第37卷 第14期 计 算 机 工 程 2011年7月
Vol.37 No.14 Computer Engineering July 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)14—0178—02 文献标识码：A 中图分类号：TP391

Bagging组合的不平衡数据分类方法

秦姣龙，王 蔚

(南京师范大学教育科学学院机器学习与认知实验室，南京 210097)

摘 要：提出一种基于Bagging组合的不平衡数据分类方法CombineBagging，采用少数类过抽样算法SMOTE进行数据预处理，在此基础上利用C-SVM、径向基函数神经网络、Random Forests 3种不同的基分类器学习算法，分别对采样后的数据样本进行Bagging集成学习，通过投票规则集成学习结果。实验结果表明，该方法能够提高少数类的分类准确率，有效处理不平衡数据分类问题。

关键词：Bagging组合；不平衡数据分类；支持向量机；神经网络；Random Forests算法

Imbalanced Data Classification Method for Bagging Combination

QIN Jiao-long, WANG Wei

(Machine Learning and Cognition Lab, School of Education Science, Nanjing Normal University, Nanjing 210097, China)

【Abstract】CombineBagging is designed as a new classification method based on bagging combination for imbalanced data. The main points are as follows: using three different base classifiers learning algorithms, such as C-SVM, Radial Basis Function(RBF) neural network and random forests, to carry out bagging ensemble learning respectively, integrating the three different learning results above into one as the final result by applying voting rule. Experimental results show that CombineBagging method can enhance the minority data's classification accuracy rate on the five different regions imbalanced data. It is proved that the method can deal with the problem of imbalanced data.

【Key words】Bagging combination; imbalanced data classification; Support Vector Machine(SVM); neural network; Random Forests algorithm

DOI: 10.3969/j.issn.1000-3428.2011.14.059

1 概述

在机器学习和数据挖掘研究中，通常假设各类别训练样本的数目是大致相同的，但这一假设在实际问题中通常是不成立的，因为真实情况下会出现类不平衡问题，即由于训练集中某一类样本的数量相对于其他类别过少，导致分类器倾向于将测试样本全部判别为多数类，因此分类器在少数类上表现出较高的错误率。对于少数类的错分，在医学数据分类、信用卡交易、基因序列中编码信息发现、对涉及恐怖信息和安全信息的文本分类及雷暴天气预报[1]等领域，很可能意味着丢失了重要的信息，这将比错分多数类时所付出的代价更大。因此，如何有效地对不平衡数据进行分类是当今机器学习和数据挖掘领域研究的一个热点问题。目前，不平衡数据分类问题的解决策略分为两大类[2]：一类是从训练集入手，改变训练集样本分布，降低不平衡程度，如使用各种采样方法和训练集划分方法；另一类是从学习算法入手，根据算法在解决不平衡分类问题时的缺陷，适当修改算法使之适应不平衡分类问题，如分类器集成和代价敏感学习。在这些方法中，集成学习方法不仅可以提高分类的正确率，而且可以提高基分类器的泛化能力，对于处理不平衡数据的分类问题相当有效。Bagging作为一种集成学习的方法，被广泛用于不平衡数据分类中。文献[3]提出一种针对2类不平衡数据分类的BEV算法，该算法主要基于数据集划分的Bagging思想，即先将多数类样本划分，并与少数类两两组合，从而构成了多个平衡的子数据集；然后在每一个子数据集上分别进行分类；最后用投票的规则决选出最终的分类结果。文献[4]基于不平衡样本的临近样本会给类别边缘的判断提供有价值信息的假设，提出一种LB(Lazy Bagging)算法，其主要思想是在修剪引导袋时将选取的样本点作为测试样本点的临近样本。以上是从数据集和Bagging算法本身进行改进的。大量的实验表明，不同分类算法在不同数据集上采用不同的分类器评价标准时，各算法的分类效果不一，其优劣各异。现有的分类算法在不平衡数据集上的效果仍有待提高。

为此，本文提出一种新的基于Bagging组合学习的方法CombineBagging，该方法的主要思想是对不平衡数据集先采用SMOTE进行数据预处理，使数据样本趋于平衡，然后采用3种不同的基分类器进行Bagging集成学习，再通过投票规则得出最终的分类结果。

