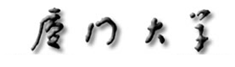
**学校编码：10384 分类号\_\_\_\_密级\_\_\_\_**

**学号:23020131153175 UDC\_\_\_\_**



硕 士 学 位 论 文

**不平衡数据的最优分类阈值研究**

**Research on Optimal Classification Threshold of Imbalanced Dataset**

谢 思 发

指导教师姓名: 吴梅红副教授

专 业 名 称：计算机科学与技术

论文提交日期：2016年 月

论文答辩时间：2016年 月

学位授予日期：2016年 月

答辩委员会主席：

评 阅 人：

年 月

**厦门大学学位论文原创性声明**

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（ ）课题（组）的研究成果，获得（ ）课题（组）经费或实验室的资助，在（ ）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年 月 日

**厦门大学学位论文著作权使用声明**

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（ ）1.经厦门大学保密委员会审查核定的保密学位论文，于 年 月 日解密，解密后适用上述授权。

（ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年 月 日

摘 要

不平衡数据分类，是目前在机器学习和数据挖掘领域里一个研究热点。不平衡分类问题在现实生活中经常出现，而传统的分类器在解决这类问题时，会出现性能大幅下降的情况。从本世纪初开始，国内外学者对不平衡分类问题进行了深入的研究，并发表了众多相关的研究成果。不过对该问题还是存在许多可以研究的方面。

在本文中，首先介绍了在数据分布不平衡场景下，分类的一些特性，简单地回顾了数据分布不平衡在机器学习及现实应用中所造成的一系列问题，介绍国内外学者对该问题的研究现状。深入地讨论数据的内在特征，在数据不平衡分类时引起的一系列问题。这些内在特征包括：小析取项问题；训练集密度不足问题；类别覆盖问题；噪音数据；数据分布偏移问题，同时给出了对应的处理方法和建议。总结目前解决数据分布不平衡分类问题的主要方法，特别地，将介绍数据预处理，代价敏感学习和集成技术三种方法。

接着深入的介绍用来衡量不平衡分类性能的指标，探讨不同应用场景下，需要使用的合适指标。特别地，将分析ROC曲线下面积(AUC)，单独做为度量分类器性能标准的不足之处:ROC只能衡量模型对样本排序的好坏，为取得好的分类结果，还需找到一个合适的划分阈值。最后提出了一个用来寻找最优分类阈值的框架:根据类别的错分代价是否相同，分别选取了F值和整体错分代价作为分类结果的衡量指标。先寻找训练集的最优分类阈值，然后用这个最优阈值去指导测试集或未知数据集的分类。将该框架应用到蛋白质的远程同源检测问题中。实验表明本文提出的框架能有效的工作，可以指导分类器的分类，提高分类器在查全率、查准率和F1值等指标的表现。

**关键词**：不平衡分类；评估指标；分类阈值；

ABSTRACT

Research on imbalanced classification problem is currently a hot topic in the field of machine learning and data mining. In the imbalanced classification scenario, the number of examples representing the class of interest is much less than that of the other classes. In many real-world applications, imbalanced classification is very common, and usually cause a dramatic drop in traditional classifier’s performance. The classifier learned from imbalanced data might has a great bias. From the beginning of this century, more and more researchers begin to pay attention to this problem. Many scholars have conducted a discussion in-depth on this issue, and published a lot research results in this aspect. However, there still exist many aspects that need to be taken into consideration in imbalanced classification problems.

In this paper, a short review of issues in machine learning and real applications of imbalance classification was firstly carried out. Then we summarized the current main methods to solve imbalanced classification through analysis of domestic and international research to this problem. After that, the main issues that related to data intrinsic characteristics in imbalanced classification problem were deeply discussed, which would contribute to improve the existing models. Those characteristics include: small disjunction problem, the lack of training data, the overlapping between different classes, the presence of noisy data and the dataset shift. Furthermore, we introduced a few methods and suggestions to solve these problems in conjunction with imbalanced classification. Summarizing the main methods to solve the problem of imbalanced classification Specially, we described preprocessing, cost-sensitive learning and ensemble techniques.

