**硕士学位论文**

针对类内不平衡样本分类方法的研究

**RESEARCH ON APPROACH FOR CLASSIFICATION OF INTRA-CLASS IMBALANCED DATA SETS**

**石凤兴**

**哈尔滨工业大学**

**2015年12月**

国内图书分类号：TP399 学校代码：10213

国际图书分类号：621.3 密级：公开

**工学硕士学位论文**

针对类内不平衡样本分类方法的研究

|  |  |  |
| --- | --- | --- |
| 硕士研究生 | ： | 石凤兴 |
| 导 师 | ： | 张春慨副教授 |
| 申请学位 | ： | 工学硕士 |
| 学科 | ： | 计算机科学与技术 |
| 所 在 单 位 | ： | 深圳研究生院 |
| 答 辩 日 期 | ： | 2015年12月 |
| 授予学位单位 | ： | 哈尔滨工业大学 |

Classified Index: TP399

U.D.C: 621.3

Dissertation for the Master Degree in Engineering

**RESEARCH ON APPROACH FOR CLASSIFICATION OF INTRA-CLASS IMBALANCED DATA SETS**

|  |  |
| --- | --- |
| **Candidate：** | Fengxing Shi |
| **Supervisor：** | Associate Prof. Chunkai Zhang |
| **Academic Degree Applied for：** | Master Degree in Engineering |
| **Speciality：** | Computer Science and Technology |
| **Affiliation：** | Shenzhen Graduate School |
| **Date of Defence：** | Dec., 2015 |
| **Degree-Conferring-Institution：** | Harbin Institute of Technology |

# 摘要

随着移动互联网技术的兴起，越来越多的原始数据被收集起来用于分析与挖掘，其中很多领域内的数据是严重不平衡的，既属于不同类别的样本数量相差极为悬殊。传统的机器学习方法通常采用全局分类准确率作为训练目标，在不平衡数据集上表现欠佳，所以不平衡数据分类算法也逐渐成为一个受人关注的课题。

目前主要的不平衡数据集分类方法分为两种，分别是数据重采样法以及提升算法，他们的主要思路是通过改变原始数据集中样本的权重或者比例来使原始数据达到一种“平衡”状态，从而适用于传统的分类算法，这类算法的结果与数据分布情况相关密切，往往需要一个复杂的调节过程才能得到一个较好的结果而且泛化能力不是很强。

本文首先介绍了不平衡数据分类中存在的难点和主要问题所在，以及不平衡数据分类的评价标准；然后简述了现有的经典不平衡分类算法，并从问题入手分析他们的主要思想，分析他们的优缺点；接着本文提出了以不平衡样本分类评价标准为最终训练目标的分类思路，通过评价标准构造一个可优化的函数作为训练目标进行训练，采用了整体数据集优化分类器的思路，使用神经网络模型，反向传播算法来使目标函数迅速收敛于一个最优解以达到不平衡样本训练的目的；最后通过与其他算法进行对比试验来证明了此算法的有效性。

关键词：不平衡样本；神经网络；最小化损失学

# Abstract

1. 绪论 ok

1、1研究背景与目的

1、2国内外研究现状

1、3主要研究内容

1、4章节结构安排

第二章 相关研究内容介绍

2、1不平衡数据集的分类评价标准

2、2classifiy phase max f measure algorithm（需要一个可以表现样本分布的概率模型）

2、3结构化支持向量结（首个在training phase 可以自己设定loss-function的算法，低秩矩阵）

2、4本章小结

第三章 算法介绍

3、1f值寻优的迭代神经网络算法

3、2f-measure loss 神经网络

第四章 实验效果

# 目 录

[摘要 I](#_Toc464304406)

[Abstract II](#_Toc464304407)

[第**1**章 绪 论 1](#_Toc464304408)

[1.1 研究背景与意义 1](#_Toc464304409)

[1.2 国内外研究现状 1](#_Toc464304410)

[1.2.1 经典不平衡处理方法 2](#_Toc464304411)

