基于样本权重更新的不平衡数据集成学习方法

陈圣灵 沈思淇 李东升

(国防科技大学并行与分布处理国家重点实验室 长沙 410073)

摘 要 不平衡数据的问题普遍存在于大数据、机器学习的各个应用领域,如医疗诊断、异常检测等。研究者提出或采用了多种方法来进行不平衡数据的学习,比如数据采样(如SMOTE)或者集成学习(如EasyEnsemble)的方法。数据采样中的过采样方法可能存在过拟合或边界样本分类准确率较低等问题,而欠采样方法则可能导致欠拟合。文中将SMOTE,Bagging,Boosting等算法的基本思想进行融合,提出了Rotation SMOTE算法。该算法通过在Boosting过程中根据基分类器的预测结果对少数类样本进行SMOTE来间接地增大少数类样本的权重,并借鉴Focal Loss的基本思想提出了根据基分类器预测结果直接优化AdaBoost权重更新策略的FocalBoost算法。对不同应用领域共11个不平衡数据集的多个评价指标进行实验测试,结果表明,相比于其他不平衡数据算法(包括SMOTEBoost算法和EasyEnsemble算法),Rotation SMOTE算法在所有数据集上具有最高的召回率,并且在大多数数据集上具有最佳或者次佳的G-mean以及F1Score;而相比于原始的AdaBoost,FocalBoost则在其中9个不平衡数据集上都获得了更优的性能指标。

关键词 SMOTE,Boosting,不平衡数据,集成学习

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.07.005

Ensemble Learning Method for Imbalanced Data Based on Sample Weight Updating

CHEN Sheng-ling SHEN Si-qi LI Dong-sheng

(National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha 410073, China)

Abstract The problem of imbalanced data is prevalent in various applications of big data and machine learning, like medical diagnosis and abnormal detection. Researchers have proposed or used a number of methods for imbalanced learning, including data sampling (e.g. SMOTE) and ensemble learning (e.g. EasyEnsemble) methods. The oversampling methods in data sampling may have problems such as over-fitting or low classification accuracy of boundary samples, while the under-sampling methods may lead to under-fitting. The Rotation SMOTE algorithm was proposed in this paper incorporating the basic idea of SMOTE, Bagging, Boosting and other algorithms, and SMOTE was used to indirectly increase the weight of minority samples based on the prediction result of the base classifier in the Boosting process. According to the basic idea of Focal Loss, this paper proposed FocalBoost algorithm that directly optimizes the sample weight updating strategy of AdaBoost based on the prediction results of the base classifier. Based on the experiment with multiple evaluation metrics on 11 unbalanced data sets in different application fields, Rotation SMOTE can obtain the highest recall score on all datasets compared with other imbalanced data learning algorithms (including SMOTEBoost and EasyEnsemble), and achieves the best or the second best G-means and F1Score on most datasets, while FocalBoost achieves better performance on 9 of these unbalanced datasets compared to the original AdaBoost.

Keywords SMOTE, Boosting, Imbalanced data, Ensemble learning

1 引言

不平衡数据广泛存在于人们的实际生产和生活中,而不平衡数据的学习通常存在较大的挑战。传统的机器学习分类算法通常假设不同类的数据比例是平衡的,对不同类样本同等对待,并以提高总体分类精度为目标。但此类算法没有将数据分布纳入考虑范围内,当其中一类数据的数量远超过其他类时,就会造成分类器偏向多数类样本,使其在多数类的分类精度较高而在少数类的分类精度很低;在最坏的情况下,少数类样本会被视为多数类的异常点而被忽略,将所有样本都

到稿日期:2017-07-30 返修日期:2017-10-27 本文受国家重点基础研究发展计划(0800067314001),国家自然科学基金项目(61602500,61502500)资助。

陈圣灵(1993-),男,硕士,主要研究方向为大数据及机器学习,E-mail:waitsl@126.com;**沈思淇**(1985-),男,博士,助理研究员,主要研究方向为大数据及机器学习,E-mail:shensiqi@nudt.edu.cn(通信作者);**李东升**(1978-),男,博士,研究员,CCF会员,主要研究方向为大数据及机器学习,E-mail:dsli@nudt.edu.cn。

分类到多数类下，这种情况下学习出的模型显然是错误的，并且可能造成严重后果。例如在医学诊断领域，如果把得了癌症的病人误判为正常，那将有可能使其付出生命的代价。因此，传统的分类方法在不平衡数据集中具有很大的局限性，我们必须采取一定的措施来确保能够从这些少数类样本中学习到重要的信息[1]。

目前，已有许多有效的技术用于缓解不平衡数据学习的问题，比较常用的有数据采样和集成学习等方法。数据采样一般是通过增加或减少样本的方法来平衡数据分布，主要分为过采样和欠采样两种方式，最简单的方法是对原始数据进行随机的重采样。随机过采样是随机重复地从少数类样本中抽取样本并将其添加到原样本中，随机欠采样则是随机重复地从多数类样本中抽取样本并将其从原样本中删除，通过设置合适的采样比例来使不同类样本最终达到平衡。随机过采样和随机欠采样都存在一定的缺陷。随机过采样由于只是简单地复制样本，容易造成过拟合；而随机欠采样由于减少了多数类样本，有可能会损失一部分有用的信息。

