doi: 10.12068/j.issn.1005-3026.2018.06.007

基于最大平衡度的自适应随机抽样算法

董立岩[1], 王越群[1], 李永丽[2], 朱 琪[1]
(1. 吉林大学 计算机科学与技术学院, 吉林 长春 130012; 2. 东北师范大学 计算机科学与信息技术学院, 吉林 长春 130117)

摘 要: 针对分类算法在非平衡数据集的情况下分类性能不理想的问题,总结了常见的数据平衡化方法,包括改造数据集与改进算法,提出一种全新的基于最大平衡度的自适应随机抽样算法,进一步优化了随机森林算法的分类效果.将其应用在随机森林算法的数据预处理阶段,并通过实验证明了该随机抽样方法的有效性,在合理的整体精度范围内能够较好地处理非平衡数据.产生的新数据比较拟合初始数据,能够提高分类器处理非平衡数据的能力.

关 键 词: 非平衡数据集;最大平衡度;随机抽样;随机森林;数据预处理
中图分类号: TP 301.6 **文献标志码**: A **文章编号**: 1005-3026(2018)06-0792-05

Adaptive Random Sampling Algorithm Based on the Balance Maximization

DONG Li-yan[1], *WANG Yue-qun*[1], *LI Yong-li*[2], *ZHU Qi*[1]
(1. College of Computer Science and Technology, Jilin University, Changchun 130012, China; 2. School of Computer Science and Technology, Northeast Normal University, Changchun 130117, China. Corresponding author: DONG Li-yan, E-mail: dongly@jlu.edu.cn)

Abstract: The problem on the classification algorithm of imbalanced datasets was analyzed. Common methods of balancing data, including improvement of datasets and the improved algorithm, were summarized. Then a novel algorithm called adaptive random sampling algorithm was put forward based on balance maximization. The classification effect of random forest algorithm was further optimized. Experiments show that the proposed algorithm performs well with the imbalanced data, the new data are fitted with the original data, and it could improve the ability of classifier to deal with the imbalanced data.

Key words: imbalanced dataset; balance maximization; random sampling; random forest; data preprocessing

在处理非平衡数据集时,基于监督式学习的随机森林算法在分类性能方面并不理想,主要表现为分类结果集中,各个类别的分类错误情况参差不齐,即对于不同类别的分类,分类错误的情况差异较明显.国内外学者针对非平衡数据的处理进行了优化[1-3],本文在经典优化方法基础上,提出了一种新的随机抽样方法,以解决训练数据集的平衡性问题,进而使得随机森林算法的分类效果显著提高.本文所提出的算法是基于最大平衡度的自适应随机抽样算法,应用于随机森林算法的预处理阶段.值得注意的是,数据训练集的平衡度与各个类别的分类错误程度成反比,一般情况下与整体分类的错误程度成正比,即训练集数据越平衡,各类别的分类错误会越小,而普遍情况下,整体的分类错误会增加.

非平衡数据的产生无非是数据本身就是非平衡的或者是平衡数据在后天因为某些限制被重新划分后变成了非平衡数据.针对本身就是非平衡数据的情况而言,多数情况是在数据集中有至少一个类别中的数据是比较少的,而对于这些数据较少的类别也正是分类器最关心的,比如癌症数据监测[1]、交通异常数据的检测、信用额度异常

收稿日期: 2017-04-21
基金项目: 国家自然科学基金资助项目(61272209).
作者简介: 董立岩(1966-),男,吉林长春人,吉林大学教授,博士生导师.

检测等. 对于后天产生的非平衡数据集,较为常见的是在解决多分类问题时,如由于支持向量机[2]的限制,需要将多分类转换为二分类问题,这就导致了二分类训练的数据集是非平衡的.

1　数据平衡化方法

非平衡数据集在数据训练、分类等过程中带来很多问题,平衡化这些非平衡数据是较为重要的,对于数据平衡化的方法,可以从改造数据集使其变得平衡[3]或通过算法对数据进行改造使之平衡化这两方面进行.

1.1　改造数据集

改造初始数据集中的数据分布,使其从非平衡状态转变为平衡状态,是处理非平衡数据集较为常见也是最为直观的方法. 数据分布的平衡性可以通过保持高频数据的采样频率,以及增加低频数据的采样范围来进行判断. 过采样及欠采样这两种随机采样方法为较常用的方法.

