逆传输类神经网络中非对称数据优化算法研究

魏传佳

（泉州轻工职业学院，福建泉州 362200）

【摘要】神经网络算法在数据分类与计算中对非对称数据的处理是一项亟待解决的难题。文章提出一种修改权重的逆传输类神经网络算法，通过修改自学习效率，对占有较少类的数据分配高权重来解决非对称平衡问题。仿真结果表明，与其他五种分类算法对比，本算法在不影响算法复杂度的情况下，提高了对非对称数据运算的精确性与有效性。

【关键词】神经网络；非对称数据；逆向传输；算法有效性

【doi:10.3969/j.issn.2095-7661.2021.03.007】

【中图分类号】TP183 【文献标识码】A 【文章编号】2095-7661(2021)03-022-04

The Optimization Algorithm Research for Asymmetrical Data of Reverse Transmission Neural Network

WEI Chuan-jia

（Quanzhou College of Technology, Quanzhou, Fujian, China 362200）

Abstract: Neural network algorithm in data classification and calculation of asymmetric data processing is an urgent problem. Therefore, this paper proposes a modification of the weights inverse transmission neural network algorithm. By modifying the efficiency of self-learning and assigning high weights to data that occupies fewer classes, the asymmetric balance problem is solved. Compared with the other five classification algorithms by simulation data, this algorithm does not affect the algorithm complexity of the situation, improving the operation of asymmetric data accuracy and effectiveness.

Keywords: neural network; asymmetric data; reverse transmission; effectiveness of the algorithm

随着数据的不断增长，机器学习算法在处理海量数据时，其运算速度会逐渐下降，非对称数据分类是否有达到平衡状态是一个研究热点，所以产生了修改训练集[1]的算法去处理非对称问题，以往的研究中主要专注于二重数据类别的分类，即只研究两种分类的结果，如支付设计中的成功与否等。本文主要解决算法在处理非对称数据时，分类效果不佳的问题，利用修改逆传输类神经网络的分类权重方法，在处理非对称数据情况下，仍能保持较高的精准度与较低的时间冗余度。

1 非对称数据分类

1.1 非对称数据问题描述

非对称数据的问题主要体现在，一个数集中每一个类别的实例与其他样本的数量存在着显著的差异。计算处理有关非对称问题的时候，主要是针对二元次的数据分类，二元次分类指的是如果其中一个实例的数量相对来说是较多的，那么把它称为多数类或叫做负类，另外，如果一个实例的数量是较少的，就把它称之为少数类或叫做正类。这种分类真实存在于如学生的修学可能性、生物信息、破产评估、疾病发现等领域。而这些应用中重要的信息都属于少数类，如果无法辨识出这些少数类结果，成本代价会非常高昂。出现这种问题的主要原因是，针对目标的机器学习算法是基于全局的搜索，并不会考虑每一个实例的数量。这会造成对少数类的忽略，因少数类反映出的特征较少，会导致错判率的上升。

以往通常使用不平衡率（IR）去判断一个数据

[收稿日期] 2021-08-06

[作者简介] 魏传佳(1983-)，男，黑龙江佳木斯人，泉州轻工职业学院副教授，在读博士，研究方向：人工智能、神经网络。

[基金项目] 福建省教育厅2020年科研项目“基于模拟退火算法的无线网络优化算法研究”（项目编号：JAT201502）。

集的不平衡程度。首先，判断数据集中哪个属于多数类，哪个属于少数类，然后用多数类去除以少数类，得到不平衡率IR，因此就能够去判断此数据集的不平衡的程度。当进行分类时，不仅不平衡数会影响分类的精准度，数据固有的特征也会影响到分类的精准度，如：样本的大小，离群值，边界样本，缺失值等。

1.2 处理非对称数据方法

处理非对称数据方法很多，技术特点通常分为三个层次：第一是数据层级的方法，主要修改原始训练集，得到一个近似于对称的数据集，这个数据集可被使用在标准的机器学习方法中；第二是修改演算法，主要是修改现有的演算法的内部操作，使其可以处理非对称数据；第三是成本敏感性分析，给予少数类实例较高的误判成本，对于其他的类别给予较低的误判成本。

数据层级计算方法主要有三种，oversampling过采样方法、under sampling欠采样方法和hybrid杂交方法。这三种方法中，相对简单的有random oversampling，它随机的从原始数据集中产生少数类（即正类）数据，直到少数类与多数类数据达到平衡。另一个是random under sampling方法，它会随机删除多数类，直到多数类与少数类相趋近。SMOTE算法与random oversampling和random under sampling相比是一种较为复杂的方法，属于过采样方法的一种，通过合成少数类数据，使得原数据集趋于平衡。