[1.2.2 传统分类器优化法 3](#_Toc464304412)

[1.3 主要研究内容 3](#_Toc464304413)

[1.4 章节结构安排 4](#_Toc464304414)

[第**2**章 相关研究内容介绍 5](#_Toc464304415)

[参考文献 12](#_Toc464304416)

[哈尔滨工业大学学位论文原创性声明和使用权限 15](#_Toc464304417)

[致 谢 16](#_Toc464304418)

1. 第**1**章 绪 论

## 研究背景与意义

随着大数据时代的到来以及各种网络化的系统变得规模越来越大，结构越来越复杂，各种监视、安全、金融等系统每天都获取巨量的数据，所以原始数据的分析与潜在信息的挖掘对分类与决策过程起着至关重要的作用。尽管现有的方法在数据挖掘领域已经获得了巨大的成功，但是在许多的实际问题中还存在着各种各样的问题，不平衡数据便是其中一个相对比较新而且比较艰巨的挑战[[1]](#endnote-2)。

不平衡数据分类问题是指不同类别样本数目相差悬殊情况下的分类学习问题。 以二分类为例，若其中有一类（正类、多数类） 的学习样本比另一类（负类、少数类）的学习样本多很多，那么就称这样的分类问题为不平衡样本分类问题[[2]](#endnote-3)。随着数据挖掘与机器学习从研究性课题逐渐变为了应用技术，以及更多的分类或预测等实际需求的出现，人们发现在某些领域，不平衡数据集往往比平衡数据集更常见，并且人们对不平衡数据集中的少数类样本的重视程度通常也高于多数类，如文本分类[[3]](#endnote-4)，欺诈识别[[4]](#endnote-5)，生物信息学[[5]](#endnote-6)等。

分类是机器学习和数据挖掘中重要的知识获取手段之一。常见的分类算法如决策树、贝叶斯网络、支持向量机和神经网络等已经被广泛的应用在了各类分类问题当中。然而经典的分类算法在面对不平衡数据时通常会出现少数类样本难以召回的情况，这种现象是由于现有的分类算法大多假定面对的训练数据集是大致均衡的，既各类所含的样例数大致相等[[6]](#endnote-7)。在此前提下经典的分类器算法的训练目标是取得最高的整体分类准确率，这也就导致了当数据集的分布情况不平衡并且多数类与少数类存在边缘略微重叠时，传统分类器算法会将边缘部分的样本直接归为多数类以提高全局准确率，但是这样会大大降低了少数类的召回率，尤其面对一些实际问题如疾病监测时[[7]](#endnote-8)，在数据批量处理环节，将一位患者误诊为无病的损失要远大于将一位没有患病的人误诊。所以，传统的经典分类器算法和普通的分类评价标准在用于不平衡数据集时，由于存在明显的缺陷所以已经无法直接应用[[8]](#endnote-9)。

综上，随着人们对于预测与分类的要求越来越高，不平衡数据集出现的越来越多，以及传统分类方法的效果无法被接收，所以不平衡样本分类算法的研究具有着重要的理论意义和实用价值。

## 国内外研究现状

不平衡样本分类问题从提出至今约有十余年[[9]](#endnote-10)，期间有很多的相关研究，在这里大体可以将所有的解决方法其分为两类：a）经典不平衡分类处理方法，这一类算法的主要思路是将不平衡数据集经过一系列处理或者分类过程中区别对待数据集中的实例，从而将原始数据转变为一个相对的“平衡态”；b）传统分类器优化法，这一类算法并没有经典不平衡处理方法那么主流，并且没有一个普遍的处理规律，他们通常是将传统分类算法的训练结果进行进一步的针对不平衡数据的调整，从而使训练结果可以适应或者处理不平衡数据集。

