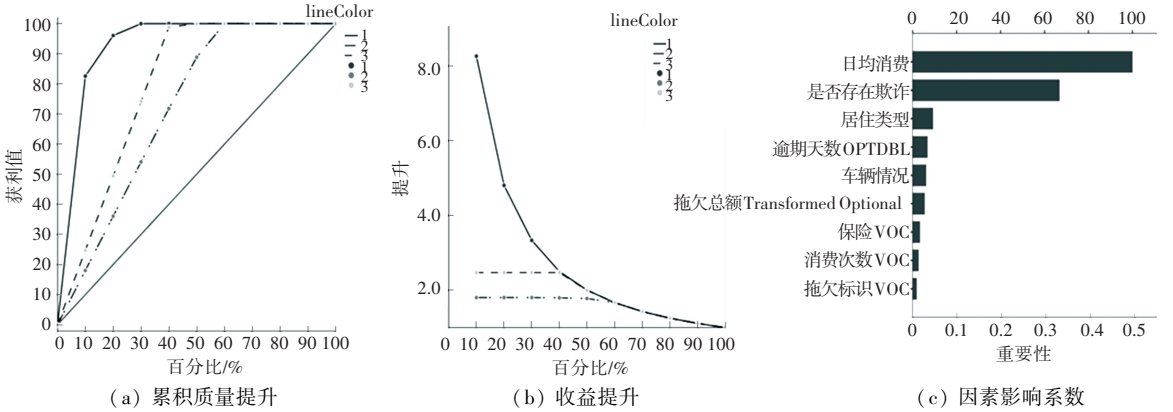
表 3 利用 LDKB 法在 BP 网络节点经过迭代输出最优结果。显而易见,聚类循环过程中特殊类 1* 中为因素完全分开互补,可使用其判断好坏,无法用作因素识别。因目标中存在完全互补的量,因此网络识别中总存在存储该效果的簇。输出最优结果如图 2 所示。

表 3 BP 网络迭代节点

Tab. 3 BP network iteration node

	Train Acc	Test Acc	ROC			
特殊识别						
1*	100	100	1		1	
四簇网络识别						
1	69.2	68.6	0.857	0.809	0.849	1
2	71.4	72	0.891	0.845	0.864	1
3	73	72.4	0.902	0.853	0.878	1
三簇网络识别						
1	97	96.5	0.927	0.987		1
2	96.9	97	0.944	0.991		1
3	96.9	97.1	0.936	0.990		1
4	96.8	97.3	0.930	0.989		1

注:表中 * 为特殊值训练、ROC 个数为循环中聚类数。



(a) 累积质量提升 (b) 收益提升 (c) 因素影响系数

图 2 识别结果输出

Fig. 2 Identification result output

2.2　实验分析

通过采用 LDKB 法识别信用等级，得到以下研究成果：

LDKB 循环回归模型初步判定环节中，通过调整分别使用 Forward、Remove、Backward、Stepwise 4 种方法优化，利用 Chi－Square、－2 Log likelihood、BIC、AIC 4 种因素衡量模型状态，并在循环中利用决策树判定效果，通过公式计算父、子树的最优最小记录数进行优化模型，并在 LDKB 循环法中不断验证数据流，输出变量因素同比优化 232.54%，达 72.13%。

LDKB 循环聚类环节共计调用 3 次，测试出特殊聚类质量数并验证，数据存在 2 类为因素完全分开互补情况，同步输出正常聚类效果结果并作为 BP 网络识别因素。

LDKB 循环 BP 网络环节中，采用构建的 BP 模型，利用计算得到隐藏层最小单位为 23，最大单位为 102，同时计算最初学习率为 0.43，在保证误差精确度为0.001 0的前提下，利用参数参与迭代优化后的 BP 模型收敛速度提高 37.29%。

采用基于模型组合的网络迭代法对目标进行识别，总体 Acc 达 97.3%，ROC 总体检测效果优于任意单个模型或单个调参后模型。

综上，LDKB 法有效实现了客户信用等级因素的评估，并有效根据历史记录识别信用等级。

3　结束语

本文提出的基于模型组合的网络迭代法（LDKB），使用拉格朗日乘数法取得最优组合权重向量，并作为选用模型组合的依据。实验证明，利用回归、聚类加以辅助，提高了 LDKB 的容错性；迭代优化数据流用于识别与多种单模型算法识别相比，LDKB 法得到的信息有着更少的精度损失，在决策识别时可以利用更小的决策成本得到更好的模型压缩效果。

根据实例分析可知，结合了 LDKB 算法的信用等级识别系统性能提升较高，输出也更接近实际值，可达到评定需求。

进一步探索在循环迭代过程中同步迭代模型最优参数[13]，实现实时调优；当前迭代过程中无法实现实时了解模型间的相互影响因素，其中决策树模型剪枝效果是对于预训练权重的数据环境敏感，因此迭代终止前可能产生当前迭代环境整体优化，而调参后的单模型效果并非最优。因此，未来可以尝试调整实现单模型最优，以查看模型间的潜在影响因素[14]。

参考文献

[1] 王琦. 个人信用评分制度在美国的应用[J]. 现代金融，2003(4)：38-39.

[2] 姜瑞. 大数据背景下的个人信用管理体系研究[J]. 中国市场，2018(29)：1-3.

[3] 刘春平.神经网络的应用与发展[J]. 电子工艺技术，2005(6)：53-54.

[4] 杨晓帆，陈廷槐. 人工神经网络固有的优点和缺点[J]. 计算机科学，1994(2)：23-26.

[5] 余本国. BP 神经网络局限性及其改进的研究[J]. 山西农业大学学报(自然科学版)，2009，29(1)：89-93.

[6] 涂占新. 数据挖掘方法及其应用展望[J]. 中南财经政法大学学报，2003(2)：117-120.

[7] 盛蕾，陈希亮，康凯. 神经网络非梯度优化方法研究进展[J/OL]. 计算机工程与应用：1-19[2022-06-07].http://kns.cnki.net/kcms/detail/11.2127.TP.20220523.1747.026.html

[8] 张桂梅，龙邦耀，曾接贤，等. 基于去冗余特征和语义关系约束的零样本属性识别[J]. 模式识别与人工智能，2021，34(9)：809-823.

[9] LIU Z，LI J G，SHEN Z Q，et al. Learning efficient convolutional networks through network slimming[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision，Washington：IEEE Computer Society，2017：2755-2763.

[10] 米雅婷. 基于 GA-BP 神经网络的温室番茄病害诊断研究[D]. 哈尔滨：东北林业大学，2016.

[11] 曲海成，张雪聪，王宇萍. 基于信息融合策略的卷积神经网络剪枝方法[J]. 计算机工程与应用，2022，58(24)：125-133.

[12] YU L，XIE L，LIU C，et al. Optimization of BP neural network model by chaotic krill herd algorithm[J]. Alexandria Engineering Journal，2022，61(12)：9769-9777.

[13] 张逸方，吴佩芬. IMDb 电影影评之单类神经网络与改良型 CNN 模型准确率差异性研究[J]. 电影评介，2021(7)：59-62.

[14] 吴赟. 大数据环境下的个人征信及其边界研究[D]. 北京：中国社会科学院研究生院，2020.