基于改进 BP 神经网络的非均衡数据分类算法①

张文东, 吕扇扇, 张兴森

(中国石油大学(华东) 计算机与通信工程学院, 青岛 266580)

摘 要: 传统的分类算法大都默认所有类别的分类代价一致, 导致样本数据非均衡时产生分类性能急剧下降的问题. 对于非均衡数据分类问题, 结合神经网络与降噪自编码器, 提出一种改进的神经网络实现非均衡数据分类算法, 在神经网络模型输入层与隐层之间加入一层特征受损层, 致使部分冗余特征值丢失, 降低数据集的不平衡度, 训练模型得到最优参数后进行特征分类得到结果. 选取 UCI 标准数据集的 3 组非均衡数据集进行实验, 结果表明采用该算法对小数据集的分类精度有明显改善, 但是数据集较大时, 分类效果低于某些分类器. 该算法的整体分类效果要优于其他分类器.

关键词: 非均衡数据; 神经网络; 降噪自编码器; 分类

Unbalanced Data Classification Algorithm Based on Improved BP Neural Network

ZHANG Wen-Dong, LV Shan-Shan, ZHANG Xing-Sen

(College of Computer and Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract: Most of the traditional classifications algorithms have the same classification cost of all categories, which results in a sharp decline in classification performance when the sample data are unbalanced. As to the problem of unbalanced data classification, we combine neural network with denoising auto-encoder and put forward a kind of improved neural network to realize unbalanced data classification algorithm. The algorithm adds a layer called feature damaged layer between input layer and hidden layer. Thus some redundant feature values are lost, and the unbalance degree of data set is reduced. And the results can be obtained after training model obtains optimal parameters and deals with the classification based on feature. It selects three sets of UCI standard unbalanced data sets for experiment. The results show that the accuracy of the algorithm for small data set classification is improved obviously, but when the data set is larger, the classification effect is lower than some classifier. And the overall classification performance of the proposed algorithm is better than other classifiers.

Key words: unbalanced data ; neural network; denoising auto-encoder; classification

在实践中, 有许多的不平衡数据分类问题, 这种情况是指二类分类问题二类分问题中, 一类样本的数量要比另一类样本数量小的多. 传统的分类算法大多默认所有类别分类代价一致, 处理非平衡数据时会偏向于多数类, 对于多数类样本实例过拟合, 对于少数类样本实例欠学习, 导致少数类的分类精度很低. 然而, 在实际生活中, 人们往往需要知道少数类是否被正确分类. 因此不平衡数据分类问题是近年来机器学习和数据挖掘的研究热点, 在各个研究领域比较常见, 例如信用卡欺骗, 癌症发病率等[1-3].

针对上述问题, 研究者提出了多种分类方法, 陶新民等人[4]提出基于 ODR 和 BSMOTE 结合的不均衡数据 SVM 分类算法, 该算法不但能有效提高 SVM 算法在不均衡数据中少数类的分类性能, 而且总体分类性能也有所提高. 但该算法中某些参数的设置仍然是经验性的, 无法确定其最优效果. 韩敏等人[5]提出一种径向基函数神经网络和随机森林集成的混合分类算法, 该算法先让样本达到均衡状态, 再采用针对小类的神经网络集成学习方法进行训练和测试, 有效提高了小类样本的分类性能. 但扩充样本会花费大量内存

① 收稿时间:2016-09-22;收到修改稿时间:2016-11-03 [doi:10.15888/j.cnki.csa.005797]

© 中国科学院软件研究所 http://www.c-s-a.org.cn

及运算时间, 导致效率降低. Wang 等人[6]提出了采用直推式支持向量机和编辑紧邻规则解决不平衡数据的相结合的两个改进算法, 第一种改进是选择一些有用的正面的可能包含嘈杂的测试样本用来补充缺乏训练样本, 然后用编辑最近邻规则去除嘈杂的样本. 第二种改进选择一些有用的测试文件的正面和负面的类样本添加到训练集, 去除嘈杂样本的方法类似于第一种改进算法. 这两种算法都添加了测试样本, 以保持少数类和大多数类之间的平衡, 经实验证明了这两种算法的可行性和有效性, 并与支持向量机相比均有明显的提高.

基于以上分析, 本文给出了一种基于 BP 神经网络的非均衡数分类算法来处理非均衡数据的分类, 并使其性能得到了明显的提高. BP 网络是一种按误差逆传播算法训练的多层前馈网络, 能学习和存贮大量的输入-输出模式映射关系, 而无需事前揭示描述这种映射关系的数学方程. 该算法利用 BP 神经网络进行输入输出映射, 并使用随机梯度下降方法, 通过反向传播来不断调整网络的权值和阈值. 本文关键点在于, 训练阶段在输入数据中加入一定量的噪声数据, 以增加其泛化能力. 经实验表明该方法对于解决非均衡数据的分类问题有较明显的效果, 特别是对于小数据集样本的分类准确率.