2 数据预处理和Bagging集成学习方法

2.1 数据预处理

文献[5]提出了一种向上采样的方法SMOTE。该方法使分类器的分类平面向多数类的空间伸展，同时可有效地避免随机向上采样的过学习问题。该方法建立在如下的假设之上：相距较近的正例样本之间的样本依然是正例。其主要思想为，首先为每个稀有类样本随机选出几个邻近样本，并且在该样本与这些邻近样本的连线上随机取点，生成无重复的新样本。具体算法如下：首先计算样本集中每个少数类别的 k 个最近多数类邻点；然后随机抽出少数类样本 i 的 k 个最近多数类邻点中的一个多数类样本 j；接下来在 i 与 j 之间构造新的少数类样本 l：$l=i+random(0,1)\times(j-i);random(0,1)$，其中，$random(0,1)$ 表示产生一个0~1的随机数。重复以上步骤，直至所有少数类样本均处理完为止。

2.2 Bagging集成学习方法

集成学习方法按个体学习器之间依赖关系的强弱可以归

基金项目：教育部留学回国人员科研启动基金资助项目

作者简介：秦姣龙(1985－)，女，硕士研究生，主研方向：机器学习；王 蔚，教授

收稿日期：2010-12-30 **E-mail：**wangwei5@njnu.edu.cn

为 2 种：个体学习器之间不存在强的依赖关系，可以并行生成算法；个体学习器之间存在强的依赖关系，必须串行生成算法。这 2 类的典型代表分别为 Bagging 算法和 Boosting 算法。在一些实际问题的处理中，Boosting 算法会出现过拟合现象，从而导致得到的分类结果比单个分类器分类的结果还差。因此，本文选用了 Bagging 算法。Bagging 算法的基本思想是使用同一种基分类器对样本进行多次训练，在抽取样本时，采用有放回的抽取方式，最终分类结果的裁决方式为多数投票法[6]。

研究表明，Bagging 可以显著提高不稳定的基学习器的泛化能力[5]。要使 Bagging 有效，基分类器的学习算法必须是不稳定的，即算法对训练数据敏感，且基分类器的学习算法对训练数据越敏感，Bagging 的效果越好[6]。因此，分别选取支持支持向量机(Support Vector Machine, SVM)中的 C-SVM 算法、人工神经网络中的 RBF 神经网络算法、决策树中的 Random Forests 算法作为 Bagging 的基分类器的学习算法。Bagging 集成学习时，由于各学习算法对训练数据的敏感性不同，采用不同的分类算法作为基分类器的学习算法，因此其分类结果有较大的差异。为尽量减少 Bagging 对训练数据的敏感性所造成的分类错误，本文提出一种基于 Bagging 组合的不平衡数据分类方法 CombineBagging。

3　CombineBagging 集成方法

CombineBagging 集成方法先采用 SMOTE 进行数据预处理，然后采用 C-SVM、RBF Neural Network、Random Forests 3 种不同的基分类器学习算法，分别对采样后的数据样本进行 Bagging 集成学习，最后通过投票规则集成学习结果。下面先简单介绍 C-SVM、RBF Neural Network、Random Forests 3 种学习算法，然后阐述 Combine Bagging 方法的思想。

SVM 的核心思想是，对于输入空间中非线性可分的情形，选择适当的非线性映射，将输入空间中的样本点映射到一个高维空间，然后通过一系列核函数、参数因子的选择得到最优分界面。它在解决小样本、非线性及高维模式识别中表现出许多特有的优势。RBF 神经网络是一种局部逼近网络，可以以任意精度逼近任一连续函数。其最基本的构成包括 3 层：(1)输入层，由一些感知单元组成，它们将网络与外界环境连接起来；(2)隐层，其作用是从输入空间到隐层空间之间进行非线性变换，在大多数情况下，隐层空间有较高的维数；(3)输出层，输出层是线性的，它为作用于输入层的激活模式提供响应。Random Forests 是一个包含多个决策树的分类器，主要是由一系列的树状分类器 $h(x,\boldsymbol{\theta k})$(即决策树)构成的，其中，$\boldsymbol{\theta k}$ 是独立同分布的随机向量，每个树状分类器对于输入给出独立的分类，作为对最终分类结果的投票。该分类器被验证有很好的分类性能，不会产生过适应的问题。