Finally, we carried out a thorough study of the metrics that measured the performance of classification, and discussed the appropriate metrics to be used in different application scenarios. In particular, we would start by analyzing the drawback of using AUC(Area Under ROC Curve) as the single measurement in imbalance data classification problem. AUC only shows the ranking power of positive prediction probability. To obtain a good classification result, a proper prediction probability threshold was also as important as a perfect ROC curve. According to whether the misclassification costs are equal, F-score or overall misclassification costs were recommending as another measurement besides AUC. And a novel framework was proposed to find the best classification threshold, which was applied to bioinformatics. The experiments indicated that the proposed framework can work effectively, which can be used to guide classification, and improved the classifier’s performance in precision and recall.

**Key Words:** imbalance classification; evaluation criteria; classification threshold;

目录

[摘 要 I](#_Toc451266409)

[ABSTRACT II](#_Toc451266410)

[第一章 绪 论 1](#_Toc451266411)

[1.1 课题背景及意义 1](#_Toc451266412)

[1.2 不平衡分类问题的挑战 2](#_Toc451266413)

[1.3 国内外研究现状 3](#_Toc451266414)

[1.4 本文组织结构 4](#_Toc451266415)

[第二章 不平衡数据分类的相关研究综述 6](#_Toc451266416)

[2.1 数据内在特征与分布不平衡问题 6](#_Toc451266417)

[2.1.1 小析取项问题 6](#_Toc451266418)

[2.1.2 密度不足 8](#_Toc451266419)

[2.1.3 类别覆盖 10](#_Toc451266420)

[2.1.4 噪音数据 12](#_Toc451266421)

[2.1.5 数据偏移 14](#_Toc451266422)

[2.2 解决不平衡数据分类的策略 15](#_Toc451266423)

[2.2.1 重采样技术 15](#_Toc451266424)

[2.2.2 代价敏感学习 18](#_Toc451266425)

[2.2.3 集成方法 20](#_Toc451266426)

[2.3 本章小结 21](#_Toc451266427)

[第三章 衡量不平衡分类性能的指标 22](#_Toc451266428)

[3.1 单一的评估指标 22](#_Toc451266429)

[3.1.1 准确率(Accuracy)与错误率(ErrorRate) 22](#_Toc451266430)

[3.1.2 查准率(Precision)和查全率(Recall) 23](#_Toc451266431)

[3.1.3 F-measure 23](#_Toc451266432)

[3.1.4 敏感性(Sensitivity，SN)和特异性(Specificity，SP) 24](#_Toc451266433)

[3.1.5 几何均数(Geometic mean，GM) 25](#_Toc451266434)

[3.2 ROC曲线与AUC值 25](#_Toc451266435)

[3.3 Precision-Recall曲线 27](#_Toc451266436)

[3.4 代价敏感错误率与代价曲线 28](#_Toc451266437)

[3.5 不平衡多分类评价指标 30](#_Toc451266438)

[3.6 本章小结 31](#_Toc451266439)

[第四章 最优分类阈值 32](#_Toc451266440)

[4.1 ROC 的缺陷 32](#_Toc451266441)

[4.2 寻找最优分类阈值 34](#_Toc451266442)

[4.2.1 错分代价相同的最优分类阈值 35](#_Toc451266443)

[4.2.2 错分代价不同的最优分类阈值 40](#_Toc451266444)

[4.2.3 最优分类阈值运用 42](#_Toc451266445)

[4.3 本章小结 44](#_Toc451266446)

[第五章 蛋白质远程同源检测 45](#_Toc451266447)

[5.1 数据集 45](#_Toc451266448)

[5.2 特征提取 47](#_Toc451266449)

[5.3 实验比较 49](#_Toc451266450)

[5.3.1 统一阈值 49](#_Toc451266451)

[5.3.2 最优阈值的使用方式 49](#_Toc451266452)