1　相关工作

1.1 BP 神经网络

BP(Back Propagation)神经网络是 1986 年由 Rumelhart 和 McCelland 为首的科学家小组提出, 是一种按误差逆传播算法训练的多层前馈网络, 一般包括单输入层、单输出层与若干隐藏层, 是目前应用最广泛的神经网络模型之一. 特征向量由输入层传入网络中, 经由隐藏层处理后, 输出层对处理结果输出. 利用误差函数计算输出结果与期望值的误差, 然后利用反向传播将计算得到的误差沿神经网络传送使用梯度下降算法进行神经元节点权值的调整, 使其沿梯度方向下降. 反复训练直至训练的迭代次数达到设定阈值或者训练样本的误差计算值符合预期误差估计时, 神经网络训练结束, 此时网络的参数最优, 能够对输入特征进行预测或者分类[7,8].

1.2 随机梯度下降算法

梯度下降是 BP 神经网络中损失函数计算最常用方法, 而随机梯度下降是梯度下降算法中一种迭代求解思路[9]. 假设 $h(x)$为待拟合的函数, 表示为公式(1), $J(\theta)$为损失函数, 表示为公式(2), θ 是模型参数, 需要迭代求最优解, 如公式(3)所示:

$$h(x)=h_\theta(x)=\sum_{j=0}^{n}\theta_j x_j \tag{1}$$

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(y^i-h_\theta(x^i))^2 \tag{2}$$

$$\min_\theta J_\theta \tag{3}$$

其中 m 是训练集的记录条数, j 是参数的个数.

从一个随机点以最快的速度达到损失最少的点, 用数学描述就是一个求偏导的过程, 公式如下:

$$\frac{\partial}{\partial\theta}J(\theta)=\frac{\partial}{\partial\theta}\frac{1}{2}\sum_{i=1}^{m}(h_\theta(x)-y)^2=(h_\theta(x)-y)x^{(i)} \tag{4}$$

这样 θ 的更新过程如公式(5)(l 表示算法的学习速率)所示:

$$\theta_i=\theta_i-l\frac{\partial}{\partial\theta}J(\theta)=\theta_i-l(h_\theta(x)-y)x^{(i)} \tag{5}$$

2　改进的BP神经网络模型

针对上文中提及对于非均衡数据分类中存在的参数的不确定性以及扩充样本降低效率等问题, 结合神经网络的较强非线性拟合能力与降噪自编码器[10]对特征更加鲁棒表达的优点, 本文提出只有一个隐层的降噪 BP 神经网络的非均衡数据分类算法, 模型结构如图 1 所示. 在输入层与隐层之间加入一层特征受损层, 使得神经网络对于非平衡数据分类的泛化能力更强, 从而增加了非平衡数据的分类性能.

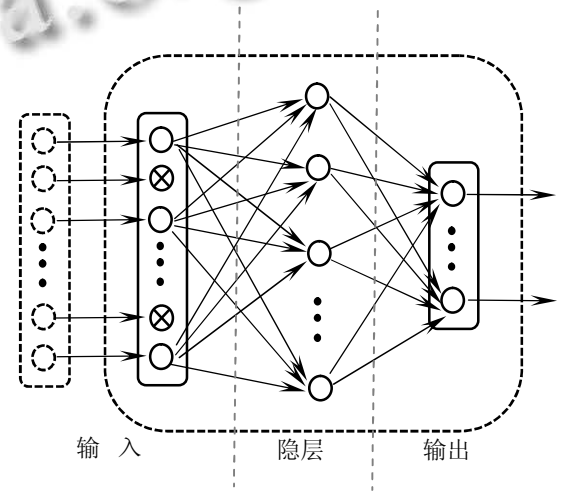

图 1　改进的 BP 神经网络模型

2.1 模型核心思想

© 中国科学院软件研究所 http://www.c-s-a.org.cn

在传统 BP 神经网络基础上, 在输入样本特征 X 时加入噪声, 添加噪声的做法采用按一定比例随机将部分特征输入置为 0, 符合人类大脑的信息处理机制, 致使部分冗余特征值丢失, 降低数据集的不平衡度, 得到 X', 再输入到神经网络进行分类并进行误差计算与权值更新.

