文章编号: 1000−8152(2011)10−1485−05

不平衡数据分类的混合算法

韩 敏, 朱新荣

(大连理工大学 电子信息与电气工程学部, 辽宁 大连 116024)

摘要: 针对传统分类算法处理不平衡数据时, 小类的分类精度过低问题, 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. 在小类样本之间用随机插值方式平衡数据集的分布, 利用受试者特征曲线在置信度为95%下的面积为标准去除冗余特征; 之后对输入数据用Bagging技术进行扰动, 并以径向基函数神经网络作为随机森林中的基分类器, 采用绝大多数投票方法进行决策的融合和输出. 将该算法应用于UCI数据, 以G均值和受试者特征曲线下的面积为评判标准, 结果表明该方法能够有效地提高中度和高度不平衡数据的分类精度.

关键词: 不平衡数据; 随机森林; 径向基函数神经网络; 受试者特征曲线

中图分类号: TP751 **文献标识码**: A

Hybrid algorithm for classification of unbalanced datasets

HAN Min, ZHU Xin-rong

(Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian Liaoning 116024, China)

Abstract: A novel hybrid algorithm of radial basis function neural network(RBFNN) integrated with the random forest algorithm is proposed to improve the poor classification result produced by traditional algorithm in classifying minor class of unbalanced datasets. Firstly, random interpolations are inserted between adjacent data in the minor dataset to balance the data distribution. Receiver operator characteristics(ROC) with degree of confidence less than 95% are considered the redundant characteristic and are deleted. The input data are perturbed by the Bagging technique. Radial Basis Function Neural Network is employed to be the basic classifier in the random forest. The fusion of decisions and the outputs are determined by the vast majority of votes. This method is applied to UCI dataset. The precision of G-mean and the area under the ROC demonstrate the improvement of the accuracy in the classifications of medium-size unbalanced and large-size unbalance class data sets.

Key words: imbalanced data; random forest; radial basis function neural network(RBFNN); receiver operator characteristics(ROC)

1 引言(Introduction)

不平衡分类问题是指数据样本中某些类的数量远远少于其他类时的分类研究. 虽然现有的分类研究可以取得较好的分类效果, 但是它们的缺陷在于假设训练的数据集是平衡的, 这一假设往往是不成立的. 在现实应用领域中存在不平衡数据集的情况大致有以下几种: 医疗诊断、信息检索、文本分类、石油泄漏[1~3]等, 在这些情况的处理过程中少数类识别的准确率更为重要. 传统的分类方法为保证分类总体精度, 通常将小类通过阈值或规则选择误分到大类中来保证整体的分类准确率[4], 从而导致少数类问题的研究和解决遇到重重困难.

鉴于不平衡分类研究的重要性, 国内外学者进行了大量的研究, 主要有以下4个方面: 1) 对已有的分类器通过改变分类阈值进行相关改进. 例如采用模糊规则对已有的智能算法和决策树算法规则的修改[5]或是在传统的分类算法基础上引入代价敏感因子进行改进[6]等; 2) 设计新的适应不平衡数据的分类方法, 例如采用新的向上采样方法或者向上向下结合的采样方法对少类数据的处理[7]; 3) 设计新的分类器性能评价准则, 通常是通过引入混淆矩阵, 比较其中的查全率, F值等来进行[8]; 4) 改变数据的分布, 常见的有随机向上采样, 随机向下采样或者是更加智能的采样方法, 其中最成熟的方法是采用少类样本合成重采样技术(synthetic minority over-sampling technique, SMOTE) 将不平衡数据通过插值来改变数据的分布情况[9]从而影响分类效果.

本文提出的基于随机森林 (random forest, RF)[10]和径向基函数的混合算法应用于不平衡数据, 利用SMOTE对少类数据进行扩充, 用受试者特征曲线下面积进行特征冗余度的去除之后再用于分类操作, 以G平均值, 受试者特征曲线下面积来评价

收稿日期: 2010−06−25; 收修改稿日期: 2010−11−02.

基金项目: 国家自然科学基金资助项目(61074096); 国家科技支撑计划资助项目(2006BAB14B05); 国家重点基础研究资助项目(2006CB403405).