[5.3.3调节因子 50](#_Toc451266453)

[5.3.4 最优分类结果上界 51](#_Toc451266454)

[5.3.5 AUC值与F1值 52](#_Toc451266455)

[5.4 本章小结 53](#_Toc451266456)

[第六章 总结与展望 54](#_Toc451266457)

[6.1 本文总结 54](#_Toc451266458)

[6.2 未来展望 54](#_Toc451266459)

[参考文献 56](#_Toc451266460)

[攻读硕士学位期间发表的学术论文 62](#_Toc451266461)

[致 谢 63](#_Toc451266462)

CONTENTS

[Abstract(CN) I](#_Toc451266863)

[Abstract(EN) II](#_Toc451266864)

[Chapter 1 Introduction 1](#_Toc451266865)

[1.1 Background and significance 1](#_Toc451266866)

[1.2 Challenges of imbalance classification 2](#_Toc451266867)

[1.3 Research Status At Home and Abroad 3](#_Toc451266868)

[1.4 Structure of This Paper 4](#_Toc451266869)

[Chapter 2 A Review For Imbalanced Classification Problem 6](#_Toc451266870)

[2.1 Problems related to data instrinsic characteristics in imbalanced classification 6](#_Toc451266871)

[2.1.1 Small disjuncts 6](#_Toc451266872)

[2.1.2 Lack of density 8](#_Toc451266873)

[2.1.3 Class overlapping 10](#_Toc451266874)

[2.1.4 Noisy data 12](#_Toc451266875)

[2.1.5 Dataset shift 14](#_Toc451266876)

[2.2 Main Methods To Solve Imbalanced Classification Problem 15](#_Toc451266877)

[2.2.1 Sampling methods 15](#_Toc451266878)

[2.2.2 Cost-sensitive methods 18](#_Toc451266879)

[2.2.3 Ensemble methods 20](#_Toc451266880)

[2.3 Summary 21](#_Toc451266881)

[Chapter 3 Assessment metrics for imbalanced learning 22](#_Toc451266882)

[3.1 Singular assessment metrics 22](#_Toc451266883)

[3.1.1 Accuracy and ErrorRate 22](#_Toc451266884)

[3.1.2 Precision and Recall 23](#_Toc451266885)

[3.1.3 F-measure 23](#_Toc451266886)

[3.1.4 Sensitivity and Specificity 24](#_Toc451266887)

[3.1.5 Geometic mean 25](#_Toc451266888)

[3.2 ROC curves and AUC value 25](#_Toc451266889)

[3.3 Precision-Recall curves 27](#_Toc451266890)

[3.4 Cost curves 28](#_Toc451266891)

[3.5 Assessment metrics for multiclass imbalanced learning 30](#_Toc451266892)

[3.6 Summary 31](#_Toc451266893)

[Chaptrt 4 Best Classification threshold 32](#_Toc451266894)

[4.1 Shortage of ROC 32](#_Toc451266895)

[4.2 Finding best classification threshold 34](#_Toc451266896)

[4.2.1 Best classification threshold for training sets 35](#_Toc451266897)

[4.2.2 Different misclassification costs 40](#_Toc451266898)

[4.2.3 Applicatin of best classification threshold 42](#_Toc451266899)

[4.3 Summary 44](#_Toc451266900)

[Chapter 5 Protein remote homology detection 45](#_Toc451266901)

[5.1 Data sets 45](#_Toc451266902)

[5.2 Feature extraction methods 47](#_Toc451266903)

[5.3 Experimental methodology 49](#_Toc451266904)

[5.3.1 Unify threshold 49](#_Toc451266905)

[5.3.2 Application way 49](#_Toc451266906)

[5.3.3 Regulation factor 50](#_Toc451266907)

[5.3.4 Upper bound of optimal classification result 51](#_Toc451266908)

[5.3.5 AUC value and F1 value 52](#_Toc451266909)

[5.4 Summary 53](#_Toc451266910)