欠采样方法指的是从上向下地随机删除一些高频数据,直到每个类别中的数据数目达到平衡. 过采样方法的最终目的也是为了让各个类别中的数据一致,使数据具有平衡性,满足传统算法对数据的需求,只不过是采用随机复制低频数据的方法. 这两种方法虽然可以满足数据的平衡性,但是也会让数据集出现新的问题,比如对于欠采样方法,因为是随机删除数据,在数据处理过程中加深数据稀疏度的同时也可能删除某些重要的高频数据. 对于过采样算法,复制增加的低频数据可能会没有任何作用,几乎无法在本质上解决数据不平衡的问题.

对于欠采样和过采样方法的不足,文献[4]详细阐述了改进的方法,通过构造数据,改善数据的稀疏性的同时增加数据的平衡性,避免了随机性盲目抽取的缺点. 本文也根据该思想进行了改进,提出了适用于非平衡数据的自适应随机抽样方法.

1.2　改造算法

改进算法处理数据集非平衡问题,往往会给每个数据赋予一个权重,以适应不同的类别,如 Boosting 算法. 代价敏感[5]方法是一个较为常见的算法改进,在处理非平衡数据时,针对不同的类别赋予不同的错分惩罚值,在下次迭代的过程中,这些惩罚值较高的数据样本会被着重处理,这样低频的分类样本就会获得较高的关注,从而提高低频分类的分类精度.

但是对于这类赋予错分惩罚值的方法,惩罚值的给定仍没有一个很好的策略.

2　基于最大平衡度的自适应随机抽样算法

2.1　平衡度定义

数据的平衡性在二分类的情况下是非常直观的. 若两个类别的样本数目基本一致,即可判断数据集具有良好的平衡性. 本文所要讨论的是多分类问题中的数据平衡问题,根据统计学中的标准差概念,引出了本文对于数据集平衡度(Bal)的定义:

$$\mathrm{Bal}(\boldsymbol{S}) = -\sqrt{\frac{\sum_{i=1}^{k}(c_i-\bar{c})^2}{k}}. \tag{1}$$

其中:k 为样本类别数目; c_i 为 i 类别的数据数目;$\bar{c}$ 为所有类别中样本数目的均值,定义如下:

$$\bar{c} = \frac{\sum_{i=1}^{k} c_i}{k}. \tag{2}$$

由式(1)可知,Bal 强依赖于数据集的大小,将 Bal(S)除以“不同类别样本数目的平均值”,从而实现了归一化.

$$\mathrm{Balance}(\boldsymbol{S}) = \frac{\mathrm{Bal}(\boldsymbol{S})}{\bar{c}}. \tag{3}$$

数据集的平衡度(Balance)与类别样本数目的标准差成反比,与数据稳定程度成正相关. 即平衡度越大,数据集越平衡,越稳定,类别样本数目的标准差越小. 这也恰恰说明通过最大化平衡度的方法来使数据达到平衡状态是可行的. 本文对方法的改进就是以数据集的平衡度为基础的,并提出了一个基于最大平衡度的自适应随机抽样算法(ARSA - BM).

根据 ID3 算法的思想,参考信息熵与信息增益的相关内容,本文提出了可以作为待构造类别选择标准的平衡度增益的相关概念. 进行每次随机采样时,平衡度最大的类别因为其平衡度增益也是最大的,故将此类别将作为待构造类别. 平衡度增益的定义如下.

$$\mathrm{Gain}(\boldsymbol{S},y)= \mathrm{Balance}(\boldsymbol{S}\cup\{(\hat{x},y)\}) - \mathrm{Balance}(\boldsymbol{S}). \tag{4}$$

式(4)中的$(\hat{x},y)$代表新构造出的数据样本,这个数据样本的类别是以 y 命名的.

对于每次随机采样,都需要利用式(4)求得平衡度最大的类别,所以给出计算平衡度率最大的类别的公式:

$$\text{argmaxBalance}(T,k) = \text{argmax}_{y\in(1,\cdots,k)}\text{Balance}(T\cup\{(\hat{x},y)\}) . \quad (5)$$

其中$(\hat{x},y)$代表了新构造出的数据样本,这个数据样本的类别是以 y 命名的.

2.2 ARSA – BM 算法介绍

SMOTE 算法[6]是处理二分类问题中数据不平衡的经典算法之一. 但是对于多分类问题中的非平衡数据,SMOTE 算法却无法直接使用,而且 SMOTE 算法仍存在一些不足和缺陷[7],如:构造新数据时,仅仅考虑低频类别数据的复制,而忽略了构造高频数据样本的概率;SMOTE 算法在构建低频类别数据时,有很大概率选择了低频数据中的噪声数据或边界数据. 由此可见,SMOTE 算法还无法处理随机森林多分类问题中的非平衡数据.