2.2 算法主要流程

输入: 带有标签的样本数据 X, 以及一定噪声比例 ρ.

输出: 网络最优参数 θ.

步骤 1. 随机初始化参数 $\theta=\{W, b\}$,其中 W 为网络的权值, b 为偏置项, 输入层节点数目为 n, 输出层节点数目为 m, 以及隐藏层节点数目 k, 迭代次数, 学习速率以及期望误差等;

步骤 2. 将训练数据 $X=(x_1,x_2,\ldots,x_n)^T$ 以公式(6)加入人工噪声, 其中 $R(\cdot)$ 为噪声矩阵, $size(\cdot)$ 计算矩阵的大小, ρ 为噪声系数. 得到噪声样本数据 $X'=(x_1',x_2',\ldots,x_n')^T$;

$$X'=R(size(X),\rho)\cdot X \tag{6}$$

步骤 3. 将步骤 2 得到的 X'输入到神经网络中进行训练, 得到输出 Y';

步骤 4. θ 根据梯度下降算法中的公式(5)的方式进行变化, 计算迭代次数: 当训练次数达到指定值时, 停止训练, 得到最优参数 θ; 若未达到训练迭代次数指定值, 根据梯度下降算法中的公式(2)计算损失函数, 若达到制定损失值, 则停止训练得到最优参数 θ; 否则重复步骤 2~4, 直至停止训练, 得到最优参数 θ.

3 非均衡数据分类实验

3.1 数据集选择

文中实验的数据采用 UCI 机器学习库中的 3 个常用的数据集. 其中 ionosphere、wpbc 数据集是二分类不平衡数据集[11], 选择 satimage 数据集并将 class2 作为少数类. 其具体描述如表 1 所示.

表 1 实验数据集

数据集	样本总数	特征个数	少数类	多数类	非均衡比
ionosphere	351	34	126	225	1:1.8
wpbc	198	33	47	151	1:3.2
satimage	2236	36	200	1211	1:6

3.2 实验参数设置

在实验与测试环节, 依托的硬件条件为: Win7 64 位系统, 双核 CPU3.2GHZ, 内存 8G 等. 使用 Python 的集成开发环境 Spyder 实现模型训练与测试对比. 本文模型中对于数据集初始化参数设置, 首先对数据集使用少数类样本合成过采样技术处理, 输入节点数目为数据集特征数, 输出节点数目根据输出分类数目按照二进制编码确定, 隐层节点数目依据公式 $k=\sqrt{m+n}+a$, k 为隐层节点数目, m 为输出节点数目, n 为输入节点数目, $a\in[0,10]$, 并在实验中进行微调. 激活函数使用 $tanh: f(x)=tanh(x)$. 迭代次数设置为 5000 次, 学习速率为 0.005, 噪声比例 $\rho=0.3$.

4 实验结果分析

4.1 非均衡数据分类结果评价标准

在传统的分类方法中衡量分类器性能一般以错误率(分类正确的测试样本数目与总测试样本数目的比率)为指标. 针对不平衡数据, 传统的评价方法掩盖了对少数类识别的不敏感性, 目前公认的为混淆矩阵, 如表 2 所示, 它可以帮助人们更好的了解分类中的错误. 针对二分类问题给出了一个混淆矩阵, 如果将一个正例判为正例, 那么就可以认为产生了一个真正例(True Positive, TP,也称真阳); 如果对一个反例正确的判为反例, 则认为产生了一个真反例(True Negative, TN,也称真阴). 相应地, 另外两种情况分别称为伪反例(False Negative, FN,也称假阴)和伪正例(False Positive, FP,也称假阳). 其中正例为少数类样本.

表 2 二分类混淆矩阵

	预测正例	预测反例
真实正例	真正例(TP)	伪反例(FN)
真实反例	伪正例(FP)	真反例(TN)

利用混淆矩阵定义来定义出多个比传统错误率更好的新指标, 例如准确率(少数类预测准确率)accuracy、召回率(多数类预测准确率)recall、F1 与 G-mean 等, 其公式定义如下:

$$accuracy=\frac{TP}{TP+FP}$$

$$recall=\frac{TP}{TP+FN}$$

$$F1=\frac{2\cdot recall\cdot accuracy}{recall+accuracy}$$

$$G-means=\sqrt{\frac{TN}{TN+FP}\cdot\frac{TP}{TP+FN}}$$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

其中准确率表示预测为正例的样本特征中真正正例的比例, 召回率表示预测为正例的真实正例占所有真实正例的比例, F1 表示准确率与召回率之间的调和平均, G-means 表示少数类分类精度和多数类分类精度的几何平均值, 用来评价分类器的整体性能.

