第37卷 第15期 计 算 机 工 程 2011年8月
Vol.37 No.15 Computer Engineering August 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)15—0122—03 文献标识码：A 中图分类号：TP393

一种改进的不平衡数据集分类方法

赵秀宽[1]，阳建宏[2]，黎 敏[2]，徐金梧[2]

(1. 中国科学院地质与地球物理研究所，北京 100029；2. 北京科技大学机械工程学院，北京 100083)

摘 要：传统的机器学习方法在解决不平衡分类问题时，得到的分类器具有很大的偏向性，表现为少数类识别率远低于多数类。为此，在旋转森林分类方法的基础上，提出一种改进的不平衡数据处理方法——偏转森林。通过对少数类进行过抽样改变训练数据的分布以减小数据的不平衡，采用随机抽取的方式确保生成偏转矩阵的样本间存在差异，从而提高集成分类器的分类精度。实验结果表明，该方法能取得较好的分类性能，具有较高的少数类识别正确率和较低的多数类识别错误率。

关键词：不平衡数据集；偏转森林；集成分类器；过抽样

Improved Unbalanced Dataset Classification Method

ZHAO Xiu-kuan[1], YANG Jian-hong[2], LI Min[2], XU Jin-wu[2]

(1. Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China;
2. School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China)

【Abstract】Referring to unbalanced dataset, the traditional machine learning methods will achieve biased performance. Using traditional methods, the recognition rate of minority class will be lower than the recognition rate of majority class. In this paper, based on rotation forest, it proposes an improved unbalanced dataset learning method, which is called deflection forest. It reduces data unbalance by over-sampling the data of minority class. It uses random resampling to increase diversity of samples which generate the deflection matrix and then improves the accuracy rate. Experimental results show that the deflection forest method achieves better performance, which carries out higher recognition rate of minority class and also lower recognition error rate of majority class.

【Key words】unbalanced dataset; deflection forest; integrated classifier; over-sampling

DOI: 10.3969/j.issn.1000-3428.2011.15.038

1 概述

在许多实际的模式分类问题中，存在大量数据不平衡分类问题，用传统的分类方法对少数类的分类比较困难。针对数据不平衡问题的相关研究主要围绕以下 3 个方面展开：(1)改变数据的分布；(2)设计新的分类方法；(3)设计新的分类器性能评价准则。本文提出的偏转森林是一种通过改变数据的分布提高不平衡数据分类精度的方法。

2 不平衡分类方法

所谓不平衡分类问题是指训练样本数量在类间分布不平衡的模式分类问题，具体地说是某些类的样本数量远少于其他类。当传统的机器学习方法用于解决这些不平衡分类问题时，通常出现分类器性能的大幅度下降，得到的分类器具有很大的偏向性，表现为少数类的识别率远低于多数类。而在大部分不平衡分类问题中，少数类是分类的重点，即正确识别出少数类的样本比识别出多数类的样本更有价值。不平衡数据集分类问题的解决方法分为数据层的方法和算法层的方法。数据层的方法又称为重抽样，它直接对训练集进行操作，处理后的样本用来训练分类器。算法层的方法对分类算法进行操作，即修改已有的分类算法或者提出新的算法。对已有的算法，通过调节不同类样本之间的成本函数、改变概率密度、调整分类边界等措施使其更有利于少数类的分类。

重抽样分为过抽样和欠抽样。最简单的过抽样通过复制少数类样本增加少数类的规模，缺点是没有增加任何新的知识，使分类器学到的决策域变小，从而导致过学习。改进的过抽样方法通过在少数类中加入随机高斯噪声或产生新的合成样本等，一定程度上解决了上述问题，如文献[1]提出的SMOTE(Synthetic Minority Over-sampling Technology)算法。最简单的欠抽样方法随机地去除某些多数类样本减小多数类的规模，缺点是会丢失多数类的一些重要信息。文献[2]提出一种粒度支持向量机学习算法，根据粒度计算思想对多数类样本进行粒划分并从中获取信息粒，以使数据趋于平衡，其方法本质上是一种欠抽样方法。