本文根据 SMOTE 算法的思想,对过采样方法进行优化,提出了最大平衡度的自适应随机抽样(ARSA – BM)算法. 该算法的主要目的是对随机森林算法中的非平衡数据集进行预处理,使之平衡化. ARSA – BM 算法根据式(1)中对数据集平衡度的定义,提出了抽样要使得平衡度最大化这个主要目标. 为了达到这个目标,采用了自适应反馈机制,取缔了盲目随机抽样的过程,而是在抽样过程中一致关注最大化平衡度类别的反馈信息.

2.3 ARSA – BM 算法描述

ARSA – BM

输入:M—初始数据集,b—数据集平衡度目标值(初始值为 0)
输出:T—平衡化数据集

```
1  T←M
2  tb←Balance(T)
3  while(tb > b)
4      i←argmaxBalance(T,k)
5    (x̂,i) ←createNew(T,i)
6      T←T∪{(x̂,i)}
7      tb←Balance(T)
8  endwhile
9  return T
```

在上述的算法框架中,通过第 3 步可以设定关于平衡度的阈值以达到自适应采样的目的,通过第 4 步进行自适应反馈,选择当前平衡度增益最大的类别. 通过第 3 步和第 4 步可以提升数据集的平衡度.

数据集平衡度并不能满足算法的需求,还需要通过 createNew 函数构造新的样本.

createNew

输入:数据集 T,新构造样本的类别 i
输出:构造出的新样本$(\hat{x},i)$

```
1  X_i←{x|(x,i) ∈ T}
2  x̄←Center(X_i)//计算 X_i 样本集合的中心 x̄
3  x←RandomSelect(S_i)
4  x̂← x + random(0,1)·(x̄ − x)
5  return(x̂,i)
```

在 createNew 函数中,通过第 3 步与第 4 步可以保证新样本的随机性. 而且第 4 步也避免了随机差值的盲目性,减少了噪声数据与边界数据.

3 ARSA – BM 算法实验分析

3.1 性能指标与评估方法

在以往平衡度测量的过程中,往往采用信息熵来判定平衡度. 但是信息熵对于指标值变动较小或指标值变动不规律的情况有很大的局限,而且对于单位指标的时间序列数据稀疏的情况,测量结果也不理想. 而且在非平衡数据的分类问题中,数据稀疏的类别可能具有更多的价值,在此情况下,信息熵作为测量标准也是不适合的.

二分类的分类性能评价指标包括分类精确度、分类准确率、平均召回率、宏平均召回率以及宏平均精确度[8]等,在这些指标中召回率越高,精确度的值就越低,反之召回率越低,精确度的值就越高,综合以上两个评测指标,定义了 $F1$ 指标用于整体的评测:

$$F1 = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} . \quad (6)$$

$F1$ 指标对于非平衡数据集而言是一个很有效的评价指标[9],但是 $F1$ 无法适用于数据集大于两个类别的多分类问题. 因此在多分类问题中,本文定义了两个指标用来评测分类性能,$F1(i)$为类别 i 的 $F1$ 指标值,Macro_$F1$ 表示所有类别的 $F1$ 指标值的平均值,这两个评测指标的公式如下:

$$F1(i) = \frac{2\cdot\text{Precision}(i)\cdot\text{Recall}(i)}{\text{Precision}(i)+\text{Recall}(i)} , \quad (7)$$

$$\text{Macro\_}F1 = \frac{1}{k}\sum_{i=1}^{k}F1(i) . \quad (8)$$

分类器评估方法多数是将数据集分为两部分,一部分数据集用来训练,另一部分数据集用于测试模型. 交叉验证法(cross validation,CV)和预留法(holdout)也是这样处理数据集的. 这两种方法是较为常见的分类器评估方法. CV 方法[10]是由统计学家 Seymour Geisser 提出的,主要用于评

估算法的性能与泛化能力，在一定程度上减少了过拟合现象.

CV 方法在评估算法时，需要不停地将数据集进行划分，一部分用来测试，一部分用来训练，这就导致评估过程的计算复杂度与空间复杂度增加，降低了评估的运行效率. 对于预留方法，会以一定的比例将原始数据集分为两部分，分别作为训练集与测试集，由于其操作简单而且运行效率高，所以本文采用预留法作为评估方法之一. 本文将原始数据集以 7∶3 的比例进行分割，其中 70% 的数据作为训练集，其余的作为测试集.