[Chapter 6 Conclusion and feature work 54](#_Toc451266911)

[6.1 Conclusion 54](#_Toc451266912)

[6.2 Feature work 54](#_Toc451266913)

[References 56](#_Toc451266914)

[Publications 62](#_Toc451266915)

[Acknowledgement 63](#_Toc451266916)

# 绪 论

## 课题背景及意义

在机器学习和数据挖掘领域中，分类是其中一个热门的研究方向。它属于有监督的学习方式，算法首先从已经标好类别的训练数据集中学习分类函数，建立分类模型。然后对未知的测试样本进行预测，力求预测错误率尽可能的小。已经有很多分类算法被提出：例如感知机、逻辑斯谛回归、朴素贝叶斯、决策树、K近邻分类器、支持向量机等等。这些算法已经被运用到数据挖掘的很多领域中，并且都取得很好的分类效果。但是它们大都是建立在数据集分布均衡的前提下的，即数据集中各类样本的数目大体一致。然而这一前提在很多现实问题中是不成立的。

在许多监督学习的应用中，不同类别的数据先验概率差别很大，比如分类问题中，一个样本属于不同类别的概率可能会非常不同，这就是我们常说的类别不平衡问题。这个问题在现实世界中经常出现，譬如在信用卡欺诈检测问题中[[1](#_ENREF_1)]，一般来说，大多数客户的交易行为都是诚信的，只有很少的一部分客户存在欺诈行为，可能10万笔交易中才存在1笔欺诈行为；另外在电信业、生态学、生物学、网络入侵、医疗诊断等其他领域中也存在类别不平衡现象，这已经成为现今数据挖掘领域中的一个突出问题。

我们称样本数多的类为多数类，样本数少的类为少数类。需要特别指出的是，从学习观点来看，少数类往往是人们最关注的，而且错分少数类样本往往会造成巨大的损失。比如在医疗诊断中，我们更关心的是诊断为有疾病的情况，并且如果把患病病人误诊为正常人，耽误了最佳治疗时间，后果将非常严重。所以对不平衡数据的正确分类，尤其是提高少数类的分类性能，已成为机器学习和数据挖掘领域里亟待解决的课题。对不平衡数据分类问题的研究具有很重要的现实意义和应用价值，针对这一问题的研究成果可以大大推动机器学习方法在实际工程中应用的步伐，扩大应用范围，真正为工业生产和社会经济带来较大的经济效益。

## 不平衡分类问题的挑战

其实在分类问题中，很多数据集都存在类别分布不相等的现象。但一般说的不平衡分类问题是指：类别分布明显不均衡，有时是差别巨大的分类问题。具体地，在不平衡分类问题中，多数类与少数类的样本比可能达到100：1，1000：1甚至更大的10000：1。每种情形下，都有某一类的样本数会远大于其他类别的样本数[[2](#_ENREF_2)]的现象。最多数的情况下，不平衡问题是二分类问题。多分类问题也有，而且因为可能存在多个少数类，使得这类问题变得更加复杂[[3](#_ENREF_3), [4](#_ENREF_4)]。

因为大多数标准的学习算法考虑的是一个平衡数据集，当这些算法用到不平衡数据时，可能会产生一个局部优化的分类模型。比如覆盖大多数的多数类样本，而错分大部分的少数类样本。所以这些在平衡数据框架下能取得很好表现的算法，在处理不平衡数据时，并不一定能取得最好的性能。造成这个问题的原因有以下3点[[5](#_ENREF_5)]：

1. 用以指导学习过程的，评估全局性能的指标，比如正确率，会偏向多数类。仍以上述信用卡欺诈为例，假设把所有欺诈用户也认定为合法用户，那么分类器的正确率可以达到99%以上，但这显然是没有意义的。
2. 预测样本为少数类的分类规则可能非常的特殊，其覆盖率很低，所以为了支持更一般化的规则(比如那些预测样本为多数类的规则)，这些特殊规则可能被遗弃。
3. 少数类中，规模非常小的类别簇可能会被错认为噪音数据，从而被分类器遗弃。而另一方面，因为用来训练少数类的样本数很少，少数几个噪音数据就可能会降低少数类的识别。