### 经典不平衡处理方法

根据现有的研究成果，经典的解决不平衡数据集的主要方法可以总结为以下几种：a）原始训练数据的重构，即通过重采样技术使原始数据变得相对平衡；b）通过集成学习的方法，通过训练多个弱分类器的，最后通过投票或权值相加的方法得出分类结果；c）敏感代价学习方法，也就是改变原始数据在评判标准中的权值，通常使用敏感代价矩阵来协助计算分类损失。

重采样方法是一种思想较为简单并且相对行之有效的方法，按照采样过程重采样方法又可以分为三类，分别是过采样，欠采样和混合采样法[[10]](#endnote-11)。其中过采样法是通过使用现有的少数类样本点生成一些新的少数类样本，可以是简单的复制，也可是进行一些插值，比较有代表性的方法为Chaudhuri等人提出的随机过采样法[[11]](#endnote-12)，Chawla等人提出的smote法[[12]](#endnote-13)，以及一些在此基础上优化的过采样方法[[13]](#endnote-14)[[14]](#endnote-15)，然而仅仅通过原始的样本点生成新的样本可能存在一些问题比如过拟合和改变原始分布等。与过采样法相反的就是欠采样法，欠采样法便是通过去掉一部分多数类样本点已达到一个数据的平衡，欠采样法相对于过采样法稍微复杂并且在许多场景下也更合理一些[[15]](#endnote-16)，除了早期简单的随机欠采样法之外，又有多项研究分别设置了不同的欠采样规则，在相应数据集上或者相应的经典分类算法均取得了一定的效果。如Zhang等人提出的基于聚类的欠采样方法[[16]](#endnote-17)，Kubat等人提出的基于距离的单边欠采样法[[17]](#endnote-18)，以及去除临近点的Tomek-link法[[18]](#endnote-19)等等。然而当多数类与少数类样本相差过于悬殊的时候，单独使用过采样方法和欠采样方法可能已经无法满足需求了，所以这时需要混合使用两种采样方法，许多研究表明[[19]](#endnote-20)[[20]](#endnote-21)，混合使用两种采样方法通常效果会好于单独使用一种方法。

集成学习方法主要分为两类，分别是训练集分割方法以及提升算法。训练集分割方法是指将多数类分割为多个子集，每一个子集的样本数目与少数类样本数目相当，然后用同种或不同种的多个分类器分别训练，每个分类器的训练集为少数类和一个多数类子集的组合，这类方法[[21]](#endnote-22)与普通的重采样法相比的优点是使用了多个分类器投票机制。另一类提升算法也是通过训练多个分类器通过使用权重的机制来控制分类结果，不过与分割法的区别是他每次训练的时候都是采用整个训练集，不过每次训练之后根据当前分类器的分类结果来更新训练集中每个样本对下次训练时的影响权重，其中比较有代表性的方法为bagging算法和Adaboos算法[[22]](#endnote-23)。

敏感代价方法的思想是通过人为设置分类错误的损失值来调整不平衡数据分类过程，以二分类为例可以调高正例（少数类）样本的分类错误损失比例，降低负例（多数类）的分类错误损失比例来解决不平衡问题，虽然使用敏感代价方法或者基于敏感代价方法的研究也有很多，不过在不平衡样本分类问题中，其本质和效果与重采样法并无二致。

以上方法本文中称之为经典不平衡处理方法，可以发现他们都是从不平衡数据分类问题的数据角度入手，采用直接设法改变样本数目的方法或者改变样本权重的方法，旨在一次性或逐渐降低原始的中的不平衡问题，来达到一种相对的“平衡”状态，然后在构建模型期间采用的都是传统的分类算法，并来按照普通问题进行处理。

### 传统分类器优化法

传统分类器优化算法有一个共同的特点，就是针对不平衡样本分类问题，他们不再是想经典不平衡处理方法那样针对不平衡的数据集进行一些平衡化的调整，他们的主要思想是通过修改分类器的训练过程或者分类过程来适应不平衡的数据集，既通过优化算法的训练过程来减轻不平衡分布对训练过程的影响，或者采用正常的训练思路训练模型后，通过一系列其他过程进行对模型的调整，又或是得到普通的模型之后在分类阶段采用与经典分类阶段不同的方法来解决不平衡问题。