无论针对哪种分类算法的评估方法，其实质是评估算法的泛化能力，即评判算法过拟合或欠拟合的程度. 泛化能力越高，对于其他数据集的分类结果也更容易符合预期标准. 泛化误差描述了这种真实数据下分类误差的情况. TreeBagging 算法是随机森林算法的基础，总数据集随机抽样获得了每棵树的训练集，每个样本被抽取的概率约为 63% 左右，剩余的 37% 数据是无法被抽取出来的，这些无法抽取的数据称为 OOB 数据. 这些数据不能以训练数据的形式存在，而是被用来进行模型测试. 这样就可以运用较少的数据进行算法泛化能力的评估. OOB 误差评估方法保留 CV 方法中精度与粒度的同时，也获得了预留法的验证效率. OOB 误差估计法也被本文用来作为评估方法之一.

3.2　实验数据集

实验采用了搜狗实验室的 SogouC 数据集作为算法的实验数据. 数据包括了数十种类别数据以及十万篇文本数据，主要来源于搜狗新闻网站的文本信息（http://www.sogou.com/labs/dl/c.html）.

提供的初始数据集是较为平衡的数据集，为了达到非平衡数据集的状态，要对数据进行不放回抽样等预处理过程. 将初始完整数据集、抽样后的非平衡数据集以及算法平衡化的非平衡数据集进行比较，判断本文所提出算法是否具有在保证分类性能的前提下解决数据平衡性的能力.

3.3　实验结果与分析

本文将原始数据集以 7∶3 的比例进行分割，其中 70% 的数据作为训练集 S_{org}，其余的作为测试集. 按照这种数据分配进行预留法的评估，以分类准确度 $F1(i)$、宏平均指标 Macro_$F1$、宏平均精确度、OOB 误差以及测试集误差等作为评价指标.

将预处理后得到的非平衡数据集 S_{ub} 通过 ARSA－BM 算法进行平衡化，可以获得一个保证了分类性能的新的平衡数据集 S_b. 因此可以将 S_{org}，S_{ub}，S_b 这三个数据集用于分类性能的评估.

本文进行多次实验，将获得的多个数据集的评价数据取平均值作为最终评价数据，以减小随机森林的随机性给实验结果造成的随机误差.

在本次试验中，选定 10 个类别，250 个最大深度为 10 的构建树，特征子集以平方根的形式作为选择策略，并以 Gini 指标作为节点的分裂策略，样本共有 262 144 个特征维度. 测试结果如图 1～6 所示.

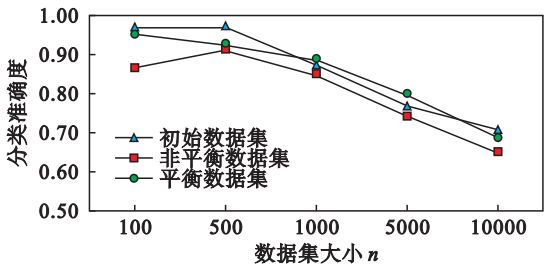

图 1　数据集平衡性对分类准确度的影响

Fig. 1　Effect of balanced dataset on the classification accuracy

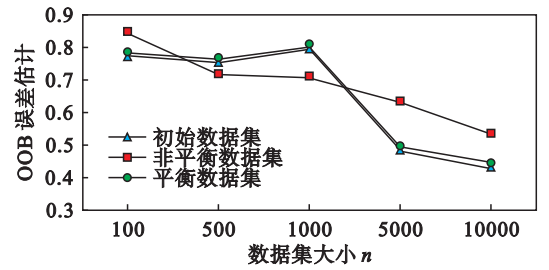

图 2　数据集平衡性对 OOB 误差的影响

Fig. 2　Effect of balanced dataset on the OOB error

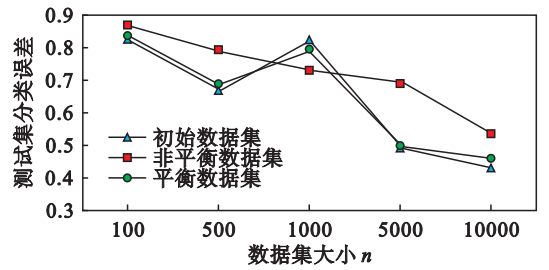

图 3　数据集平衡性对测试集分类误差的影响

Fig. 3　Effect of balanced dataset on the classification error of test set

通过实验结果可以发现，分类精确度与数据集的大小是成负相关的，OOB 误差与测试集的分类误差是正相关的. 由图 1～3 可知，当数据集大小为 100 时，分类准确度达到最高值，但是误差值是最高的，所以其泛化能力也是最低的，这就表明模型出现了过拟合的现象. 由图 2 与图 3 可知，当数据的规模大于 5000 后，分类误差以及泛化误差

呈下降趋势,且小于 50% ,表明模型具有了较好的泛化能力.