最终分类精度. 结果表明, 该方法对不平衡数据具有较好的效果, 并且随着不平衡度的增加, 所提分类算法在各项指标中的精度更高, 效果也更显著.

2 小类的扩充和均一化处理(Minor data expansion and uniformization)

为了解决常用分类方法将小类误分为大类的问题, 考虑对输入的不平衡数据进行向上采样方法, 对输入数据中的小类进行扩充.

首先, 计算大类与小类之间的不平衡程度(imbalanced level, IL), 通过对小类样本进行随机插值来完成向上采样, 即SMOTE方法. 对于每个小类样本x, 找到对应该样本的k个最近邻同类样本, 根据向上采样的倍率N, 从该k个最近邻样本中随机选择N个样本, 记为$y_1, y_2, \cdots, y_N$; 在x与y_j之间进行随机插值, 形成新的小类样本c_j. 整个过程可由下式所示:

$$c_j = x + \mathrm{rand}(0,1) * (y_j - x),\ j = 1, 2, \cdots, N, \tag{1}$$

其中rand(0,1)表示区间(0,1)内的一个随机数. 通过随机插值, 直到小类样本数与大类的相差不多, 为了使结果与其他文献方法基准相同, 设置同样的采样倍率N. 根据数据集的不平衡程度IL来确定该值:

$$N = \mathrm{round}(\mathrm{IL}) - 1, \tag{2}$$

其中round(IL)表示对IL四舍五入得到的数值.

其次, 选用受试者特征[11](receiver operator characteristic, ROC)曲线对扩充后的平衡数据进行处理, 分析其中对分类无用的相关特征, 通过去除的方式降低数据的冗余度. 其基本原理是: 通过判断点的移动, 获得多对灵敏度和误判率, 以灵敏度为纵轴, 误判率为横轴, 连接各点绘制曲线. 其中灵敏度是把实际为真值的判断为真值的概率; 特异度是把实际为假值的判断为假值的概率; 误判率是把实际为假值的判断为真值的概率, 其值等于(1–特异度). 若二者相等即把实际为真判为真的概率与实际为假判为真的概率相等, 因此说斜45°线附近的曲线是对整体判断状态特征的一个冗余因子. 以置信度为95%时的渐近P值为约束条件, 用曲线下面积(area under the curve, AUC)与0.5的大小来进行特征的选取, 去除掉对状态特征相关性最小的特征值, 达到降低数据冗余度的目的.

最后, 为了消除由于量纲的不同所造成的影响, 对输入数据采用最大最小归一化方法, 将所有数据归一化到[–1,1]之间, 其公式如下所示:

$$y_i = \frac{2x_i - (x_{\min} + x_{\max})}{x_{\max} - x_{\min}}. \tag{3}$$

其中: y_i表示对应输入值归一化后的数值, $y_i \in [-1, 1]$; x_i表示对应该输入点的数值; $x_{\min}, x_{\max}$分别表示对应该输入数据的最大值和最小值.

3 随机森林与RBFNN的混合算法(The hybrid algorithm of random forest and RBFNN)

通过借用RF的框架结构, 对输入数据用Bagging技术对输入数据集进行扰动, 从原始数据集中进行n次重采样, n是原始训练数据集中实例被随机替换的个数. 森林中的基分类器最终通过均等权重的投票给出分类结果. 模型中以RBFNN作为森林中的基分类器对数据进行训练和测试. 该算法框架由多个单元组成, 如图1, 2所示, 其中$i = 1, 2, \cdots, n$.