集成学习是通过构建并结合多个学习器来完成学习任务的方法，通过将多个学习器进行结合，常常可获得比单一学习器更优越的泛化性能。根据集成的个体学习器的生成方式，集成学习主要可以分为两大类：1）Bagging，个体学习器间不存在强依赖关系、可同时生成的并行化方法；2）Boosting，个体学习器间存在强依赖关系、必须串行生成的序列化方法[2]。Bagging 采用 bootstraping 技术对原始样本进行不放回的抽样，在得到一系列子样本后，对每个子样本进行训练，以得到多个学习器，再对其进行集成。随机森林即是 Bagging 的一个变种。Boosting 则是迭代地训练原始样本，每次迭代根据上一次的训练结果修改原始样本的分布，在得到不同的学习器后再进行集成。比较经典的 Boosting 方法有 AdaBoost。

集成学习不是专门针对不平衡数据进行处理的方法，它能一定程度地改善对所有数据的学习。为了更加突出它对不平衡数据集中少数类样本准确率的提升效果，有研究者将其与 SMOTE[3] 等数据采样方法相结合，如 SMOTEBoost[4] 即是通过 SMOTE 修改 AdaBoost 中的样本分布，使其每次 Boosting 训练的样本都是平衡的。由于 AdaBoost 每次迭代都会调整样本的权重分布，尤其是对于错误样本，大大增加了其权重，使得在下次迭代训练中会更多地关注错误样本。SMOTEBoost 的改进增加了少数类样本，相当于变相地改变了样本的权重分布；但是它对所有少数类样本一视同仁，并没有考虑误分类样本的特殊性，因此从另一个角度来说其又模糊了 AdaBoost 中错误分类样本与正确分类样本的非同等重要性的区别。

本文在结合 SMOTE 过采样技术、Bagging 和 Boosting 集成学习思想的基础上，提出了一种新的混合集成采样算法 Rotation SMOTE。该算法采用了一种新的 Ensemble 方式，通过对原始样本进行线性特征变换得到了多个不同的新样本，之后再学习得到多个不同的模型并将其进行集成。其核心组成部分是基于 AdaBoost.M2 对基分类器预测结果进行改进的算法，称之为 boostSMOTE。该算法是在每次 Boosting 时，根据上一次分类器的预测结果对正样本（少数类样本）采用 ADASYN[5] 这个关注正负样本边界信息的 SMOTE 方法进行重采样。研究表明，这种混合的 Ensemble 方法能够获得更好的分类性能，尤其是对提升少数类样本的准确率有很好的帮助。另外，本文提出了一种新的权重更新策略来改进 AdaBoost 算法。研究表明，除通过采样算法来影响 Boosting 过程中不平衡样本的权重外，还有更简单有效的方式有助于区分不平衡样本分类的难易程度。本文仅考虑二分类问题，对于多分类的情况（如 bilibili 视频用户的分类[6]），可以基于该算法利用 one-vs-rest 或者 one-vs-one 方法进行分类。

2 相关工作

目前，已有许多通过数据采样和集成学习来缓解不平衡问题的方法，在数据采样技术中，SMOTE（合成少数类过采样技术）是一种比较有效的过采样方法。它的基本思想是根据少数类样本的特征空间相似性来合成新样本。具体做法是：针对每个少数类样本，计算其在该类中的 K 个近邻，并从中随机选择一个或多个样本与原样本进行线性差值计算得到新的样本。这种方式能够有效缓解随机过采样采取简单复制样本的方式所造成的模型过拟合问题。原始的 SMOTE 算法将每个少数类样本都合成新样本，没有考虑多数类的邻近点，容易发生类间的样本重叠，引入额外的噪音，且实际建模过程中发现那些处于分类边界位置的样本更容易被错分。因此，SMOTE 衍生出了很多变种，比较有效的有 Borderline-SMOTE[7] 及 ADASYN。Borderline-SMOTE 对少数类中每个样本求得的 K 近邻不再仅仅是少数类样本，而是属于整个样本集合，且只对能够代表接近分类边界的样本进行 SMOTE 合成新样本。ADASYN 则是根据各个少数类样本最近邻中多数类样本的数量来生成相应比例的新样本，邻近样本集合中多数类样本越多表示越靠近分类边界，合成的样本也就越多[5]。这种利用边界位置样本信息产生新样本的方式可以给模型带来更大的提升。除利用边界位置样本信息的 SMOTE 外，还有其他与聚类算法相结合的样本生成方式，如 kmeans-SMOTE[8]，CURE-SMOTE[9]，ASCB-DmSMOTE[10] 等。这种方式是先对样本数据进行聚类得到多个簇，再分别对每一个簇进行 SMOTE，比较适用于数据集表现为多个团簇分布的情况。

SMOTE 系列采样算法是为了缓解随机过采样发生的模型过拟合问题，而对于欠采样造成的数据信息丢失问题，也有相应的缓解方式，EasyEnsemble[11] 就是其中的一种集成学习方法。其基本思想是：对于多数类样本，利用 BoostStrapping 技术，通过 N 次有放回的抽样生成 N 份与少数类样本同等比例的子集，再将少数类样本和这 N 份子集分别合并得到新的 N 个平衡子样本；通过每个子样本训练一个 AdaBoost 模型，每个 AdaBoost 模型又是多个弱学习器（如决策树）的集成，其训练过程会根据预测结果给每个基学习器生成一个相应的权重，最终的 EasyEnsemble 模型则是这多个 AdaBoost 模型中所有弱学习器的加权平均。EasyEnsemble 算法融合了 Bagging 和 Boosting 的思想，能够获得更好的模型性能。