4.2 实验结果及分析

本文对比算法有一个隐藏层的SVM算法, 把输入的样本映射到一个高维的特征空间, 在这个特征空间中构造出最优分类超平面, 实现分类. 而 SAE 算法与 SMOTE-SDAE 算法均有 4 层网络结构, 其中包括 2 个隐层, SAE 通过组合低层特征产生更加抽象的高层特征, 以发现数据的内在特征表示; SMOTE-SDAE 算法结合少数类样本合成过采样技术(SMOTE)和降噪自编码器(SDAE)将合成新少数类样本作为训练集, 利用无监督学习方式, 预训练初始化参数, 微调抽象出最能表达原始数据的本质特征. 本文提出的一种改进的神经网络实现非均衡数据分类算法, 在神经网络模型输入层与隐层之间加入一层特征受损层, 致使部分冗余特征值丢失, 降低数据集的不平衡度, 训练模型得到最优参数后进行特征分类得到结果, 提高了分类功能.

文中采用F1与G-means两项非均衡数据分类器性能评价指标对 3 个不平衡数据集(ionosphere、wpbc、satimage)采用 SVM、SAE、SMOTE-SDAE 和提出的改进 BP 神经网络分类算法进行比较, 表 3 和 4 为各算法的 F1 与 G-means, 对比结果.

表 3 各种算法的 F1 对比(%)

数据集	SVM	SAE	SMOTE-SDAE	本文
ionosphere	57.38	56.48	78.42	78.57
wpbc	61.50	58.57	83.71	85.97
satimage	86.55	76.12	89.48	87.43
平均	68.48	63.72	83.87	83.99

表 4 各种算法的 G-means 对比(%)

数据集	SVM	SAE	SMOTE-SDAE	本文
ionosphere	60.22	68.12	76.15	76.30
wpbc	48.30	46.77	65.24	68.29
satimage	84.62	77.22	90.77	88.71
平均	64.38	64.04	77.39	77.77

从两个表可以看出, 本文所提出的方法在较低的非均衡比下作用在上述三个数据集中, 对于前两个数据集(ionosphere、wpbc), 使用的两个度量指标均优于其他算法, 而在数据集 staimage 中两个度量指标均低于SMOTE-SDAE算法. 实验结果表明: 该算法在处理小数据集的非均衡数据分类问题中较为有效, 但是处理数据集较大的非均衡问题时效果会弱于某些分类算法, 但是该算法总体的分类效果较好.

5 结语

对于不平衡数据分类问题, 本文结合神经网络与降噪自编码器提出了一种基于 BP 神经网络的非均衡数分类算法, 经过与其他算法比较表明该算法对于较小数据集的非平衡数据分类精度较其它方法有明显的提高, 当数据集中训练样本增多时, 该算法较深度学习方法优势不再明显.

参考文献

1 Zhang C, Gao W, Song J, et al. An imbalanced data classification algorithm of improved autoencoder neural network. Eighth International Conference on Advanced Computational Intelligence. 2016. 95–99.

2 章少平,梁雪春.优化的支持向量机集成分类器在非平衡数据集分类中的应用.计算机应用,2015,35(5):1306–1309.

3 许丹丹,王勇,蔡立军.面向不均衡数据集的 ISMOTE 算法.计算机应用,2011,31(9):2399–2401.

4 陶新民,童智靖,刘玉,等.基于ODR和BSMOTE结合的不均衡数据 SVM 分类算法.控制与决策,2011,26(10):1535–1541.

5 韩敏,朱新荣.不平衡数据分类的混合算法.控制理论与应用,2011,28(10):1485–1489.

6 Wang A, Liu L, Jin X, et al. Adapting TSVM for fault diagnosis with imbalance class data. Control and Decision Conference (CCDC). China. IEEE. 2016. 2919–2923.

7 王洪元,史国栋.人工神经网络技术及其应用.北京:中国石化出版社,2002.

8 夏战国,王珂,王志晓.基于网络复杂性的 BP 算法.计算机工程与设计,2008,29(15):3867–3869.

9 汪宝彬,汪玉霞.随机梯度下降法的一些性质(英文).数学杂志,2011,31(6):1041–1044.

10 Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders, Proc. of the Twenty-fifth International Conference on Machine Learning (ICML‘08). ACM. 2008. 1096–1103.

11 张枭山,罗强.一种基于聚类融合欠抽样的不平衡数据分类方法.计算机科学,2015,42(11A):63–66.

© 中国科学院软件研究所 http://www.c-s-a.org.cn