成本敏感分析是充分考虑错误分类的成本，假设数据资料中有M种类别，错误分类成本可以用一个M×M矩阵来表示，如表1所示，设$\cos t(i,j)$为实际类别是 j 的资料被分为 i 的错误分类成本。在表1中，$\cos t(T,F)=2000$，$\cos t(F,T)=8000$，当且仅当只有两个类别时，$\cos t(i,j)$也表示为将实际类别标记为 j 的错误分类成本。

表1　成本敏感矩阵举例

类别	T	F
T	0	2000
F	8000	0

对于错误成本的分类设计方法，通常不去改变传统的方法，而是对训练数据进行预处理或者对结果进行结果处理，让传统算法去考虑错误分类的结果，这样就不需要去修改原有的分类方法，因此可以直接套用在许多算法里。常用的方法有修改权重法、门限法等。

1.3 评定非对称数据分类精准度

对于分类的精准度，需要通过一个混合矩阵去说明，混合矩阵的组成是通过每个类别的错误分类或者正确分类的区分而组成，如表2所示。

表2　成本敏感矩阵举例

	Positive-prediction	Negative- prediction
Positive class	True Positive	False Negative
Negative class	False Positive	True Negative

当进行分类精度判断的时候，将其分为两种计算方式，一种是不区分少数类和多数类的计算，一种是区分少数类和多数类的计算。不区分类方法公式如下：

$$\frac{TP+TN}{TP+TN+FP+FN}=ACCURACY \tag{1}$$

此种方法对少数类的分类结果不是很满意，但对分类的结果的评价却很高。对于分类精度判断的时候，采用区分少数类与多数类的计算方法，本文也采取此方法进行计算。定义公式如下：

$$GM=\sqrt{sensitivity\cdot specificity} \tag{2}$$

公式（2）中的*sensitivity*计算公式为（3）所示：

$$sensitivity=\frac{TP}{TP+FN} \tag{3}$$

公式（2）中的specificity计算公式为（4）所示：

$$specificity=\frac{TN}{FP+TN} \tag{4}$$

此种操作方法可最大化两个类别的精准度，并且具有很好的平衡性。

2 逆传输类神经网络

2.1 类神经网络

人类脑海中有超过1000亿个神经细胞，每个神经细胞中，含有神经元[2]，细胞核，轴突，树突，突触等。细胞核为处理器，轴突为信号传输介质，树突为信号线，突触是神经元之间的连接点。图1为神经元，图2为人工类神经元运算模型，图3为类神经网络模型。类神经网络是仿神经网络去处理复杂网络问题，信息来源从人工神经元与外在环境中获得，根据不同的网络拓扑结构和机器自学习算法去训练类神经网络，最终获得目标值。

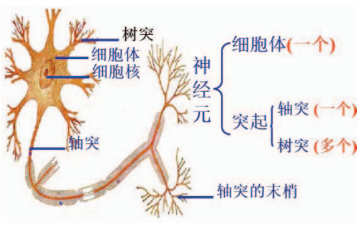

图1　神经元结构图

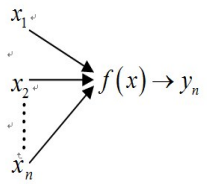

图2 人工类神经元运算模型图

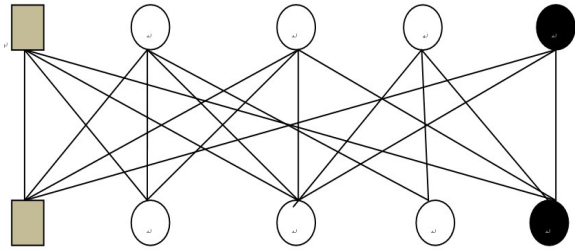

图3 类神经网络模型图

2.2 逆传输类神经网络

逆向传输类神经网络[3-4]是一种检测型自学习算法，利用多层次架构模式去构造模型。本文定义公式如下：

i input层的第i个节点，$i = 1,2,\cdots,p$

j hide层的第j个节点，$j = 1,2,\cdots,h$

k output层的第k个节点，$k = 1,2,\cdots,q$

l 第l个训练元素，$l = 1,2,\cdots,n$

w_{kj} hide层与output层之间节点的权重

z_j^l 第l个训练元素在hide层中节点j的output值

y_k^l 第l个训练元素在output层中节点k的output值

f 神经元动态化函数

d_k^l 第l个训练元素在output层节点k的目标值

w_{j0} hide层的门限值

w_{k0} input层的门限值

w_{ji} input层与output层之间节点的权重

η 自学习效率

x_i^l 第l个训练元素中节点i的input值

δ_k output层的误差

2.3 逆传输类神经网络算法

1)初始化各项参数；

2)初始化w_{ji}与w_{kj}，选定动态化函数；

3)选取训练集$x_i^l = \left(x_1^l, x_2^l \cdots x_p^l\right)$，目标集$d_i^l = \left(d_1^l, d_2^l \cdots d_p^l\right)$；