由于传统的分类算法会有很大的偏向性，导致其在少数类上的分类效果很差。但正如之前强调的，在实际问题中，往往需要重点检测的是少数类。因此，如何在不平衡数据的分类问题中，修正这种偏向性，提高分类器在少数类的识别，已成为机器学习和数据挖掘领域的一大挑战。

## 国内外研究现状

自21世纪开始，国内外很多学者不断地对这个课题进行深入探讨。在影响分类器性能的因素研究方面、处理不平衡数据集分类的策略方面以及用来评估不平衡分类的性能指标等都取得了很大进展。很多学者根据影响分类器性能的因素对算法进行了有针对性的改进，还有一些学者通过有机集成各种方法，多学科互相渗透，协同解决实际应用中的复杂问题，在各领域进行了有针对性的应用。

在研究不平衡数据集对分类器性能的影响因素方面，Japkowicz N和Stephen S对不平衡数据问题进行了系统研究，分别从概念复杂度、训练样本规模、类间不平衡程度等方面进行实验研究，试图找出数据对分类器性能影响的原因。研究发现当概念复杂度较低时，类间不平衡程度对分类器性能产生的影响不大，提高训练样本规模可以减少不平衡数据对分类器性能的影响，不同的不平衡数据分布特征对不同的分类方法影响程度有差异[[6](#_ENREF_6)]。Batista G E和 Prati R C等人通过多组对比实验，研究类间和类内不平衡性在影响分类器性能方面的程度差异，发现分类器性能下降有时不完全是不平衡数据引起，还可能是受多类之间的样本覆盖的影响[[7](#_ENREF_7)]。Zhou Z H和Zhang M L通过实验证明了训练集不平衡程度、学习任务的复杂程度、训练集规模和分类算法是影响分类器性能的主要因素[[8](#_ENREF_8)]。

针对训练样本的类别倾斜对传统分类器的性能影响，国内外学者提出了重构数据集的思想，进而提出了多种改进的采样算法。从基本的形式来分，有过采样算法和降采样算法，如Chawla N V和Bowyer K W提出了虚拟少数类过采样算法 SMOTE(synthetic minority over-sampling technique)，通过生成人工样本来对小类样本向上采样[[9](#_ENREF_9)]。而Liu X Y、Wu J X和Zhou Z H提出了两种有效的启发式降采样方法:EasyEnsemble和BalanceCascade[[10](#_ENREF_10)]。通过多次对多数类样本采样，学习构建多个学习器，并在最后级联所有学习器，完成最终的分类任务。在尽量保留原数据集信息的前提下，对多数类样本进行降采样。

此外，还有学者研究对传统分类器进行改进，从而使其能应用到不平衡数据集。根据现有分类算法如代价敏感学习、支持向量机算法、决策树算法、神经网络算法在处理不平衡分类时的不足，对算法进行改进，通过调节各个类别之间的代价函数、引入惩罚因子等有利于小类的措施，或者算法间取长补短进行集成，进一步提高了分类算法在不平衡数据分类时的性能。

分类器评估指标指导着学习器的建模方向，所以指标是否合理对分类器的性能有直接影响。不平衡数据分类因为特有的类别倾斜特点，使得常规的一些分类器评估指标不能有效地指导学习器的建模过程。Spackman将ROC（Receiver Operating Characteristic）[[11](#_ENREF_11)]曲线引入数据挖掘领域，提出将其作为评估指标。ROC曲线相比于常规的指标，在衡量不平衡数据分类时，有更好的表现。另外还有类似其他曲线图，如Davis Jesse等人提出Presicion-Recall(PR)曲线[[12](#_ENREF_12)]以及Drummond Chris等人提出的代价曲线[[13](#_ENREF_13)]，这些曲线都是针对ROC曲线在数据集严重不平衡时，会出现过高评价分类器性能的问题而提出的。

