**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 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

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合理一些[[15]](#endnote-16)，除了早期的随机欠采样

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。

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。

第2章 相关研究内容介绍

**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

参考文献

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

转眼之间，两年半的硕士研究生生涯已然接近尾声。值此论文完成之际，向在此期间给我无私帮助的所有老师、同学以及亲朋好友表示由衷的感谢。

首先，我衷心的感谢我的导师张春慨副教授，在这两年半时光里对我的精心培养。从本课题最开始的选题、开题、中期检查、以及最后的定稿与答辩，我都深深的受益于张老师地精心指导与谆谆教诲。 张老师治学严谨、为人谦逊，在论文完成期间，提出了很多建设性的意见；生活上，他更是以一位长辈的身份悉心照料，使我收获颇丰。

其次，我在这里要感谢我的父母和大哥。感谢你们一直以来在生活和精神上给我的莫大支持，是你们赋予了我不断进步的力量，决心与信心。

然后，还要感谢一直伴随我成长的实验室兄弟们——陈龙飞，肖鹏，熊腾飞和胡志亮。在学习上，课题研究与撰写中我们相互借鉴，共同成长，带给我很多建设性的建议和灵感。我的毕业论文可以顺利完成也有一定程度上要归功于你们所给予的无私帮助；短暂同窗，一生难忘。

最后，还要感谢我学习过，生活过的母校——哈工大。“规格严格，功夫到家”的校训，将时刻鞭策着我不断向前，使我受益终生。

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)