算法层的方法主要通过引入代价敏感函数修改已有的分类算法或者提出新的算法。文献[3]提出的 Metacost 是一种代价敏感学习方法，通过赋予各个类别不同的错分代价，对错分少数类样本做更大惩罚，迫使最终分类器对少数类样本有更高的识别率。代价敏感学习能有效地提高少数类的识别率，但是在多数情况下，真实的错分代价很难被准确地估计。文献[4]提出一种新的不平衡处理算法，通过对偏最小二乘分类器进行改进，改善了分类器在处理不平衡数据时的性能。

集成学习通过综合多个分类器的输出，可以得出更准确的分类模型。目前，基于集成学习的不平衡数据算法大多是基于 AdaBoost 进行的改进算法，SMOTEBoost[5]是结合

基金项目：国家自然科学基金资助项目(50705069, 50905013, 50934007)；高等学校博士学科点专项科研基金资助项目(20090006120007)；中央高校基本科研业务费专项基金资助项目(FRF-TP-09-014A)

作者简介：赵秀宽(1982－)，男，工程师、博士，主研方向：模式识别，智能监测；阳建宏，副教授、博士；黎 敏，讲师、博士；徐金梧，教授、博士

收稿日期：2011-02-17 **E-mail：**kevin2008.zhao@gmail.com

SMOTE 和 Adaboost 方法，本质上是一种重采样方法；AdaCost 通过在 Boosting 算法的权值更新规则中引入每个训练样本的错分代价，提高 Boosting 算法对少数类的查全率和查准率，本质上是代价敏感学习方法。另外，文献[6]提出的 asBagging 算法是对 Bagging 集成学习算法的改进。

3　偏转森林

旋转森林[7]是一种新提出的多分类器集成系统，其特点在于使用线性变换方法生成旋转矩阵，使数据可以投影到不同坐标系中，从而构建有差异的分类器。其基本思路是随机打乱原始数据集的特征排序，并分割成若干个子集，然后对每个子集分别进行线性变换，再将得到的变换分量重新按这些特征子集原来对应的顺序合并。这样，每次随机分割后得到的数据都被投影到不同坐标空间中，因此形成差别较大的数据子集。当用这些数据子集训练分类器时，能够得到差异度较大的分类器。由于可以构建更加多样的基分类器，因此其性能优于 Bagging、AdaBoost 等多分类器系统。

但是旋转森林是基于平衡数据构建多分类器，在处理不平衡数据时效果较差。本文在旋转森林分类方法的基础上，提出偏转森林的分类方法，以处理不平衡数据的分类问题。给定训练集 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$，其中，$N$ 是训练集的大小；$\boldsymbol{x}_i$ 由 n 个特征组成的向量，可以表示为 $\langle x_{i,1}, x_{i,2}, \cdots, x_{i,n}\rangle$；$y_i$ 取自类别标签 $\{\omega_1, \omega_2, \cdots, \omega_c\}$。定义 $C_1, C_2, \cdots, C_L$ 为偏转森林的基分类器，其中，L 是偏转森林中基分类器的个数。定义 $\boldsymbol{F}$ 为 n 维特征集向量。基分类器 C_l 的训练集构建具体如下：

(1)利用过抽样方法 SMOTE 对数据集矩阵 $\boldsymbol{X}$ 中的少数类进行过抽样得到新的数据集矩阵 $\boldsymbol{X}_l$，过抽样参数为 T%。即按式(1)在 $\boldsymbol{x}_i$ 与其 P 个近邻间产生新的合成样本，合成样本个数与原有少数类样本个数比值为 T%：

$$synthetic_p = \boldsymbol{x}_i + \boldsymbol{r}_p \times \boldsymbol{dif}_p \quad p = 1, 2, \cdots, P \tag{1}$$