4)计算hide层output值z_j^l，output层的output值y_k^l；

5)求出误差函数E；

6)求出output层的差距量δ_k公式(5)，hide层的差距量δ_j公式(6)；

$$\delta_k = \left(d_k - y_k\right) y_k \left(1 - y_k\right) \tag{5}$$

$$\delta_j = \sum_k \left(\delta_k w_{kj}\right) z_j \left(1 - z_j\right) \tag{6}$$

7)求出output层与hide层之间的权重修正值Δw_{kj}^l公式(7)，input层与hide层之间的权重修正值Δw_{ji}^l公式(8)；

$$\Delta w_{kj}^l = \eta \delta_k z_j \tag{7}$$

$$\Delta w_{ji}^l = -\frac{\partial E}{\partial w_{ji}} = \eta \delta_j x_i \tag{8}$$

8)权重更新：

$$w_{kj}^{l+1} = w_{kj}^l + \Delta w_{kj}^l \tag{9}$$

$$w_{ji}^{l+1} = w_{ji}^l + \Delta w_{ji}^l \tag{10}$$

9)返回步骤3，循环计算，令$l = 1 + l$，直到训练完所有元素；

10)循环计算，直到循环次数达到最大为止。

3 改进优化逆传输类神经网络算法

针对非对称数据分类问题[5-7]，提出改进权重的逆传输类神经网络算法(MWRTNN)，此算法通过修改自学习效率，对占有较少类的数据分配高权重[8]来解决非对称平衡问题。算法步骤如下：

1)网络拓扑模型建立，初始化各层数据以及最大自学习周期，令$l = 1$；

2)权重初始化w_{ji}与w_{kj}，选定节点输出活化函数；

3)选取训练集，令$x_i^l = \left(x_1^l, x_2^l \cdots x_p^l\right)$，$d_i^l = \left(d_1^l, d_2^l \cdots d_p^l\right)$；

4)求出hide层各个节点的输出z_j^l，output层各个节点的输出y_k^l；

5)求出误差函数E；

6)求出output层的差距量δ_k公式(11)，hide层的差距量δ_j公式(12)；

$$\delta_k = \left(d_k - y_k\right) y_k \left(1 - y_k\right) \tag{11}$$

$$\delta_j = \sum_k \left(\delta_k w_{kj}\right) z_j \left(1 - z_j\right) \tag{12}$$

7)在训练数据时，对于连接点的权重修改问题，先去判断训练集属于正分类还是负分类。如果属于正分类，就使用正分类权重，如果属于负分类，就使用负分类权重。求出output层与hide层之间的连接点修正权重，正分类修正权重为Δw_{kj}^{l+}公式(13)，负分类修正权重为Δw_{kj}^{l-}(14)；求出hide层与input层之间的连接点修正权重，正分类修正权重为Δw_{ji}^{l+}公式(15)，负分类修正权重为Δw_{ji}^{l-}公式(16)；

$$\Delta w_{kj}^{l+} = \eta IR \delta_k z_j \tag{13}$$

$$\Delta w_{kj}^{l-} = \eta \delta_k z_j \tag{14}$$

$$\Delta w_{ji}^{l+} = \eta IR\delta_j x_i \quad (15)$$

$$\Delta w_{ji}^{l-} = \eta \delta_j x_i \quad (16)$$

8)连接点权重更新,求出output层与hide层之间的连接点权重更新,正分类修正权重更新为w_{kj}^{l+1}公式(17),负分类修正权重更新为w_{kj}^{l+1}公式(18);求出hide层与input层之间的连接点权重更新,正分类修正权重更新为w_{kj}^{l+1}公式(19),负分类修正权重更新为w_{kj}^{l+1}公式(20);

$$w_{kj}^{l+1} = w_{kj}^{l} + \Delta w_{kj}^{l+} \quad (17)$$

$$w_{kj}^{l+1} = w_{kj}^{l} + \Delta w_{kj}^{l-} \quad (18)$$

$$w_{ji}^{l+1} = w_{ji}^{l} + \Delta w_{ji}^{l+} \quad (19)$$

$$w_{ji}^{l+1} = w_{ji}^{l} + \Delta w_{ji}^{l-} \quad (20)$$

9)令$l = l + 1$,返回到步骤3,直至训练完成;