除了 EasyEnsemble 这种将欠采样技术与 AdaBoost 简单结合的方式外，还有利用采样技术改进 AdaBoost 中 Boosting

过程的方法。SMOTEBoost 就是一种通过 SMOTE 影响 AdaBoost. M2 算法中权重更新的算法，它在每一次 Boosting 之前对原始样本进行 SMOTE 合成以得到新的样本，并赋予每个合成样本一个初始权重，每次训练的样本基于原始样本与合成样本，并根据分类错误率更新原始样本的权重。RUSBoost[12] 沿用了 SMOTEBoost 的思想，但在每个 Boosting 过程中应用随机欠采样而非 SMOTE。这种将采样算法应用于 Boosting 过程的方式使得每次训练的样本都是平衡的，因此 RUSBoost 和 SMOTEBoost 都比原始的 AdaBoost. M2 算法更适用于不平衡数据。

对于集成学习而言，Bagging 和 Boosting 一般都是基于“重采样法”或“重赋权法”来处理原始样本的。“重采样法”是利用数据采样技术对原始样本进行采样，从而得到多个子样本；“重赋权法”则是对原始样本赋予不同大小的权重来影响样本分布。基于决策树集成的方法 Rotation Forest[13]，则是另外一种类型。它采取了一种投影变换的方式，利用 PCA 对原始样本进行特征变换，以得到多个不同的新样本（新样本并非原样本的子样本，而是原样本的多种不同特征表现形式），通过对多个不同的新样本进行训练，得到若干个弱学习器后再集成。

近年来，深度学习技术被广泛应用于机器学习的各个领域。在目标检测领域中，有研究者针对样本类别不平衡问题提出了一种新的损失函数 Focal Loss。该损失函数是在标准交叉熵的基础上添加了一个与预测概率值 P 相关的控制函数，样本越容易分类，则 P 越大，其贡献的损失就越小，难分样本所占的比重会越大，以此区分出易分样本和难分样本[14]。

GAN（生成对抗网络）是神经网络中一种能够从训练样本中学习出新样本的网络结构[15]，它的基本思想是根据训练集估计其样本分布，再通过该样本分布生成与训练集类似的样本。我们或许可以将 GAN 技术应用于少数类样本的生成中来解决不平衡数据的问题。

3　Rotation SMOTE

本文提出的 Rotation SMOTE 算法是一种融合了 EasyEnsemble，RotationForest，SMOTEBoost 等多种算法思想的 Ensemble 方法。为了增加子样本的多样性，Ensemble 的方式并没有采用传统的“重采样法”或“重赋权法”，而是依据 RotationForest 中的方法利用 PCA 对原始样本进行 Rotation 转换，以得到一系列新的样本。原始的 RotationForest 并不适用于不平衡数据，本文将其中的基分类器决策树替换成基于 AdaBoost. M2 改进的 boostSMOTE 算法。算法 1 描述了 Rotation SMOTE 的具体流程。

算法 1　Rotation SMOTE

Given: Set S of examples $\{(x_1,y_1),\cdots,(x_m,y_m)\}$, $x_i\in X$, with minority (positive) class $y^r\in Y$, $|Y|=2$.

1. Get new training dataset $S'(X',Y')$ by creating N synthetic examples from minority class using the ADASYN algorithm if S is imbalanced.
2. Do for i=1…L
   2.1. Prepare the rotation matrix R_i^a based on S'.
   Split F (the feature set) into K subsets: $F_{i,j}$ (j=1…K)
   For j=1…K
   Apply PCA on $X'_{i,j}$ (a random subset from $X_{i,j}$ based on X') to obtain the coefficients in a matrix $C_{i,j}$
   Arrange the $C_{i,j}$ (for =j=1…K) in a rotation matrix R_i
   construct R_i^a by rearranging the columns of so as to match the order of features in F.
   2.2. Build a boostSMOTE classifier $h_i: X'\times Y'\to[0,1]$ using $(X'R_i^a, Y')$ as the training set
3. Output the final hypothesis: $H(x)=\arg\max\limits_{y'\in Y'}\sum\limits_{i=1}^{L} h_i(x'R_i^a,y')$

1）首先，利用 ADASYN 算法生成原始样本的新样本 $S'(X',Y')$，使得转换之后得到的新样本保持平衡。

2）然后，利用不同的旋转矩阵变换得到若干份新样本。旋转矩阵的生成过程如下：先按特征维度将原样本数据分割成 K 个子集。对于每一个特征样本子集 $X_{i,j}$，从中随机去除若干样本，之后使用 bootstraping 随机采样得到 $X'_{i,j}$，再对其进行 PCA，将这样的多个特征子集的 PCA 结果进行重新排列即可得到旋转矩阵 R_i^a。新样本即为 $(X'R_i^a,Y')$。

3）得到若干个新样本之后，再对每一份样本进行训练，从而得到 boostSMOTE 模型。使用模型进行预测时，需要先对测试集特征矩阵右乘与训练集相同的旋转矩阵 R_i^a 进行相应的变换，Rotation SMOTE 模型的最终预测结果即为若干个 boostSMOTE 模型预测结果的平均。

算法 2 给出了 Rotation SMOTE 算法的核心 boostSMOTE。它采用 SMOTE 来改进 AdaBoost. M2，以适应不平衡数据，并对其中合成样本的策略做了针对性的改动，以区分各个样本分类的难易程度。

算法 2　boostSMOTE

Given: Set S of examples $\{(x_1,y_1),\cdots,(x_m,y_m)\}$, $x_i\in X$, with minority (positive) class $y^r\in Y$, $|Y|=2$.