其中，$\boldsymbol{r}_p$ 是(0,1)之间的随机数组成的 n 维向量；$\boldsymbol{dif}_p$ 是 $\boldsymbol{x}_i$ 与第 p 个近邻的矢量差。重复使用式(1)直至合成足够多的样本数，即达到过抽样参数 T%。

(2)将特征集 $\boldsymbol{F}$ 随机分为 K 个子集，子集之间可以是相交的也可以是不相交的。本文假设各个特征子集不相交，且分别包含 $M = n/K$ 个特征。假设 $\boldsymbol{F}_{l,k}$ 为基分类器 C_l 的第 k 个特征子集。对应于特征子集 $\boldsymbol{F}_{l,k}$，从新数据集 $\boldsymbol{X}_l$ 中选择相应的数据子集 $\boldsymbol{X}_{l,k}$。

(3)在数据子集 $\boldsymbol{X}_{l,k}$ 上随机选择 50%的数据集，通过主成分分析(Principal Component Analysis, PCA)提取其主成分系数 $\boldsymbol{a}_{l,k}^{(1)}, \boldsymbol{a}_{l,k}^{(2)}, \cdots, \boldsymbol{a}_{l,k}^{(M_k)} (M_k \leqslant M)$。采用随机抽取 50%数据样本的方式，能进一步确保用于生成偏转矩阵的样本间存在差异。

(4)由得到的主成分系数组成基分类器 C_l 的偏转矩阵 $\boldsymbol{D}_l$：

$$\boldsymbol{D}_l = \begin{bmatrix} \boldsymbol{a}_{l,1}^{(1)}, \boldsymbol{a}_{l,1}^{(2)}, \cdots, \boldsymbol{a}_{l,1}^{(M_1)} & [0] & \cdots & [0] \\ [0] & \boldsymbol{a}_{l,2}^{(1)}, \boldsymbol{a}_{l,2}^{(2)}, \cdots, \boldsymbol{a}_{l,2}^{(M_2)} & \cdots & [0] \\ \vdots & \vdots & & \vdots \\ [0] & [0] & \cdots & \boldsymbol{a}_{l,K}^{(1)}, \boldsymbol{a}_{l,K}^{(2)}, \cdots, \boldsymbol{a}_{l,K}^{(M_K)} \end{bmatrix} \tag{2}$$

其中，偏转矩阵 $\boldsymbol{D}_l$ 的维数是 $n \times \sum_k M_k$。为了对应原始特征的顺序，将 $\boldsymbol{D}_l$ 的列数据重新进行排列得到新的偏转矩阵 $\boldsymbol{D}_l^a$。由以上步骤可以获得基分类器 C_l 的训练集 $\boldsymbol{X}_l\boldsymbol{D}_l^a$，$\boldsymbol{X}_l$ 为包含合成样本的训练数据集。

本文选择决策树 C4.5 训练基分类器，在训练阶段由训练集 $\boldsymbol{X}_l\boldsymbol{D}_l^a$ 得到基分类器 C_l。在分类阶段，给定一个测试样本 $\boldsymbol{x}$，分别求取 L 个基分类器的输出，并对输出结果取平均得到分类器的总输出 $\mu_j(\boldsymbol{x})$：

$$\mu_j(\boldsymbol{x}) = \frac{1}{L}\sum_{l=1}^{L} d_{l,j}\left(\boldsymbol{x}\boldsymbol{D}_l^a\right) \quad j = 1, 2, \cdots, c \tag{3}$$

在式(3)中，$d_{l,j}\left(\boldsymbol{x}\boldsymbol{D}_l^a\right)$ 为分类器 C_l 赋予的 $\boldsymbol{x}$ 属于类别 ω_j 的概率。偏转森林的总输出 $\mu_j(\boldsymbol{x})$，表示 $\boldsymbol{x}$ 属于类别 ω_j 的概率，由对所有的基分类器的结果取平均得到。最终 $\boldsymbol{x}$ 的类别标签为取得最大概率的类别。

偏转森林方法比旋转森林的优势在于其通过对少数类进行过抽样改变了训练数据的分布减小数据的不平衡，每个基分类器合成样本的差异性使构成的偏转矩阵多样性更强，这种处理方法提高了基分类器的多样性，从而提高了分类精度。