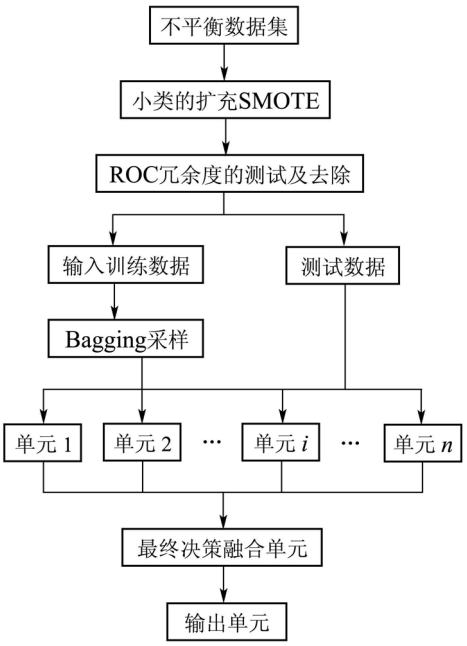

图 1 RF+RBFNN混合算法示意图

Fig. 1 Hybrid of RF and RBFNN algorithm

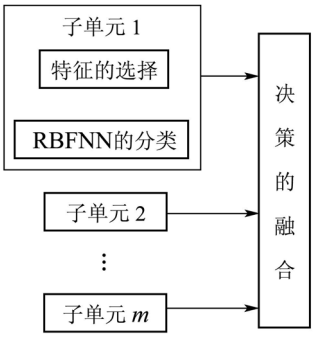

图 2 单元i的内部示意图

Fig. 2 The schematic diagram of unit i

该算法需要以下几个步骤来构成: 首先将输入的不平衡数据经过SMOTE方法的小类扩充之后再选用ROC曲线进行冗余度的测试, 比较AUC与0.5之间的大小, 用渐近P值作为判断标准, 去除与状态特征相关性最弱的特征, 即在置信度为95%以内的特

征得到的AUC值处于0.5附近需要进行去除, 以此来达到对输入特征中冗余特征去除的目的.

其次将输入的训练数据用Bagging方法, 通过有放回的抽样方式生成n组数据, 将这每一组数据都用来训练m个由RBFNN构成的基分类器. 训练好之后, 每个测试数据都送入由子分类器集构成的模型, 由这m个模型给出各自的结果, 进行投票, 即给出单个模型的决策融合结果. 再将这n个单元的结果进行多数投票方法, 给出最终测试数据的类别归属.

最后由于绝大多数投票方法对于数据的类型要求较少, 而且适用的条件也相对宽松, 因此最终的决策融合单元采用绝大多数投票[12]方法进行. RF+RBFNN混合算法包含了几个中间的分类步骤, 这几个分类步骤的输出共同构成了一个向量.

假设该混合模型需要解决的是一个n类的模式分类问题, 集成的规模为N, 各模式类分别记作c_1, $c_2, \cdots, c_n$, 类别i的输出编码为$[0, \cdots, 0, 1, 0, \cdots 0]$, 即除第$i$个元素为1外, 其余元素全为0. 通过期望的输入输出编码映射关系对N个成员网络进行训练. 训练之后对于测试模式P, 每个成员网络给出对应的一个输出列向量, 第k个成员网络的输出为$[x_{1,k} \quad x_{2,k} \quad \cdots \quad x_{n,k}]^{\mathrm{T}}$, $k = 1, \cdots, N$, 由投票法得到该N个网络的集成输出如式(4)所示:

$$\boldsymbol{P}_{\text{ensem\_out}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} + x_{1,2} + \cdots x_{1,N} \\ x_{2,1} + x_{2,2} + \cdots x_{2,N} \\ \vdots \\ x_{n,1} + x_{n,2} + \cdots x_{n,N} \end{bmatrix}. \quad (4)$$

上述矩阵中各向量的组合方式, 一般可分为多数通过、一致通过和无反对票3种形式. 采用绝大多数投票原则, 即如果$x_{i,k} = \max\limits_{j}(x_{j,k})$, 则$x_{i,k} = 1$且$x_{j,k} = 0, j \neq i$. 那么当$y_j = \max\limits_{i}(y_i)$时, 模式$P$就被划分到第$j$类$(c_j)$.