目前已有很多不平衡分类算法应用于实际问题中。1998年Kubat M和 Holte R C等人就对卫星图像进行分析，应用改进的分类方法对石油喷井进行自动监测[[14](#_ENREF_14)]。2003年Cohen G和Hilario M 等人分析了药物治疗检测中不平衡数据的分类问题，并将改进分类方法应用于生物信息学方面[[15](#_ENREF_15)]。2004年Phua C和Alahakoon D等人在金融欺诈的模式识别中讨论了不平衡数据分类问题[[16](#_ENREF_16)]，Zheng Z和Wu X等人分析了文本分类的特点，并提出针对不平衡数据分类的新方法[[17](#_ENREF_17)]。2005年Peng Y和Huang Q等人应用不平衡分类算法对基因数据进行分析来对乳腺癌进行诊断[[18](#_ENREF_18)]。2012年边婧和彭新光等人提出了新的大规模数据分层预处理LDSP算法，用于处理入侵检测大规模不平衡数据集分类问题[[19](#_ENREF_19)]这些应用均取得了不错的效果。

## 本文组织结构

尽管类别不平衡学习已经取得了很多研究成果，但仍有很多值得研究的问题。本文首先回顾了不平衡数据分类问题，介绍了该问题产生的原因，解决该问题面临的挑战，并总结了国内外对该问题的研究现状。为了对不平衡分类问题有一个更好的理解，我们还列举了一些与数据复杂度相关的数据内在特征，分析了这些内在特征与不平衡数据之间的联系。在总结前人工作的基础上，梳理了以往解决不平衡数据分类问题的方法。

介绍了适用不平衡分类问题的性能评估指标，重点分析了ROC曲线。最后也是最重要的，我们提出了一个提高分类器性能的框架，通过寻找最优的分类阈值，使分类器在正例的准确率和召回率上取得最佳性能。论文的各章节安排如下：

第一章：简要说明了本文的研究背景和现实意义，引入了不平衡分类问题。分析不平衡分类面临的挑战，总结了国内外学者对不平衡数据分类的研究现状。

第二章：对不平衡数据分类问题的相关进行综述。首先深入地分析与数据复杂度相关的5个数据内在特征。研究这些数据特征在不平衡分类中所造成的困境，并相应地给出了每种数据内在特征对应的解决方法和建议。其次概述了目前学术界提出的用以解决不平衡分类问题的策略，这些策略从数据层面或算法层面，探索解决分类器在不平衡数据分类时，偏好多数类的问题。

第三章：介绍评估不平衡分类性能的指标，指出每种指标的适用情形。重点地，介绍了ROC曲线和其对应的AUC值，分析在评估不平衡分类问题中的优势和缺陷。

第四章：针对ROC曲线的不足之处，提出了一套寻找最优分类阈值的框架。大多数学习器的输出结果是每个样本的一个类别分布概率，ROC曲线衡量的是学习器对样本的一个排序能力。在排序相对较好的情况下，还需确定一个合适的划分阈值，将样本分类。本文重点关注如何设定划分阈值，使得分类结果在其他指标也能取得好的表现。

第五章：将本文提出的框架运用到现实问题中，在生物信息学的蛋白质远程同源检测问题上做了大量实验，结果表明了我们提出的框架的有效性。

第六章：总结了本文所做的工作，对不平衡多分类问题的最优分类阈值寻找做了展望，介绍了未来可能的研究方向。

# 不平衡数据分类的相关研究综述

## 数据内在特征与分布不平衡问题

在研究中，有实验证明，简单的类别分布偏斜本身并不会影响学习任务[[2](#_ENREF_2), [5](#_ENREF_5), [20](#_ENREF_20)]。一个感性的认识是，如果不平衡数据是线性可分，则传统分类器，在该数据集上的性能不会有太大的降低。Victoria等人设计了一组对比实验[[5](#_ENREF_5)]，从同一份数据里抽取不同正负比的子集。考察C4.5在这些不同正负比数据集上的表现。结果显示正负比和分类器的性能之间并没有存在明显的关系。无论是正负比很高的数据，或是正负分布大致均衡的数据，C4.5都有表现不好的情况。