4　不同分类器方法对比

本文选取 3 种具有代表性的不平衡数据学习方法与偏转森林方法进行对比，不平衡数据学习方法分别是：SMOTE 方法，SMOTEBoost 方法和代价敏感学习 MetaCost 方法。同时，基于平衡数据的旋转森林方法被用来与偏转森林方法进行对比。实验中使用 UCI 标准数据集，对 5 个不同对象的不平衡数据集进行测试。各数据集的信息如表 1 所示，其中，不平衡数据集 Iris3 的含义是对于 UCI 标准 Iris 数据集(有 3 个类、4 个特征、150 个样本)，将其中的第 3 类 50 个样本作为少数类，其余合成一类作为多数类，构成不平衡数据集。

表 1　不平衡数据集信息

名称	特征个数	少数类样本数	多数类样本数	多数类样本数/少数类样本数
Iris3	4	50	100	2.0000
Vowel0	13	90	900	10.0000
Segment1	19	330	1 980	6.0000
Glass7	9	29	185	6.3793
Letter17	16	734	19 266	26.2480

将实验数据随机均分为 10 组，采用交叉验证的方法，每个结果均是独立进行 100 次实验取平均值获得。ROC (Receiver Operating Characteristic)曲线能够全面描述分类器的性能，已成为不平衡数据分类器性能评价的准则，ROC 曲线越靠近左上角表示分类器性能越好。图 1 给出了 Iris3 不平衡数据集的 ROC 曲线对比。偏转森林、SMOTE 和 SMOTEBoost 的过抽样参数依次设为 200、300、400、500、600、700、800、900、1 000、2 000、3 000，MetaCost 的惩罚因子参数依次设为 4、5、6、7、8、9、10、20、30、40、50。