4　仿真结果及分析(Simulation and analysis)

通常情况下, 将不平衡度在[1.5, 3.5)之间的称之为低度不平衡范围; 在[3.5, 9.5)之间的称之为中度不平衡范围; 在$[9.5, +\infty)$之间的称之为高度不平衡范围. 选取UCI[13]数据库中的12组数据, 按不平衡度大小排列, 如表1所示. 对于多类标的数据集, 采用与其他文献类似的方法: 将其中的一类作为小类, 其他的看作一个整体成为大类. 以Glass 3为例, 即认为将Glass数据中的第3类标(ve-window-float)作为小类, 其他类标作为一个整体看成一个大类, 最终的不平衡度为10.39. 冗余特征一栏中横线表示经过SMOTE扩充后, 各特征值与状态变量相关性强, 且渐近P值(asymptotic P value)小于0.5, 在置信度95%范围内, 无需去除. 对应的V4, V5等表示应当考虑去除的特征是该组数据中第4和5特征变量. 以 Ecoli4 数据为例, 需要去除的特征 V4, 渐近P值0.644, AUC值0.511; 去除特征 V5, 渐近P值0.945, AUC值0.498.

为比较方便起见, 实验中对数据采用五折交叉检验(5-fold cross-validation)方式, 为保证数据组在进行分组的过程中不平衡度一致, 采用分层采样, 即: 将数据集中的小类和大类样本分别随机分为5等份, 两两随机组合得到5个大小相差不多的子集. 将其中1份作为测试集, 其余4个子集作为训练集. 重复5次仿真, 以平均值作为最终的分类结果. 采用G均值来评价分类的好坏, 其公式如下:

$$G - \text{mean} = \sqrt{\text{acc}^{+} * \text{acc}^{-}}, \quad (5)$$

其中: acc^{+}表示小类的分类精度, acc^{-}表示大类的分类精度.

以表1中的12组UCI数据集作为比较对象, 所提算法与以下文献中的算法进行比较: 文献[14]中通过判断小类样本的真假并修改假小类样本标签的方式对Adaboost进行的改进方法Ada+B; 文献[15]中基于语言规则迭代方法的自适应推断系统Chi3+AIS算法; 文献[16]中通过全局配置语义型的基于机器学习的模糊集成迭代算法FH+GBML+LTR; 文献[17]中基于合作与竞争的进化RBFNN 模型在SMOTE扩充小类之后的算法CO2RBFN; 文献[18]中利用随机向上采样和 Boost 技术结合的方式提出的算法RUSBoost. 对上述预处理过的UCI数据进行仿真, 结果依据不平衡度大小排列, 为方便结果查看, 选用4个文献中UCI重叠率最高的9组进行对比, 如表2所示. 对应每一组UCI数据, 用黑色粗体表示其中精度最高的一个, 表格的最后一行是不同算法在这9组UCI数据中的平均值, 表格的省略号表示文献中未给出对应的值.

由表2的结果来看, 所提方法RF+RBFNN在低度不平衡范围条件下要比Ada+B, Chi3+AIS和CO2-RBFN方法好, 与FH+GBML+LTR方法相比, 两者结果相差不多, 最高精度相差不超过2%. 随着不平衡度的增加, 在中度不平衡和高度不平衡条件下, 除了Abalone9～18数据集外, 所提方法都要比另外的4种方法得到的精度要高出许多, 尤其是在Glass 5数据集中高出了约12%的精度. 此外, 通过计算这9组UCI数据在不同算法条件下的平均值, 可以看出所提方法要明显的优于其余的4种算法, 平均值达到88.5033%, 其次是融合了模糊集理论与统计学非参数检验方法的FH+GBML+LTR方法, 其平均值为84.71%.

此外, 对于不平衡数据分类效果的评价, 还常用到AUC值进行比较, 现以FH+GBML+LTR方法,

Ada+B方法以及RUSBoost方法与所提方法进行比较, 结果按不平衡度程度排列, 如表3所示. 为查看结果方便起见, 选用3个文献中UCI重叠率最高的7组进行对比. 从表3的结果可以看出, 在这7组UCI数据中, 除了Vehicle0数据集在FH+GBML+LTR中的AUC较高, Segment1在RUSBoost方法中AUC数值较高外, 所提方法在低度、中度和高度不平衡范围条件下, AUC值都远高于另外3种算法, 尤其是在Glass3的处理中, 所提方法在AUC的值要比FH+GBML+LTR高出22%左右. AUC的值越高, 说明对应的ROC曲线越靠近纵轴, 对应分类器的效果也就越好, 该分类器的泛化性能也就越高.