10)重复执行步骤2至步骤9,直到匹配最大周期。

4 实验结果比较

为了分析算法执行效果,从UCI库中选取三种不同的DATA集,KDD CUP99、RLCP、POKE HAND DATA SET。这三种数据集属于多类别分类集[8],为此将其分为多个数据集,用来解决类间的分类问题。三个数据集信息如表3所示,其中包含实例数量(*Ex.),属性数量(*Att.),类别名称组(Class(Maj;Min)),各类别实例数量(*Class(Maj;Min)),各类别所占百分比(%Class(Maj;Min)),不平衡比率(Ubr)。为扩展实验,本文使用5-Fold Cross Validation进行验证。

表3　非对称数据表

Datasets	*Ex	%Class(Maj;Min)	Ubr
Ku U2L	972832	(99.994%;0.006%)	18707.2
Ku R2L	973906	(99.883%;0.117%)	863.925
Ku PRB	1013882	(95.945%;4.055%)	23.666
Pr 0-2	562531	(91.32%;8.68%)	10.52
Pr 0-3	535335	(95.958%;4.042%)	23.74
Pr 0-4	517681	(99.231%;0.769%)	129.12
Pr 1-2	481924	(89.867%;10.133%)	8.86
Pr 1-3	454730	(95.241%;4.759%)	20.01
RP	5749131	(99.635%;0.365%)	273.66

为证明RTNN对非对称数据分类效果的提升,本文以六种算法做比较,结果如表4与表5所示。

表4　RTNN与MWRTNN比较

	RTNN		MWRTNN	
	Train	Test	Train	Test
Pr 0-2	0.282	0.281	0.537	0.535
Pr 0-3	0.222	0.224	0.503	0.502
Pr 0-4	0.289	0.292	0.487	0.487
Pr 1-2	0.046	0.044	0.489	0.493
Pr 1-3	0.033	0.031	0.512	0.514
Ku 2r	0.597	0.599	0.761	0.762
Ku R2L	0.442	0.442	0.557	0.563
Ku PRB	0.000	0.000	0.754	0.741
RP	0.000	0.000	0.613	0.613

由表4可知,RTNN网络在处理非对称数据时,效率不高,而本文提出的MWRTNN算法,大大改善了分类上的效率,对比其他算法,分类效果有了较大的改善,如表5所示。

表5　MWRTNN与其他算法比较

	MWRTNN		RTNN-ROS		RTNN-RUS		RTNN-SMOTE	
	Train	Test	Train	Test	Train	Test	Train	Test
Pr 0-2	0.536	0.535	0.321	0.321	0.421	0.424	0.517	0.517
Pr 0-3	0.503	0.502	0.483	0.483	0.489	0.492	0.495	0.494
Pr 0-4	0.487	0.487	0.568	0.568	0.677	0.676	0.518	0.518
Pr 1-2	0.489	0.493	0.453	0.452	0.479	0.483	0.264	0.264
Pr 1-3	0.512	0.514	0.506	0.506	0.503	0.504	0.434	0.434
Ku 2r	0.761	0.762	0.600	0.600	0.631	0.628	0.600	0.599
Ku R2L	0.557	0.563	0.593	0.593	0.583	0.583	0.605	0.605
Ku PRB	0.754	0.741	0.605	0.606	0.573	0.556	0.648	0.648
RP	0.613	0.613	0.410	0.410	0.501	0.502	0.457	0.456

5 结束语

对于逆向传输类神经网络处理非对称数据时效率低下的问题,本文提出了一种基于权重修改的逆向传输类神经网络算法,该算法切实解决数据分类的问题,对比其他算法,效果显著。但该算法在处理海量数据时候,时间冗余度会比较大,如何处理此问题,将在以后的研究中继续改进。

【参考文献】

[1]Labovitz C, Iekel-Johnson S, McPherson D, et. al. Internet inter-domain traffic [J]. Computer communication review, 2010 (4):75-86.

[2]豆育升,崔晟圆,唐红,李鸿健.云数据中心高能效的虚拟机放置算法[J].小型微型计算机系统,2014(11):2543-2547.