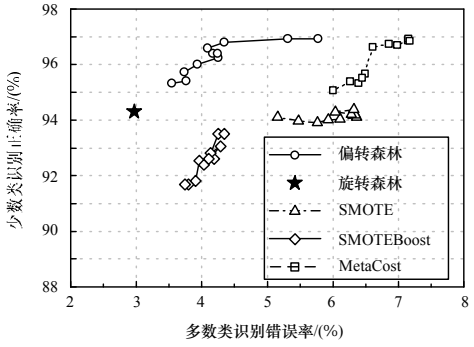

图 1　Iris3 数据集的 ROC 曲线

从图 1 可以看出，本文提出的偏转森林方法具有最好的分类性能，具有最高的少数类识别正确率，同时很好地控制多数类的识别错误率。

其他 4 个不平衡数据集在不同方法下的分类结果对比见表 2，每个结果均是独立进行 100 次实验取平均值后获得。

表 2 不平衡数据集在不同方法下的分类结果对比

数据集	方法	过抽样参数/惩罚因子	少数类正确率	多数类错误率
Vowel0	偏转森林	1 000	0.992 3	0.002 0
		2 000	0.993 0	0.002 3
		4 000	0.993 6	0.002 9
	旋转森林	-	0.978 6	0.000 4
	MetaCost	10	0.938 3	0.027 6
		50	0.940 2	0.044 4
		100	0.949 0	0.065 6
	SMOTEBoost	1 000	0.978 0	0.003 6
		2 000	0.982 8	0.003 9
		4 000	0.983 4	0.004 3
	SMOTE	1 000	0.957 1	0.012 8
		2 000	0.966 4	0.012 2
		4 000	0.963 6	0.012 8
Segment1	偏转森林	1 000	0.991 3	0.000 4
		2 000	0.991 8	0.000 6
		4 000	0.992 2	0.000 7
	旋转森林	-	0.987 2	0.000 4
	MetaCost	10	0.977 5	0.009 7
		50	0.979 0	0.011 4
		100	0.981 0	0.015 5
	SMOTEBoost	1 000	0.991 5	0.001 6
		2 000	0.990 9	0.001 7
		4 000	0.991 8	0.001 9
	SMOTE	1 000	0.987 3	0.004 1
		2 000	0.987 6	0.004 4
		4 000	0.987 8	0.004 9
Glass7	偏转森林	1 000	0.884 1	0.029 7
		2 000	0.896 6	0.037 0
		4 000	0.903 8	0.045 4
	旋转森林	-	0.813 4	0.015 1
	MetaCost	10	0.898 6	0.053 3
		50	0.899 0	0.068 4
		100	0.904 5	0.196 7
	SMOTEBoost	1 000	0.870 3	0.025 7
		2 000	0.872 1	0.027 9
		4 000	0.872 8	0.027 3
	SMOTE	1 000	0.884 5	0.047 4
		2 000	0.886 6	0.053 2
		4 000	0.886 6	0.045 9
Letter17	偏转森林	1 000	0.961 5	0.000 6
		2 000	0.968 1	0.000 8
		4 000	0.973 3	0.001 1
	旋转森林	-	0.912 8	0.000 2
	MetaCost	10	0.926 8	0.006 3
		50	0.935 5	0.009 9
		100	0.950 2	0.025 1
	SMOTEBoost	1 000	0.955 1	0.000 5
		2 000	0.955 1	0.000 5
		4 000	0.957 0	0.000 6
	SMOTE	1 000	0.921 4	0.003 2
		2 000	0.922 1	0.003 4
		4 000	0.928 0	0.003 8

由表 2 数据对比可知，本文方法对不平衡数据集的处理取得了最好的分类效果。如对 Glass7 不平衡数据集，偏转森林在过抽样参数值为 4 000 时，少数类的正确率达到 0.903 8，多数类的错误率仅为 0.045 4。

在相同参数下，SMOTEBoost 少数类的正确率 0.872 8，多数类的错误率为 0.027 3，SMOTE 方法少数类的正确率 0.886 6，多数类的错误率为 0.045 9。MetaCost 方法在惩罚因子参数值为 100 时，少数类准确率为 0.904 5，多数类的错误率已经达到 0.196 7。

旋转森林的少数类的正确率为 0.813 4，多数类的错误率为 0.015 1。在其他不平衡数据集中，偏转森林方法也都取得了很好的分类效果。

5 结束语

本文提出一种不平衡数据集分类新方法，其优势在于通过对少数类进行过抽样改变训练数据的分布从而减小数据的不平衡；同时每个基分类器合成样本的差异性使构成的偏转矩阵多样性更强，这种处理方法提高了基分类器的多样性，从而提高集成分类器的分类精度。

将本文方法应用于 5 个不同对象的不平衡数据集，与其他不平衡数据分类方法进行对比，结果表明本文分类方法具有更好的分类效果。

参考文献

[1] 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(1): 321-357.

[2] 郭虎升, 亓 慧, 王文剑. 处理非平衡数据的粒度 SVM 学习算法[J]. 计算机工程, 2010, 36(2): 181-183.

[3] Domingos P. MetaCost: A General Method for Making Classifiers Cost-sensitive[C]//Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining. San Diego, California, USA: ACM Press, 1999: 155-164.

[4] Qu Haini, Li Guozheng, Xu Weisheng. An Asymmetric Classifier Based on Partial Least Squares[J]. Pattern Recognition, 2010, 43(10): 3448-3457.

[5] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving Prediction of the Minority Class in Boosting[C]// Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Cavtat-Dubrovnik, Croatia: [s. n.], 2003: 107-119.

[6] Li Guozheng, Meng Haohua, Lu Wencong, et al. Asymmetric Bagging and Feature Selection for Activities Prediction of Drug Molecules[C]//Proceedings of Symposium of Computations in Bioinformatics and Bioscience. Iowa City, Iowa, USA: [s. n.], 2008: 22-30.

[7] Rodriguez J J, Kuncheva L I, Alonso C J. Rotation Forest: A New Classifier Ensemble Method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(10): 1619-1630.

编辑 陆燕菲