表 1 仿真中采用的12组UCI数据集

Table 1 The 12 UCI datasets used in the experiment comparation

编号	原数据集	类标小类: 类标大类	小类比例: 大类比例	不平衡度(IL)	冗余特征
1	Diabetes	tested-positive: tested-negative	34.84: 66.16	1.90	—
2	Credit-G	bad: good	30.00: 70.00	2.33	V4, V10, V11, V15, V18
3	Vehicle0	opel: remainder	25.06: 74.94	2.99	—
4	Hepatitis	Diel: Live	20.64: 79.36	3.84	V9, V10, V16
5	New-thyroid2	hyper: remainder	16.28: 83.72	5.14	—
6	Ecoli3	pp: remainder	15.48: 84.52	5.46	V1, V4, V5, V7
7	Segment1	brickface: remainder	14.26: 85.74	6.01	V3
8	Ecoli4	Imu: remainder	10.88: 89.12	8.19	V4, V5
9	Glass3	ve-window-float: remainder	8.78: 91.22	10.39	—
10	Glass5	containers: remainder	6.07: 93.93	15.47	—
11	Abalone9-18	18: 9	5.65: 94.25	16.68	—
12	Abalone19	19: remainder	0.77: 99.23	128.87	—

表 2 不同算法在9组UCI数据条件下G均值的精度比较

Table 2 The comparation of G-means among 9 UCI datasets used in different algorithms

原数据集	不平衡度	Ada+B [14]	Chi3+AIS [15]	FH+GBML+LTR [16]	CO2RBFN [17]	RF+RBFNN+ROC
Diabetes	1.90	59.32	68.34	**74.57**	—	72.6834
Credit-G	2.33	53.75	—	—	—	**79.0769**
Vehicle0	2.99	55.41	71.60	73.09	—	**74.4094**
Hepatitis	3.84	62.86	—	—	—	**89.243**
New-thyroid2	5.14	—	93.40	94.01	98.01	**98.5915**
Ecoli3	5.46	—	89.42	88.26	93.14	**93.75**
Glass5	15.47	—	86.09	86.33	—	**97.9798**
Abalone9∼18	16.68	—	59.18	**74.28**	75.34	73.913
Abalone19	128.87	—	55.15	67.64	70.18	**77.1788**
平均值	—	57.853	80.8393	84.7100	84.1657	**88.5033**

表 3 7组UCI数据在4种不同算法条件下AUC的比较

Table 3 The comparation of AUC among 7 UCI datasets used in different algorithms

编号	原数据集	不平衡度(IL)	Ada+B[14]	FH+GBML+LTR[16]	RUSBoost[18]	RF+RBFNN+ROC
1	Diabetes	1.90	77.66	74.5	—	**79.1**
2	Credit-G	2.33	75.67	—	—	**83.2**
3	Vehicle0	2.99	86.13	**92.88**	83.59	90.7
4	Hepatitis	3.84	83.46	—	—	**94.1**
5	Segment1	6.01	—	99.04	**99.64**	99.1
6	Ecoli4	8.19	—	90.90	93.20	**94.9**
7	Glass3	10.39	—	59.65	76.20	**82.6**
平均值	—	—	81.23	83.394	88.1575	**92.2875**

5 总结(Conclusions)

所提出的RF+RBFNN混合算法能够有效的处理不平衡数据的问题, 该方法通过对小类样本的扩充使得整体数据达到较为均衡的状态, 然后采用针对小类的神经网络集成学习方法进行训练和测试, 最终结果表明, 该算法可以有效提高小类样本的分类性能.

但是由于数据集本身的多样性和复杂性, 样本的分布也呈现多样性, 比如小类内部可能会聚为多个小簇的情况, 即使通过SMOTE方法进行扩充, 也无助于改进小类样本过于集中的情况. 如果能实现估计小类样本潜在的分布, 根据不同的分布设置不同的合成新样本的方式, 对小类的分类性能将会提高更多. 此外, 将向下采样的智能技术和SMOTE进行结合, 进一步改善对小类的分类性能也是下一步的工作内容.