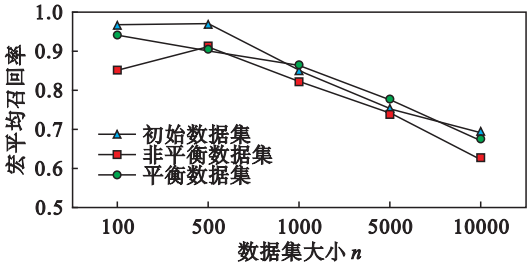

图 4 数据集平衡性对宏平均召回率的影响

Fig. 4 Effect of balanced dataset on the Macro_Recall

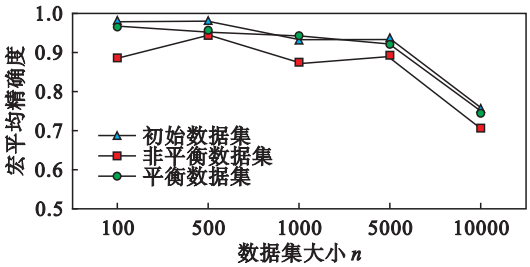

图 5 数据集平衡性对宏平均精确度的影响

Fig. 5 Effect of balanced dataset on the Macro_ Precision

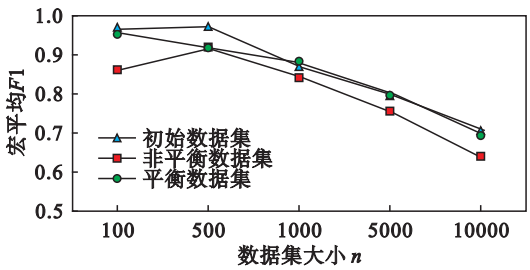

图 6 数据集平衡性对宏平均 *F*1 的影响

Fig. 6 Effect of balanced dataset on the Macro_*F*1

实验结果表明,本文提出的 ARSA – BM 算法,提升了宏平均精确度、宏平均召回率、整体评测指标 *F*1 这三个评价指标. 初始训练集以及平衡化后的非平衡数据集在 OOB 误差与训练集分类误差上趋势一致,表明 ARSA – BM 算法在保证分类性能的前提下可以解决数据平衡性问题.

总之,本文提出的 ARSA – BM 不仅可以用于二分类问题,对于多分类问题也有很好的实验效果. 而且可以在不影响分类性能的情况下,解决数据不平衡的问题,生成的平衡化数据可以很好拟合初始数据,分类性能还有所提高.

4 结 语

本文通过分析现有数据平衡化方法的不足,提出了一种基于最大平衡度的自适应随机采样算法,将搜狗文本分类语料库中数据预处理后作为实验数据. 实验结果表明,本文提出的算法可以提高随机森林算法在处理非平衡数据情况下的分类性能.

参考文献:

[1] Gray K R, Aljabar P, Heckemann R A, et al. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease [J]. *NeuroImage*, 2013, 65 (1): 167 – 175.

[2] Li X, Guo Y. Active learning with multi-label SVM classification [C]//Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Beijing, 2013: 1479 – 1485.

[3] Galar M, Fernández A, Barrenechea E, et al. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling [J]. *Pattern Recognition*, 2013, 46(12): 3460 – 3471.

[4] Zhu Q, Cao S. A novel classifier-independent feature selection algorithm for imbalanced datasets [C]//10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing. Daegu, 2009: 77 – 82.

[5] Hsu J L, Hung P C, Lin H Y, et al. Applying under-sampling techniques and cost-sensitive learning methods on risk assessment of breast cancer [J]. *Journal of Medical Systems*, 2015, 39(4): 1 – 13.

[6] Sarakit P, Theeramunkong T, Haruechaiyasak C. Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm [C]//2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA). Chonburi, 2015: 1 – 5.

[7] Ono S, Matsuyama H, Fukui K, et al. Error detection of oceanic observation data using sequential labeling [C]// IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris, 2015: 1 – 8.

[8] Kremic E, Subasi A. Performance of random forest and SVM in face recognition [J]. *The International Arab Journal of Information Technology*, 2016, 13(2): 287 – 293.

[9] Maratea A, Petrosino A, Manzo M. Adjusted F-measure and kernel scaling for imbalanced data learning [J]. *Information Sciences*, 2014, 257: 331 – 341.

[10] Geisser S. The predictive sample reuse method with applications [J]. *Journal of the American Statistical Association*, 1975, 70(350): 320 – 328.