1. Initialize the distribution D_1 over the new examples S, such that $D_1(i)=1/m$.
2. Get temporary training dataset S_1' and distribution D_1' by creating N synthetic examples from minority class using the ADASYN algorithm if S is imbalanced.
3. Do for t=1,…T
   3.1. Train a weak learner on S_t' using distribution D_t'.
   3.2. Get back a weak hypothesis $h_t: X\times Y\to[0,1]$ and predict S to get the false negative (FN) instances and the true positive (TP) instances.
   3.3. Calculate the pseudo-loss (for S and D_t):
   $$\varepsilon_t=\sum_{(i,y):y_i\neq y} D_t(i)(1-h_t(x_i,y_i)+h_t(x_i,y))$$
   3.4. Calculate the weight update parameters $\beta_t=\frac{\varepsilon_t}{1-\varepsilon_t}$, $w_t=\frac{1}{2}(1-h_t(x_i,y:y\neq y_i)+h_t(x_i,y_i))$ and then update $D_t: D_{t+1}(i)=\frac{D_t(i)}{Z_t}\cdot\beta_t^{w_t}$.
   3.5. Creating N synthetic examples from FN or TP using the ADASYN algorithm and get temporary training dataset S'_{t+1} and their weights D'_{t+1}.
4. Output the final hypothesis: $H(x)=\arg\max\limits_{y\in Y}\sum\limits_{t=1}^{T} h_t(x,y)\log\frac{1}{\beta_t}$

1）首先，对于 Rotation 转换得到的样本数据 S（转换后得

到的样本是平衡的，当然也可以适用不平衡的数据），初始化每个样本的权重为 $1/m$，其中 m 为样本总量。

2）然后，若训练的样本 S 为不平衡样本，则使用 ADASYN 算法合成样本（这样可以使第一次 Boosting 训练的样本是平衡的，否则可能会生成错误率大于 0.5 的弱分类器而导致 Boosting 过程无法继续）。

3）接着，Boosting 迭代训练得到 T 个弱分类器，与 AdaBoost.M2 算法类似，但其在每次迭代中训练的数据除了样本权重更新之外，还生成了新的样本：第 3.2 步中，基于本次训练得到弱分类器 h_t，对原始样本 S 进行预测，分别得到 FN 和 TP；第 3.5 步中，针对本次预测得到的 FN 和 TP，利用 ADASYN 算法生成新的样本，此处所采用的策略是优先对 FN 生成，若 FN 数量很少，表示该弱分类器的能力较强，则对 TP 生成。

ADASYN 算法生成样本的方式是：针对每个少数类样本，根据其邻近样本点中多数类样本的个数生成相应比例的新样本，若 FN 或 TP 邻近点中没有多数类样本，则无法生成。本文对此做出了相应的改进：当输入的所有少数类样本（一定范围内）的邻近点中没有多数类样本时，根据少数类样本在本次预测中得到的概率值生成相应比例的新样本，预测概率值越大，则生成样本的数量越少。其中，将对错误样本生成新样本的总 SMOTE 比例设为 $FN/N * 2$（分母 N 表示训练集中负样本的个数，FN 表示每次预测结果中错误分类的正样本个数），使得每次生成样本的总量为 FN（若 FN 太小，则对 TP 生成数据，与 FN 类似，设置 SMOTE 比例为 $TP/N*2$，若该值大于 1，则 SMOTE 比例设为 1）。

4）通过以上策略生成新样本之后，得到新的样本 S_1' 及其权重分布 D_{t+1}'，其中 S_1' 仅包含原始样本 S 与本次新生成的样本，在下次训练完成之后丢弃其中的生成样本，并更新权重分布。最终的模型为多个弱分类器的集成。

4 基于 Focal Loss 优化的 AdaBoost

除了 Roation SMOTE 那类通过 SMOTE 数据合成采样算法来间接地更新 Boosting 过程中的样本权重外，还可以采用通过修改样本权重更新策略的方式来处理不平衡问题。借鉴深度学习目标检测领域 Focal Loss 的基本思想，可以在 Boosting 过程中将上次训练得到的弱模型的预测概率值 P 作为控制样本权重的系数加入下次训练的权重更新计算中，使得预测概率值越大，权重更新的幅度越大，从而在一定程度上区分出每个样本分类的难易程度。

对于原始的 AdaBoost 算法，其样本权重更新策略如下：

$$D_{t+1}(x)=\frac{D_t(x)\mathrm{e}^{-\alpha_t f(x)h_t(x)}}{Z_t}$$

其中，D_t 表示第 t 次 Boosting 迭代训练样本的权重分布，h_t 为第 t 次训练得到的弱分类器，α_t 为根据错误率计算得到的弱分类器的权重，Z_t 为规范化因子，用于确保 D_t 是一个分布。

借鉴 Focal Loss 的思想，将权重更新公式修改为：

$$D_{t+1}(x)=\frac{D_t(x)\mathrm{e}^{-\alpha_t f(x)h_t(x)/(1-p_t)^2}}{Z_t}$$

其中，p_t 为 h_t 在训练样本上的预测概率值。通过加入权重控制系数 $1/(1-p_t)^2$，可以在原来的基础上增大样本权重减小或增大的幅度，使得每一个训练样本分类的难易程度都得到很好的区分。

5 实验方案与结果分析

本文进行了大量的实验来全面且系统地评估 Rotation SMOTE 算法，将其与 RandomForest，Randomforest + SMOTE，SMOTEBoost，EasyEnsemble 及其基本组成 boostSMOTE 等算法进行对比。结果表明，所提方法能够获得更好的 Recall 指标及更好的或可比较的 G-mean 和 F1Score。同时，通过比较基于 Focal Loss 思想改进的 AdaBoost-FocalBoost 与原始的 AdaBoost 算法，也能够看出这种策略对于不平衡数据能够获得一定程度的性能提升。