多数的传统分类器优化法并没有一个统一的思路和方案，他们通常采用不同的原理但是殊途同归，都可以解决不平衡样本分类问题。如Cheng等人提出的基于极限学习机（Extreme Learning Machine）的边界移动算法[[23]](#endnote-24)，就是通过极限学习机把原始问题当做普通问题训练之后在采用核密度估计以及分类边界移动的方法来解决不平衡问题；Jansche等人提出的最大化F1值期望的逻辑回归训练法[[24]](#endnote-25)，则是通过采用一种在训练过程中求解F1值的期望与当前模型的关系，并通过此关系不断的迭代去优化F1值得期望，进而去生成一个可以适应不平衡样本的模型；Dembczynski等人提出的F1值优化算法的思路则是在分类阶段与传统分类算法有所不同，这种算法可以适用于各种概率模型，通过先求解当把多少个实例划分为少数类时可以获取最大的F1值期望，之后在把所有实例按照可能属于正例的概率进行排序并分类，从而可以在近似时间内完成整体分类过程[[25]](#endnote-26)。

## 主要研究内容

本文主要进行了不平衡样本分类算法的研究与一种基于最小化损失的不平衡样本分类算法的设计。

首先通过对现有的机器学习传统方法与经典的不平衡样本分类算法进行分析，得到不平衡数据集分类时的主要问题与难点。

其次对F1值优化法进行改进，由于该方法更倾向于得到了一个对空间拟合相对准确的模型，而直接对不平衡样本数据进行训练并无法得到，所以采用了神经网络模型，并设计了一个迭代过程，在每一步通过F1值优化算法求解当前步骤的最优解，并通过最优解与目标解之间的偏差去重新优化模型，进而得到一个相对样本空间拟合相对准确的模型。

最后，由于上一算法的时间复杂度过高，每次迭代过程需要使用规模的时间进行求解，所以再次对上一算法进行优化，从不平衡样本分类的评价标准入手，其中不平衡样本的评价标准根据不同的问题有很多种，本文中采用了最常用的F1值作为优化目标，并也采用了整体优化的思路，根据F1值的计算过程，建立了一种F1值的期望与训练过程中的分类器的数学关系，并使用此关系通过梯度下降的思路和神经网络模型，反向传播算法对模型参数进行求解，使得可以得到最大的F1值。

## 章节结构安排

本文主要结构包含如下内容：

（1）绪论：首先简要的介绍了本课题的研究背景与意义，然后分类讨论了国内外的研究现状，并结合不平衡样本分类问题的关键点分析了各种算法的主要思想与方法，最后介绍了本文的主要研究内容与章节安排。

（2）相关理论及算法介绍：这一部分首先介绍了不平衡样本分类问题中的一些常见的评价标准，然后介绍了分类过程中求解最大F1值以及结构化支持向量机这两种算法，为第三章中用到的一些理论进行一定的铺垫。

（3）算法介绍：首先介绍介绍了基于最大化F1期望的神经网络训练算法，然后通过分析该方法主要存在的问题，并从直接改进传统分类算法的训练过程的以适应不平衡数据集的角度，提出了基于损失最小化参数学习的F1值神经网络训练法。

（4）实验结果与分析：首先对以上算法进行验证性实验来证明算法的有效性，然后与其他不平衡样本分类算法在相同数据集上分别进行对比试验以验证算法的可靠性。

最后得出结论，总结全部的工作并提出接下来的展望。

1. 第**2**章 相关研究内容介绍

通常，机器学习是一种通过数学模型来拟合一个样本空间的问题，由于不可能获得到样本空间中所有可能出现的点集，所以通常采用该空间中的一部分点作为模型训练的集合，通过对于不同算法，采用不同的训练方法从训练集中求解出归纳偏置从而预测或分类空间中没有被采集到的未知的样本[[26]](#endnote-27)。然而现有的传统的分类方法都是以训练集分类错误最少为训练目标的，所以他们都默认各个类别的样本数目大致均衡，或者处于两类边界部分的样本点密度相差不大。而不平衡数据集恰恰不满足这个条件，无论是评价标准以及传统分类方法均已不能完美适用于不平衡数据集，所以需要有专有的分类方法进行分类。

