**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；最小化损失学习

# Abstract

1. 绪论

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

# 绪 论

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得巨量的数据，所以原始数据的分析与潜在信息的挖掘对决策过程起着至关

## 研究背景与意义：

不平衡数据分类考虑的是各类样本数目不平衡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）的学习样本多得多，那么就称这样的分类问题为不平衡样本分类问题。在现实生活中有很多问题都属于不平衡样本分类问题，如文本分类，欺诈识别等，传统的分类算法主要适用于各类样本数量大致均衡的情况，其评价标准主要是基于精度的，而不平衡样本分类中精度已经无法准确的评价分类效果了，所以传统的分类算法在不平衡数据集上的表现也很差。

随着人们对数据挖掘研究的深入，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加常见，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，通常会出现少数类样本难以召回的情况，所以不平衡数据分类的问题越来越受到国内外学者的广泛关注，并成为一些学者在数据挖掘领域重点挑战的课题之一。

针对不平衡数据分类问题中具有这么多需要考虑的不平衡问题，传统分类器在遇到这些问题的时候往往表现的不够好。首先，由于传统的分类器算法的最终目标就是使最终全局分类准确率高，这也就导致了当面对不平衡数据分类时，传统分类器算法会倾向的将少数类数据样本直接归为多数类，大大降低了少数类的分类准确率。其次，虽然少数类的数量少，但是蕴含的价值信息却是很重要的，往往正是我们需要去关注的地方。所以，传统的经典分类器算法和普通的算法评价标准都已经不能再应用到不平衡数据分类领域。

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。

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。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，即各类所含的样例数大致相等。当遇到类数据严重不平衡时，把总体分类精度作为学习目标的传统分类算法会过多地关注多数类，而使少数类样本的分类性能下降。然而在实际应用中，有一类问题，他们的共同特点是存在着不同类别之间数据不均衡的特点，并且少数类样本的误差代价远高于多数类，例如信用卡欺诈问题，网络入侵检测，部分医疗检测，如直接套用常用的分类算法，将无法得到一个令人满意的效果。

根据现有的研究成果，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，或者使用敏感代价矩阵与混淆矩阵相乘的方法来重新评估分类结果。

然而上述方法均可以认为是从数据集入手来解决问题，采用各种不同的方案来使数据集达到一种“平衡”态，从而继续套用常用的分类算法来解决问题。本文提出了一种从分类器角度入手的方案，采用重新定义了损失函数的方法，通过不平衡分类结果来构造损失函数，来达到不平衡数据集训练的目的。

本文首先介绍了一种优化特定损失函数的思路，之后构造了一种适用于不平衡数据集的损失函数，通过采用神经网络模型以及反向传播算法来使得损失值迅速的收敛于一个最优解，来完成不平衡样本分类模型的训练，最后通过与其他算法的对比来验证了本算法的有效性。

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：

由于不平衡样本分类的特殊性，全局准确率并不能代表整个分类器在数据集上的表现，所以我们引入了一些其他的分类标准来评价不平衡样本分类结果，其中包括混淆矩阵，查全率，查准率，F-measure，G-measure，ROC曲线与AUC面积等等。

下表2-2为二分类问题的混淆矩阵，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。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。

表2-2混淆矩阵

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

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。

# Algorithm：

**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 |