5.1 实验数据集

表 1 列出了实验测试所用 11 个数据集的基本信息，包括特征维度、数据集大小、少数类样本数和多数类样本数，以及不平衡比例。这些数据集大部分来自于 UCI[16]，涵盖了多种不同的应用领域及不同程度的不平衡比例（按不平衡比例从小到大排序）。由于本文仅考虑二分类问题，因此有必要对其中的多分类数据集进行转换，以得到二分类标签。根据相关文献[5,17]中的常用方法来修改数据标签，即选择其中某一类作为正样本，并将剩余的类组合作为负样本。

表 1 不平衡数据集的基本信息

Table 1 Basic information of imbalanced dataset

数据集	特征维度	数据集大小	少数类样本数	多数类样本数	不平衡比例（多数类:少数类）
spambase	57	4601	1813	2788	1.5
ionosphere	34	351	126	225	1.8
pima	8	768	268	500	1.9
breastcancer	10	699	241	458	1.9
biodeg	41	1055	356	699	2.0
phoneme	5	5404	1586	3818	2.4
vehicle	18	846	199	647	3.3
page-blocks	10	5473	560	4913	8.8
satimage	36	6435	626	5809	9.3
fengji	6	11766	497	11269	22.7
mammography	6	11183	260	10923	42.0

5.2 性能评估方法

对于不平衡数据的学习而言，一般的总体准确率的评估方法显然是不适用的。因此，本文采取了 4 种不同的指标来对模型的性能进行评估，包括 Precision，Recall，F-measure 和 G-mean。这些指标的定义如下：

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

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

$$F\text{-}Measure=\frac{(1+\beta^2)\cdot Precision\cdot Recall}{\beta^2\cdot Precision+Recall}$$

其中，β 为调节 $Precision$ 和 $Recall$ 相对重要性的系数（取 1 即为 F1Score）。

$$G\text{-}mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}}=\sqrt{TPR\cdot TNR}$$

在众多分类应用领域中，我们总是希望尽可能地把所有少数类样本都识别出来，如在异常检测领域，我们希望尽可能

地把所有的异常点都检测出来，以免造成大的损失。这种需求表现在性能评价指标中便是更加看重 Recall 指标，期望其尽可能地接近 1（如果能获得较高的 Precision 将更完美），因此我们的目标是在允许略微牺牲 Precision 的情况下尽可能地提高 Recall，整体使得 F-measure 或 G-mean 等综合指标性能保持稳定或者得到提升。

5.3　实验设计

为了减小随机划分数据集造成的偏差，本文所有实验均采取 10-fold 交叉验证的方式，即将整个数据集分成 10 份，每次选择其中 9 份用于训练，剩余 1 份用于测试，最后的评价指标取 10 次训练的平均值。Rotation SMOTE 中 Bagging 阶段将生成 10 份训练集，每份训练集通过 Boosting 训练得到 1 个分类器，再将得到的 10 个分类器进行集成，其中 Boosting 阶段采取的基分类器为决策树，迭代次数为 10，不同的 Bagging 分类器个数和 Boosting 迭代次数对实验结果的总体情况没有造成太大的影响。

5.4　实验结果分析

表 2－表 5 列出了本文实验中 Rotation SMOTE 及其他多个分类器在 11 个数据集上的表现结果，包括 Recall，Precision，G-mean 和 F1Score，同时还给出了所有数据集的平均性能。对于每项指标，最佳结果都加粗及倾斜显示，次佳结果加粗显示。

表 2　Recall 性能对比

Table 2　Performance comparison of Recall

DataSet	Random Forest	Random Forest_ SMOTE	SMOTE Boost	Easy Ensemble	boost SMOTE	Rotation SMOTE
vehicle	0.8845	0.9650	0.8189	**0.9697**	0.8747	***0.9750***
phoneme	0.7863	**0.8594**	0.6374	0.8222	0.7004	***0.8808***
satimage	0.4460	0.5691	0.4258	**0.8868**	0.6172	***0.9297***
page-blocks	0.7964	0.8625	0.7107	**0.8786**	0.7125	***0.9768***
fengji	0.3541	0.5433	0.3379	**0.8631**	0.6415	***0.9216***
pima	0.5521	0.6155	0.6570	**0.7611**	0.6823	***0.7840***
spambase	0.8925	0.9013	0.8787	**0.9062**	0.8946	***0.9305***
biodeg	0.6875	0.7525	0.7613	**0.8735**	0.7749	***0.8791***
mammography	0.5077	0.7346	0.4385	**0.8538**	0.6077	***0.8885***
ionosphere	**0.8590**	0.8359	0.8128	0.7673	0.8365	***0.8994***
breastcancer	0.9255	0.9422	0.9092	0.9338	**0.9505**	***0.9835***
平均值	0.6992	0.7801	0.6717	**0.8651**	0.7539	***0.9135***