(下转第58页)

位，将学生的自我评价、学生互评与师生互评有机结合，采取匿名评价方式保护师生个人隐私，在提高学生参与感的同时，促进教师教学能力的提高。再次，切实落实教师互评。高校应主张匿名评价，避免教师碍于人际关系做出不真实评价。同时教师应重视教师互评评语，清楚认识教学理念与方式的优缺点，及时查缺补漏，提升教学质量。最后，引进校外体育专家参与评价。高校可以引进体育学科带头人、优秀教师作为评价主体来参与校内教师评价，学习先进的体育教学评价理念和经验，明确体育教学改革方式，创新体育教学评价体系。

3.4 注重评价功能转化，选择多元评价方式

评价功能的转化与表现取决于评价方式的创新与融合，多元评价方式有助于体育教学评价信息反馈作用的发挥。高质量、高效率的教学评价方式可以对体育教学进行精准地价值判断，监督教师教学过程及其效果，强化学生学习态度。对此，高校应改善体育教学评价方式单一现状，选择多元评价方式。首先，高校应注重阶段性评价，强化教学监督功能。阶段性评价的补充与融合，能够充分发挥评价在课堂教学过程中的监督调节功能，将期末考核与课堂评价结合在一起，监督学生课堂学习纪律，培养学生运动习惯，同时，阶段性评价可以促进教师教学方式的创新与课堂质量的提升。其次，高校应注重定性评价，深化体育育人价值。健康中国战略背景下，体育教学评价应从外向内发展，凸显体育教学对学生意识、态度及行为养成等内在因素的作用。高校应融入定性评价，凸显教书育人本质，完善体育教学评价体系。

4 结语

随着国民健康生活意识的不断提高，“健康中国”已经成为高校体育教学改革发展的全新教育理念。将传统体育运动从增强体质的角度转变为促进身心健康的角度，从更深刻的层面实现高校体育教学意义。在现代社会人们生活压力增大的背景下，高校体育发展应充分利用“健康中国”教育理念，培养高校大学生展开体育运动的自主性与能动性，进而形成终身体育的意识，促进当代大学生身心全面发展。

【参考文献】

[1]万炳军,史岩,曾肖肖.“健康中国”视域下体育的价值定位、历史使命及其实现路径[J].北京体育大学学报,2017(11):1-9.

[2]任一春.高校体育教学评价功能异化与理性回归[J].集宁师范学院学报,2019(6):62-66.

[3]潘丽萍.高校体育课程改革评价指标体系的优化研究[J].浙江体育科学,2019(6):73-77.

[4]董川,陈玲.普通高校体育教学"立德树人"的路径探索[J].兰州文理学院学报(自然科学版),2020(2):125-128.

[5]徐树礼,侯学华,林琳,郭力.对我国现行体育教学评价的质疑与建议[J].体育研究与教育,2019(2):49-52.

[6]刘莹.体育传统教学评价功能及其重构[J].教学与管理,2016(36):122-124.

[7]王国亮,詹建国.翻转课堂引入体育教学的价值及实施策略研究[J].北京体育大学学报,2016(2):104-110.

[8]张文波,赵利.价值冲突与回归:高校体育教学评价问题研究[J].广州体育学院学报,2015(1):110-112,121.

[9]王珂.高校体育教学评价的现状及改进方法[J].体育视野,2020(8):73-74.

（上接第25页）

[3]Falkenauer E, Delchambre A. A genetic algorithm for bin packing and line balancing[C]. Proceedings of IEEE International Conference on Robotics and Automation,1992.

[4]Gao Y, Guan H, Qi Z, et al. A multi-objective antcolony system algorithm for virtual machine placement in cloud computing[J]. Journal of Computer and System Sciences, 2013 (8):1230-1242.

[5]Aarts E , Laarhoven P . Statistical cooling : a general approach to combinatorial optimization problems[J]. Philips Journal of Research, 1985(4)193-226.

[6]Anshul Gandhi, Mor Harchol-Balter, Rajarshi Das, Charles Lefurgy. Optimal power allocation in sever farms[C]. Measurement and Modeling of Computer Systems, ACM, New York,NY, USA , 2009:157-168.

[7]Gong Chen, Wenbo He, Jie Liu, Suman Nath, Feng Zhao. Energy-aware server provisioning and load dispatching for connection-intensive internet services[C]. Usenix Symposium on Networked Systems Design & Implementation. USENIX Association, 2008:337-350.

[8]Duy T , Sato Y , Inoguchi Y . Performance evaluation of a Green Scheduling Algorithm for energy savings in Cloud computing[C]. IEEE International Symposium on Parallel & Distributed Processing. IEEE, 2010.