本章首先介绍了不平衡样本分类评价标准进行讨论，然后选取其中最常用的F1值进行作为本文的评价标准，最后介绍了两种第三章中用到或借鉴到的算法，并对其进行一定的分析和总结。

## 不平衡样本分类评价标准

不平衡样本分类问题通常为二分类问题（多数类和少数类），或者多分类的不平衡样本问题通常可以简化为多个二分类问题，所以评价标准通常是针对二分类问题所设计的。下表2-1为二分类问题的结果混淆矩阵（Confusion Matrix）。

表2-1 混淆矩阵

|  |  |  |
| --- | --- | --- |
| 真实类标 | 预测类标 | |
| 正例 | 负例 |
| 正例 | TP | FN |
| 负例 | FP | TN |

其中TP表示正确预测得到正例个数，FP表示把负例预测为正例的个数，FN表示把正例预测为负例的个数，TN表示正确预测得到负例的个数，二分类问题中我们通常将少数类视为正例，多数类视为负例。

对于二分类问题，无论是传统的分类问题或者是不平衡样本分类问题，他们的评价标准通常都是通过混淆矩阵中TP、FN、FP、TN四个参数的数学计算求得的。传统的分类方法通常使用的是全局准确率作为评价标准，其计算方法如公式2-1所示。

 (2-1)

通过上式可以看出，如果负例与正例的比例相差很悬殊，那么正例的分类正确个数TP对分子的影响很小，并且正例的总体个数TP+FP对于分母的影响也非常小，所以整体来说正例对于全局准确率的影响远小于负例，因此研究者们设计了一些其他的评价标准，如原子评价标准，F1值，G-measure，A-mean，以及Roc曲线，Aoc面积等来评价不同的分类问题。

Part2、相关介绍：

1. 不平衡样本分类评价标准，针对不同类型的问题，f-measure ， am-measure。
2. 针对不平衡样本的评价标准，经典算法classification error base不再适应。传统不平衡样本处理方法如采样法与集成学习法都是在经典算法的基础上改变样本空间来解决不平衡问题。引出以下两种算法框架。
3. structural support vector machine(original)：作者提出了一套对自定义损失函数的训练方法，但是针对不同的损失函数训练过程中存在不同的解决方法，主要是一步argmax的优化，作者以f-measure举例优化。
4. Gradient descent based：两篇文章采用逻辑回归，不过在算法中应用了部分搜索方面的过程，不完全是梯度下降，其一是以f-measure为目标 ，另一个是以am为训练目标

Part3、研究内容：

1. 针对ssvm，提出以am为训练目标的解决方案，从而更好的解决两类不平衡样本分类问题，但是ssvm是一种迭代算法，不能避免每次的解决qp问题以及argmax问题，导致速度慢，如果使用核函数并对误差要求比较高的话可能导致无法训练出结果，如使用liner kernel，根据作者后来提出的算法可以在多项式时间内求解出结果
2. Gradient descent based：对f-measure 和 am 采用结构化的思想进行函数拟合， 采用ann算法，以回归的训练思路进行训练。

评价标准：

通过复合计算混淆矩阵中的结果可以得到二分类问题公认较为有效的评价标准。

下列公式2-1至2-n分别列出了常用的评价标准：

真阳率（True Positive Rate）即正例召回率，查全率：

(2-1)

伪阳率（False Positive Rate）：

FPR=FP/(TN+FP) (2-2)

真阴率（True Negative Rate）：

TNR=TN/(TN+FP) (2-3)

伪阴率（False Negative Rate）：

FNR=FN/(TP+FN) (2-4)

准确率（Accuracy Rate）：

ACC=(TP+FN)/(TP+FN+FP+TN) (2-5)