参考文献(References):

[1] KUBAT M, HOLTE R C, MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. *Machine Learning*, 1998, 30(2): 195 – 215.

[2] MAZUROWSKI M A, HABAS P A, ZURADA J M, et al. Training neural network classifiers for medical decision making:the effects of imbalanced datasets on classification performance[J]. *Neural Networks*, 2008, 21(2/3): 427 – 436.

[3] 朱卫, 沈玉琨. 乳腺癌普查资料的分析[J]. 疾病控制杂志, 2002, 6(3): 253 – 254.
(ZHU Wei, SHEN Yukun. An analysis of breast cancer discovered through census[J]. *Chinese Journal of Disease Control Prevention*, 2002, 6(3): 253 – 254.)

[4] SUN Y M, KAMEL M S, WONG A K C, et al. Cost-sensitive boosting for classification of imbalanced data[J]. *Pattern Recongnition*, 2007, 40(12): 3358 – 3378.

[5] SUN A X, LIM E P, LIU Y. On strategies for imbalanced text classification using SVM: a comparative study[J]. *Decision Support Systems*, 2009, 48(1): 191 – 201.

[6] RASKUTTI B, KOWALCZYK A. Extreme rebalancing for SVMs: a case study[J]. *Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2004, 6(1): 61 – 69.

[7] YEN S J, LEE Y S. Cluster-based under-sampling approaches for imbalanced data distributions[J]. *Expert Systems with Applications*, 2009, 36(3): 5718 – 5727.

[8] BAE M H, WU T, PAN R. Mix-ratio sampling: classifying multiclass imbalanced mouse brain images using support vector machine[J]. *Expert Systems with Applications*, 2010, 37(7): 4955 – 4965.

[9] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. *Journal of Artificial Intelligence Research*, 2002, 16(2): 341 – 378.

[10] BREIMAN L. Random forests[J]. *Machine Learning*, 2001, 45(1): 5 – 32.

[11] BRADLEY A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. *Pattern Recognition*, 1997, 30(7): 1145 – 1159.

[12] LAM L, SUEN C Y. Application of majority voting to pattern recognition: an analysis of its behavior and performance[J]. *IEEE Transactions on Systems, Man and Cybernetics–Part A: Systems and Humans*, 1997, 27(5): 553 – 568.

[13] BLAKE C L, MERZ C J. UCI repository of machine learning databases[OL/DB]. 1998. http://www.ics.uci.Edu/ mlearn/MLR–epository.html.

[14] 郭乔进, 李立斌, 李宁. 一种用于不平衡数据分类的改进AdaBoost算法[J]. 计算机工程与应用, 2008, 44(21): 217 – 221.
(GUO Qiaojin, LI Libin, LI Ning. Novel modified AdaBoost algorithm for imbalanced data classification[J]. *Computer Engineering and Applications*, 2008, 44(21): 217 – 221.)

[15] FERNANDEZ A, DEL JESUS M J, HERRERA F. On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced datasets[J]. *Expert Systems with Applications*, 2009, 36(6): 9805 – 9812.

[16] FERNANDEZ A, DEL JESUS M J, HERRERA F. On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced datasets[J]. *Information Science*, 2010, 180(8): 1268 – 1291.

[17] PEREZ-GODOY M D, RIVERA A J, FERNANDEZ A, et al. A preliminar analysis of CO2RBFN in imbalanced problems[C] //*Proceedings of the 10th International Work-Conference on Artificial Neural Networks*. Berlin, Heidelberg: Springer, 2009: 57 – 64.

[18] SEIFFERT C, KHOSHGOFTAAR T M, VAN HULSE J, et al. RUSBoost: a hybrid approach to alleviating class imbalance[J]. *IEEE Transactions on Systems, Man and Cybernetics–Part A: Systems and Humans*, 2010, 40(1): 185 – 197.

作者简介:

韩　敏　(1959—), 女, 教授, 博士生导师, 研究方向为神经网络、混沌序列分析以及它们在控制和识别方面的应用, E-mail: minhan@dlut.edu.cn;

朱新荣　(1986—), 男, 硕士研究生, 研究方向为遥感影像的分类和数据处理.