表 3　Precision 性能对比

Table 3　Performance comparison of Precision

DataSet	Random Forest	Random Forest_ SMOTE	SMOTE Boost	Easy Ensemble	boost SMOTE	Rotation SMOTE
vehicle	**0.9360**	***0.9396***	0.8593	0.5153	0.8506	0.7704
phoneme	***0.8626***	**0.8219**	0.6611	0.5978	0.6294	0.5263
satimage	***0.7286***	**0.6071**	0.4057	0.3436	0.4533	0.3056
page-blocks	***0.8667***	0.7957	0.7889	0.7035	**0.7984**	0.5663
fengji	***0.6744***	0.4237	**0.6443**	0.3272	0.4506	0.2742
pima	***0.7014***	0.6384	**0.6542**	0.6164	0.6459	0.5779
spambase	***0.9322***	**0.9235**	0.9051	0.8850	0.8766	0.8233
biodeg	***0.8260***	**0.8190**	0.7003	0.6663	0.7057	0.6633
mammography	***0.8779***	0.5801	**0.7653**	0.1692	0.5810	0.1488
ionosphere	***0.9366***	0.9226	0.9026	***0.9732***	0.8793	0.9289
breastcancer	***0.9522***	0.9433	0.9362	0.9408	**0.9461**	0.9332
平均值	***0.8450***	**0.7650**	0.7475	0.6126	0.7106	0.5926

表 4　G-mean 性能对比

Table 4　Performance comparison of G-mean

DataSet	Random Forest	Random Forest_ SMOTE	SMOTE Boost	Easy Ensemble	boost SMOTE	Rotation SMOTE
vehicle	0.9303	***0.9719***	0.8830	0.8214	0.9105	**0.9386**
phoneme	**0.8631**	***0.8901***	0.7416	0.7955	0.7577	0.7666
satimage	0.6370	0.7253	0.5566	***0.8432***	0.7375	**0.8383**
page-blocks	0.8784	**0.9125**	0.8194	0.9068	0.8238	***0.9362***
fengji	0.5825	0.7049	0.5405	**0.8407**	0.7431	***0.8475***
pima	0.6912	0.6976	0.7234	***0.7515***	**0.7351**	0.7341
spambase	**0.9237**	***0.9247***	0.9081	0.9125	0.9029	0.8933
biodeg	0.7931	***0.8268***	0.7900	**0.8149**	0.7960	0.8133
mammography	0.7104	0.8506	0.6573	**0.8762**	0.7647	***0.8827***
ionosphere	**0.9084**	0.8920	0.8732	0.8616	0.8815	***0.9276***
breastcancer	0.9485	0.9539	0.9355	0.9499	**0.9592**	***0.9720***
平均值	0.8061	0.8500	0.7662	**0.8522**	0.8193	***0.8682***

表 5　F1Score 性能对比

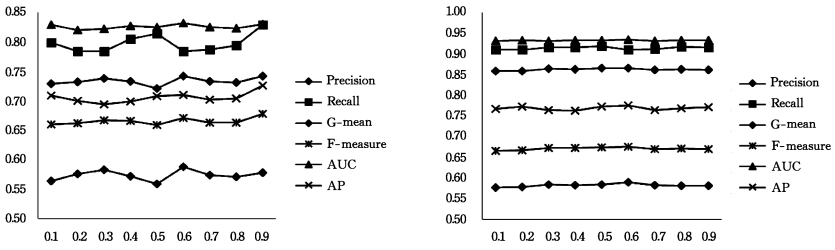
Table 5　Performance comparison of F1Score

DataSet	Random Forest	Random Forest_ SMOTE	SMOTE Boost	Easy Ensemble	boost SMOTE	Rotation SMOTE
vehicle	**0.9070**	***0.9505***	0.8345	0.6673	0.8588	0.8571
phoneme	**0.8224**	***0.8398***	0.6483	0.6920	0.6587	0.6580
satimage	**0.5357**	***0.5815***	0.4137	0.4923	0.5098	0.4575
page-blocks	**0.8104**	***0.8148***	0.7182	0.7558	0.7248	0.7038
fengji	**0.4340**	0.4202	0.3539	0.4134	***0.4454***	0.3774
pima	0.6148	0.6198	0.6492	***0.6800***	0.6605	**0.6639**
spambase	**0.9113**	***0.9114***	0.8912	0.8943	0.8840	0.8712
biodeg	0.7440	***0.7808***	0.7237	**0.7510**	0.7307	0.7505
mammography	**0.6412**	***0.6457***	0.5515	0.2823	0.5469	0.2546
ionosphere	**0.8918**	0.8720	0.8467	0.8416	0.8538	***0.9107***
breastcancer	0.9363	0.9405	0.9195	0.9357	**0.9463**	***0.9569***
平均值	**0.7499**	***0.7615***	0.6864	0.6732	0.7109	0.6783

从实验结果来看，相比于其他常用算法，本文提出的 Rotation SMOTE 方法在测试的所有数据集上的 Recall 性能是最好的。单从表 1 中 RandomForest 与 RandomForest＋SMOTE 的对比中就可以看出，SMOTE 有益于 Recall 的提升，而 boostSMOTE 算法则加强了这一效果，通过增加新的少数类样本能够从其样本分布中学到更多的信息，从而提高了 Recall。从表 2 可知，Rotation SMOTE 会在一定程度上降低 Precision，但在可接受的范围之内，与 Rotation SMOTE 相比，Random Forest 的 Precision 结果最佳，但是它的 Recall 效果是最差的。在表 3 中，大部分数据集上的 Rotation SMOTE 在 G-mean 指标上能获得最佳或次佳性能；而表 4 中 F1Score 相比于 RandomForest 等算法则略显颓势，但其在不平衡比例较低的小规模数据集上表现仍是最好的，且所有数据集的平均情况比 EasyEnsemble 好。另外，与之前的 SMOTEBoost 相比，Rotation SMOTE 算法的核心（即 boostSMOTE 算法）拥有更佳的表现，而通过 Rotation 转换的 Ensemble 方法在 boostSMOTE 的基础上又有了更进一步的提升，且这种 Ensemble 的方法比 EasyEnsemble 的平均性能更佳。