正例准确率，也是查准率：

*Precision=TP/(TP+FP)* (2-6)

G评分标准：

(2-7)

F评分准则：

(2-8)

以上最常用于不平衡样本二分类问题的准则为正例召回率，准确率，F评分准则，以及全局准确率。

数据集介绍

当前的实验数据集包括3个UCI数据集分别是YEAST，abalone及glass，一个来自威斯康星大学（University of Wisconsin ）医学院的数据集Breast Cancer，以上四个数据集经常被用于检测不平衡样本分类算法。

这五个数据集的具体参数如下表：

表2-1 数据集参数表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 数据集 | 样本个数 | 少数类比例 | 属性个数 |
| 1 | YEAST | 1484 | 12.60% | 8 |
| 2 | Abalone | 4177 | 8.02% | 8 |
| 3 | Glass | 214 | 23.83% | 10 |
| 4 | Breast Canser | 699 | 34.50% | 9 |

本课题目前研究的内容主要是针对二分类问题，所以我们将包含多类的那些数据集人工分为了多数类和少数类，其中我们人工选取多数类与少数类的区分阈值，并且优先保证相近的数据被分到了同一类以保证数据的平滑和归纳偏置的存在。

**1**

Loss-Minimizing Learning





Let is loss function

Find ANN matrix Lead to minimal



**2**

Loss function-F-measure

So let

, 

Non-linear performance

Measure

The hypothesis is from

 to 



And the loss function is



Find  lead to minimal 

**3**

Confusion matrix

Predicted

Predicted

Ture 

And let y is the true laber of train. Set,let z is the predict laber of train set





And 

**4**





**5**



 



**Case 1 output layer**









**Case2 hidden layer**





部分实验结果：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Svm | Ssvm | FL-ann | F Optimal Iterative |
| Yeast | 0.46 | 0.78 | 0.73 | 0.70 |
| Abalone | 0.28 | 0.46 | 0.49 | 0.47 |
| Glass | 0.85 | 0.89 | 0.87 | 0.85 |
| Breast | 0.94 | 0.94 | 0.94 | 0.94 |

结束

1. 参考文献
2. 哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《针对类内不平衡样本分类方法的研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名： 日期： 年 月 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名： 日期： 年 月 日

导师签名： 日期： 年 月 日

# 致 谢

逝者如斯，已逝两载星霜，朱颜辞岁，却是仰取俯拾。在此我要向身边陪伴我一路走来的人表示由衷的感谢。

首先要感谢的一定是我的导师张春慨副教授。张老师他生活中为人谦和，风趣幽默，课余时间一直像长辈一样对我们悉心照料，关爱有加。工作中认真负责，治学严谨。教学中条例清晰，循循善诱，尤其是在与我讨论论文的时候，经常一两句话就令我豁然开朗，受益匪浅。