CombineBagging 的主要思想如下：设 $S=\{\{x_1,y_1\},\{x_2,y_2\},\cdots,\{x_i,y_j\},\cdots,\{x_m,y_n\}\}$，$x_i\in X, y_j\in Y, i\leqslant m, j\leqslant n$ 为训练样本，训练样本特征集为 X，样本类别集为 Y，总训练样本数为 m，总类别数为 n。首先用 SMOTE 方法对样本集中的少数类样本进行采样，以使不平衡数据趋于平衡，新的样本集为 $S'=\{\{x_1,y_1\},\{x_2,y_2\},\cdots,\{x_k,y_j\},\cdots,\{x_m',y_n\}\}$，$x_k\in X', y_j\in Y, k\leqslant m', m'>m, j\leqslant n$；在采样后的样本 S' 上，采用 C-SVM、RBF Neural Network、Random Forests 3 种基分类器学习算法，分别对 S' 进行 Bagging 集成学习；最后通过投票规则将学习结果进行集成。具体投票的规则为：每个样本的最终分类结果以 3 种学习结果中票数最多的一类为此样本的分类结果，若 3 种学习结果都不同，则以 Random Forests 的分类结果为准，这是因为在许多不同的数据集上，Random Forests 可以产生高的分类准确率；其次，对于不平衡的数据，Random Forests 能平衡误差。算法描述如下：

输入 Training Set S, Machine Learning Algorithm Smote, Bagging, C-SVM, RBF Neural Network，Random Forests

S'=smote(S) //利用 SMOTE 技术获取新的样本集 S'
For i=1 to 3{
Ti=bagging(algorithmi(S'))
// algorithmi={C-SVM,RBF Neural Network,Random Forests}，采
//用 C-SVM、RBF 神经网络、Random Forests 3 种不同的基分类器
//学习算法，分别对 S'进行 Bagging 集成学习}

$$T'(x)=\arg\max_{y\in Y}\sum_{i=1}^{3}I(T_i(x)=y)$$

//通过投票规则，将学习结果集成

输出 classifier T' //返回结果

4　实验与分析

4.1　实验数据及方法

本文选取 UCI 数据集中的 haberman 样本集进行测试，网址是 http://www.ics.uci.edu/~mlearn/MLRepository.html，具体信息如表 1 所示。

表 1　实验使用的 haberman 数据样本集

数据集	规模	类别	主动类别	P: N
haberman	306	2	Class=2	81: 225

实验中采用十倍交叉验证，使用 C-SVM、RBF Nerual Network 和 Random Forests 3 种分类算法作为实验的比对方法，对于各 SVM 算法均选用径向基核函数，并使用网格寻优法为其选取 C 和 gamma 参数的值。分类器的性能评价标准选用 *G-mean* 和 A_{UC} 2 种方式，具体计算公式如下：

$$G\text{-}mean=\sqrt{TP/(TP+FN)\times TN/(TN+FP)}$$

$$A_{UC}=1-\frac{\frac{FN}{TN+FN}+\frac{FP}{TP+FP}}{2}$$

其中，*TP* 为正例判为正例的个数；*TN* 为负例判为负例的个数；*FP* 为正例误判为负例的个数；*FN* 为负例误判为正例的个数。

4.2　实验结果及分析

实验结果如表 2 所示，其中，每个数据集上最好的结果用粗体标出，最差的结果以斜体标出。从 *G-mean* 值可以看出，CombineBagging 对正/负例的判断准确率比其他 4 种算法好；在 A_{UC} 值方面，CombineBagging 也明显高于其他 4 种算法。由表 2 可以得出，CombineBagging 提高了少数类的分类准确率，对于处理不平衡数据问题是切实有效的。

表 2　haberman 样本集上的 *G-meam* 和 A_{UC} 值

算法	*G-mean*	A_{UC}
RBF Nerual Network	0.640	0.560
Random Forest	0.564	0.586
SVM	0.574	0.514
asBagging	*0.510*	*0.509*
CombineBagging	**0.726**	**0.716**

5　结束语

本文提出了一种基于 Bagging 组合的不平衡数据分类方法 CombineBagging，该方法先采用 SMOTE 进行数据预处理，然后采用 C-SVM、RBF 神经网络、Random Forests 3 种不同的基分类器学习算法分别对采样后的数据样本进行 Bagging 集成学习，最后通过投票规则将学习结果进行集成。

(下转第 182 页)