为了评估 SMOTE 比例对 Rotation SMOTE 方法的影响，本文通过设置 boostSMOTE 中不同的 SMOTE 比例参数（从 0.1 到 0.9）对比了算法的各项性能指标。图 1 给出了不同 SMOTE 比例下在所有数据集及 pima 数据集上多个评估指标的平均性能对比结果（其中 AUC 指标为 ROC 曲线下的面积，ROC 曲线是以 FPR 为横轴、TPR 为纵轴的性能曲线，AP 指标则为 recision-Recall 曲线下的面积）。从中可以看出，单从 pima 数据集上来看，SMOTE 比例产生了一定的影

响，当 SMOTE 比例为 0.6 时模型的性能最好；而整体上，不同的 SMOTE 比例对性能的影响较小。这可能是由于本文提出的这种有针对性的 boostSMOTE 方式能够生成有助于基分类器学习的样本，从而学到了更有效的信息，降低了基分类器的偏差，而 Rotation 的 Ensemble 方式降低了模型的整体方差，从而模糊了 SMOTE 比例带来的影响。

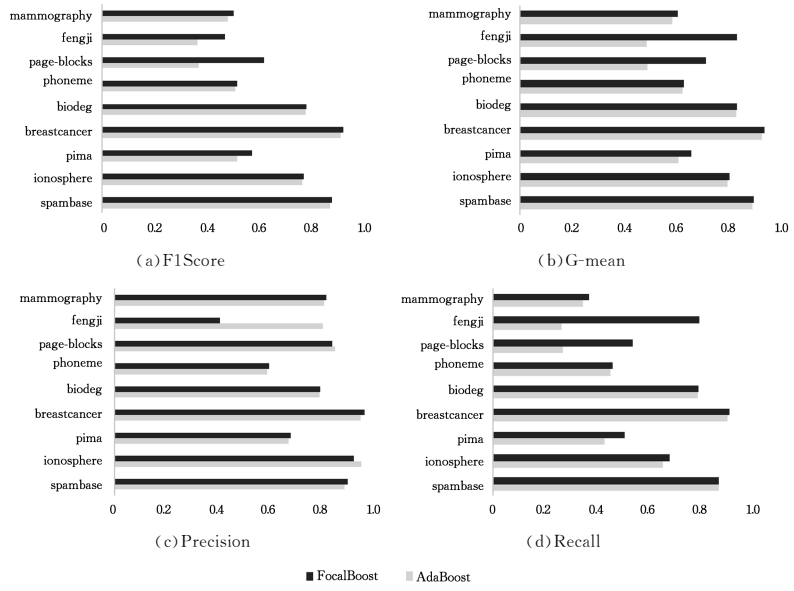


(a)PIMA (b)ALL DATASETS

图 1 SMOTE 比例对算法的影响

Fig. 1 Impact of different SMOTE ratios on performance of algorithm

图 2 给出了基于 Focal Loss 优化的 AdaBoost 算法 FocalBoost 与原始的 AdaBoost 算法的性能对比结果。可以看出，实验的 9 个数据集中 Recall，F1Score，G-mean 3 个指标都获得了不同程度的提升，且在约一半的数据集中 Precision 的结果也有小幅提升。因此，对于 Boosting 算法来说，这种与采样方法不同的样本权重更新优化策略有助于不平衡数据的学习。



(a)F1Score (b)G-mean

(c)Precision (d)Recall

图 2 基于 Focal Loss 优化的 AdaBoost 性能对比

Fig. 2 Performance comparison of AdaBoost based on Focal Loss optimization

结束语 不平衡数据普遍存在于人们生活的各个领域。本文探索了不平衡数据在机器学习中面临的问题，总结了已有的数据采样和集成学习等不平衡数据学习方法，并结合 SMOTE，Bagging，Boosting 等方法，提出了一种通过数据采样间接改变样本权重的集成学习方法 Rotation SMOTE，同时借鉴 Focal Loss 的基本思想提出了不同的样本权重更新策略来处理不平衡问题。实验结果表明，这两种有针对性地更新样本权重的方法能够获得很好的 Recall 及 G-mean 性能，为不平衡数据的学习提供了新的解决方案。未来我们将针对模型需要多次调用 SMOTE 从而使计算量较大这一问题进行优化，以提升模型的训练速度；此外，未来我们还将研究如何利用 GAN 来生成少数类样本，并将其与 Focal Loss 思想相结合，或许能够达到更好的效果。

参 考 文 献

[1] HE H，GARCIA E A. Learning from Imbalanced Data[J]. IEEE Transactions on Knowledge & Data Engineering，2009，21(9)：1263-1284.

[2] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[3] 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.

[4] CHAWLA N，LAZAREVIC A，HALL L，et al. SMOTEBoost：Improving prediction of the minority class in boosting[C]// European Conference on Knowledge Discovery in Databased：PKDD. 2003：107-119.

[5] HE H，BAI Y，GARCIA E A，et al. ADASYN：Adaptive synthetic sampling approach for imbalanced learning[C]// IEEE International Joint Conference on Neural Networks. IEEE，2008：1322-1328.

[6] JIA A L，SHEN S，CHEN S，et al. An Analysis on a YouTube-like UGC site with Enhanced Social Features[C]// Proceedings of the 26th International Conference on World Wide Web Com-

panion. 2017：1477-1483.

[7] HAN H，WANG W Y，MAO B H. Borderline-SMOTE：a new over-sampling method in imbalanced data sets learning[C]// International Conference on Intelligent Computing. Berlin，Springer，Heidelberg，2005：878-887.