其次我要感谢我的家人，长我育我，顾我复我，对我百般呵护，无微不至，是你们给了我前进的力量。

最后，要感谢的是身边的同学和朋友，无论是点头之交或是莫逆之交，无论已成为匆匆过客还是将风雨同舟，你们都已给我带来了可以永藏的回忆。

1. Menon A K, Narasimhan H, Agarwal S, et al. On the statistical consistency of algorithms for binary classification under class imbalance[C]// International Conference on Machine Learning. 2013:603-611. [↑](#endnote-ref-2)
2. He H, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on knowledge and data engineering, 2009, 21(9): 1263-1284. [↑](#endnote-ref-3)
3. Tomanek K, Hahn U. Reducing class imbalance during active learning for named entity annotation[C]//Proceedings of the fifth international conference on Knowledge capture. ACM, 2009: 105-112. [↑](#endnote-ref-4)
4. Marqués A I, García V, Sánchez J S. On the suitability of resampling techniques for the class imbalance problem in credit scoring[J]. Journal of the Operational Research Society, 2013, 64(7): 1060-1070. [↑](#endnote-ref-5)
5. Fattahi S, Othman Z, Othman Z A. NEW APPROACH FOR IMBALANCED BIOLOGICAL DATASET CLASSIFICATION[J]. Journal of Theoretical & Applied Information Technology, 2015, 72(1). [↑](#endnote-ref-6)
6. Chawla N V, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 1-6. [↑](#endnote-ref-7)
7. Yoon K, Kwek S. A data reduction approach for resolving the imbalanced data issue in functional genomics[J]. Neural Computing and Applications, 2007, 16(3): 295-306. [↑](#endnote-ref-8)
8. Huang K, Yang H, King I, et al. Learning classifiers from imbalanced data based on biased minimax probability machine[C]//Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, 2004, 2: II-558-II-563 Vol. 2. [↑](#endnote-ref-9)
9. Japkowicz N. Learning from imbalanced data sets: a comparison of various strategies[C]//AAAI workshop on learning from imbalanced data sets. 2000, 68: 10-15. [↑](#endnote-ref-10)
10. Batista G E, Prati R C, Monard M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 20-29. [↑](#endnote-ref-11)
11. Chaudhuri S, Motwani R, Narasayya V. On random sampling over joins[C]//ACM SIGMOD Record. ACM, 1999, 28(2): 263-274. [↑](#endnote-ref-12)
12. 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: 321-357. [↑](#endnote-ref-13)
13. Wang S, Li Z, Chao W, et al. Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning[C]//The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, 2012: 1-8. [↑](#endnote-ref-14)
14. Gao M, Hong X, Chen S, et al. Probability density function estimation based over-sampling for imbalanced two-class problems[C]//The 2012 International Joint Conference on Neural Networks (IJCNN). IEEE, 2012: 1-8. [↑](#endnote-ref-15)
15. Drummond C, Holte R C. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling[C]// ICML Workshop on learning from imbalanced datasets II. 2003, 11. [↑](#endnote-ref-16)
16. Zhang Y P, Zhang L N, Wang Y C. Cluster-based majority under-sampling approaches for class imbalance learning[C]//Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on. IEEE, 2010: 400-404. [↑](#endnote-ref-17)
17. Kubat M, Matwin S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection[C]//In Proceedings of the Fourteenth International Conference on Machine Learning. 1997. [↑](#endnote-ref-18)
18. Provost F, Fawcett T. Robust classification for imprecise environments[J]. Machine learning, 2001, 42(3): 203-231. [↑](#endnote-ref-19)
19. Japkowicz N, Stephen S. The class imbalance problem: A systematic study[J]. Intelligent data analysis, 2002, 6(5): 429-449. [↑](#endnote-ref-20)
20. Zhang C, Jiang J, Shi F. Research on approach for classification of Within imbalanced data sets[C]//Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016: 85. [↑](#endnote-ref-21)
21. Wang C, Hu L, Guo M, et al. imDC: an ensemble learning method for imbalanced classification with miRNA data[J]. Genetics and Molecular Research, 2015, 14(1): 123-133. [↑](#endnote-ref-22)
22. Freund Y, Schapire R E. A desicion-theoretic generalization of on-line learning and an application to boosting[C]//European conference on computational learning theory. Springer Berlin Heidelberg, 1995: 23-37. [↑](#endnote-ref-23)
23. Cheng K, Chen Q, Yang X, et al. Classification of imbalanced bioinformatics data by using boundary movement-based ELM[J]. Bio-Medical Materials and Engineering, 2015, 26(s1): S1855-S1862. [↑](#endnote-ref-24)
24. Jansche M. Maximum expected F-measure training of logistic regression models[C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005: 692-699. [↑](#endnote-ref-25)
25. Dembczynski K J, Waegeman W, Cheng W, et al. An exact algorithm for F-measure maximization[C]//Advances in neural information processing systems. 2011: 1404-1412. [↑](#endnote-ref-26)
26. Mitchell T M. The need for biases in learning generalizations[M]. New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ., 1980. [↑](#endnote-ref-27)