[8] CIESLAK D A，CHAWLA N V，STRIEGEL A. Combating imbalance in network intrusion datasets[C]// IEEE International Conference on Granular Computing. IEEE，2006：732-737.

[9] LI M，FAN S. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests[J]. Bmc Bioinformatics，2017，18(1)：169.

[10] LI J，FONG S，SUNG Y，et al. Adaptive swarm cluster-based dynamic multi-objective synthetic minority oversampling technique algorithm for tackling binary imbalanced datasets in biomedical data classification[J]. Biodata Mining，2016，9(1)：37.

[11] LIU X Y，WU J，ZHOU Z H. Exploratory Undersampling for Class-Imbalance Learning[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society，2009，39(2)：539-550.

[12] 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.

[13] RODRÍGUEZ J J，KUNCHEVA L I，ALONSO C J. Rotation forest：A new classifier ensemble method[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2006，28(10)：1619-1630.

[14] LIN T Y，GOYAL P，GIRSHICK R，et al. Focal Loss for Dense Object Detection[OL]. http://www. researchgate. net/publication/322059369-Focal-Loss-for-Dense_Object-Detection.

[15] GOODFELLOW I J，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets[C]// International Conference on Neural Information Processing Systems. MIT Press，2014：2672-2680.

[16] ARTHUR A，DAVID N. The UCI Machine Learning Repository [DB/OL]. http://archive. ics. uci. edu/ml/datasets. html.

[17] CHEN S，HE H，GARCIA E A. RAMOBoost：Ranked Minority Oversampling in Boosting[J]. IEEE Transactions on Neural Networks，2010，21(10)：1624-1642.

（上接第 15 页）

[34] RUMI G，COLELLA C，ARDAGNA D. Optimization Techniques within the Hadoop Eco-system：A Survey[C]// 2014 16th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). IEEE，2014：437-444.

[35] VERMA A，CHERKASOVA L，CAMPBELL R H. ARIA：automatic resource inference and allocation for mapreduce environments[C]// Proceedings of the 8th ACM International Conference on Autonomic Computing. ACM，2011：235-244.

[36] SANDHOLM T，LAI K. Dynamic proportional share scheduling in hadoop[C]// Workshop on Job Scheduling Strategies for Parallel Processing. Springer Berlin Heidelberg，2010：110-131.

[37] RAO B T，REDDY L S S. Survey on improved scheduling in Hadoop MapReduce in cloud environments[J]. arXiv preprint arXiv：1207. 0780，2012.

[38] KC K，ANYANWU K. Scheduling hadoop jobs to meet deadlines[C]// IEEE Second International Conference on Cloud Computing Technology and Science. IEEE，2011：388-392.

[39] VERMA A，CHERKASOVA L，KUMAR V S，et al. Deadline-based workload management for mapreduce environments：Pieces of the performance puzzle[C]// Network Operations and Management Symposium (NOMS). IEEE，2012：900-905.

[40] ZACHEILAS N，KALOGERAKI V. Real-Time Scheduling of Skewed MapReduce Jobs in Heterogeneous Environments[C]// ICAC. 2014：189-200.

[41] XU X，CAO L，WANG X. Adaptiv e task scheduling strategy based on dynamic workload adjustment for heterogeneous Hadoop clusters[J]. IEEE Systems Journal，2016，10(2)：471-482.

[42] NIGHTINGALE E B，CHEN P M，FLINN J. Speculative execution in a distributed file system [J]. ACM SIGOPS Operating Systems Review，2005，39(5)：191-205.

[43] YANG Z W，ZHENG Q，WANG S，et al. Adaptive Task Scheduling Strategy for heterogeneous Spark Cluster[J]. Computer Engineering，2016，42(1)：31-35，40. (in Chinese)
杨志伟，郑烇，王嵩，等. 异构 Spark 集群下自适应任务调度策略[J]. 计算机工程，2016，42(1)：31-35，40.

[44] KANG H M. Research on Spark Optimization Based on Fine-Grained Monitoring[D]. Harbin：Harbin Institute of Technology，2016. (in Chinese)
康海蒙. 基于细粒度监控的 Spark 优化研究[D]. 哈尔滨：哈尔滨工业大学，2016.

[45] RANA N，DESHMUKH S. Shuffle Performance in Apache Spark[C]// International Journal of Engineering Research and Technology. ESRSA Publications，2015.

[46] DAVIDSON A，OR A. Optimizing Shuffle performance in Spark [R]. University of California，Berkeley-Department of Electrical Engineering and Computer Sciences，2013.

[47] JASON D. Consolidating Shuffle Files in Spark[EB/OL]. [2017-04-28]. https://issues. apache. org/jira/browse/SPARK-751.

[48] CHERN Y Z. Analysis and optimization of Memory Scheduling Algorithm of Spark Shuffle[D]. Hangzhou：Zhejiang University，2016. (in Chinese)
陈英芝. Spark Shuffle 的内存调度算法分析及优化[D]. 杭州：浙江大学，2016.

[49] YIGITBASI N，WILLKE T L，LIAO G，et al. Towards machine learning-based auto-tuning of mapreduce[C]// 2013 IEEE 21st International Symposium on Modelling，Analysis and Simulation of Computer and Telecommunication Systems. IEEE，2013：11-20.

[50] CHEN C O，ZHUO Y Q，YEH C C，et al. Machine Learning-Based Configuration Parameter Tuning on Hadoop System[C]// 2015 IEEE International Congress on Big Data. IEEE，2015：386-392.