真正造成传统分类器在不平衡数据上性能下降的原因，是和数据的复杂度紧密相关的。在这节中，我们将讨论数据复杂度问题。具体地，我们将强调几个和复杂度相关的数据内在特征，这些特征会深刻影响不平衡分类结果。

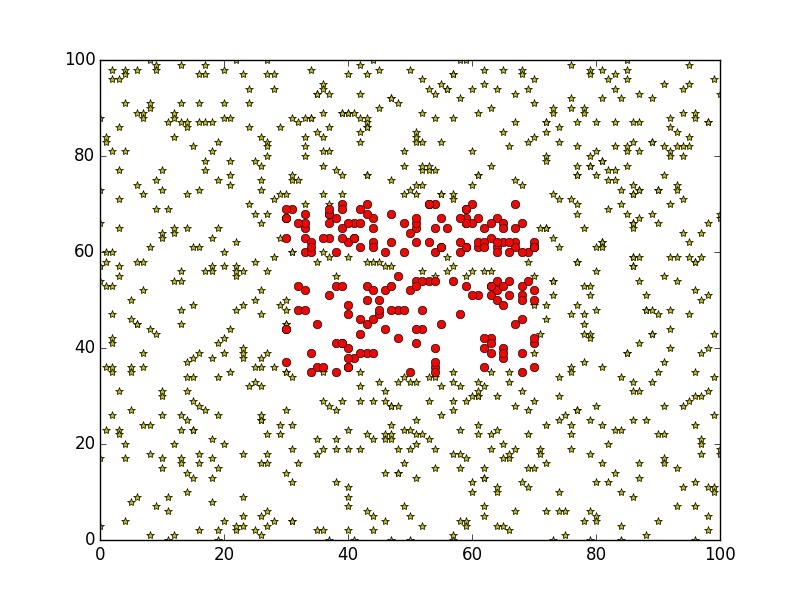
我们将枚举与每个特征相关的分类场景，分析内在特征与分布不平衡之间的联系。需要说明的是，部分内在特征之间会有一些相同的性质，而且一个通常的情况是，给定一个数据集，可能会发现有几个子问题同时出现。不过为了对这个主题有一个全局的介绍，我们只关注影响最大的那个子问题。此外，我们采用了[[21](#_ENREF_21)]中的不平衡数据集来做分析。

首先，我们讨论在不平衡数据集中的小析取项问题，然后分析训练集密度不足问题。其次，我们关注类别覆盖问题，揭示其在不平衡领域里的重要意义。然后讨论噪音数据，分析它是如何影响数据预处理技术和分类算法的性能。最后我们定义不平衡分类下的数据偏移问题。

### 小析取项问题

通常说的类别不平衡问题是指类间（between-class）不平衡，而小析取项是一种类内（within-class）不平衡问题[[22](#_ENREF_22)]。从概念学习的角度来说，当一个概念是由一系列子概念组成，且其中有一些子概念未被充分表达时，就会出现小析取项问题，它反映了同一类的若干子概念之间学习样本分布的不平衡性。

尽管大多数问题中都隐含有这些小析取项，但是在类别不平衡情形下，这些小析取项的存在极大地增加了问题的复杂性。因为在该情形下，很难知道某些样本是代表一个实际的子概念还是仅仅只是噪音数据[[22](#_ENREF_22)]。为了有个直观的感受，我们构造了一个人工数据集来说明问题。在图 2. 1中，圆圈代表正样本，五角星代表负样本，容易看出正样本是由两个子集构成的。另外需要注意的是，夹在正样本之间的几个负样本，也是一个小析取项。意即，在多数类中也是可能存在小析取项。



**图 2. 1：模拟小析取问题人工数据集**

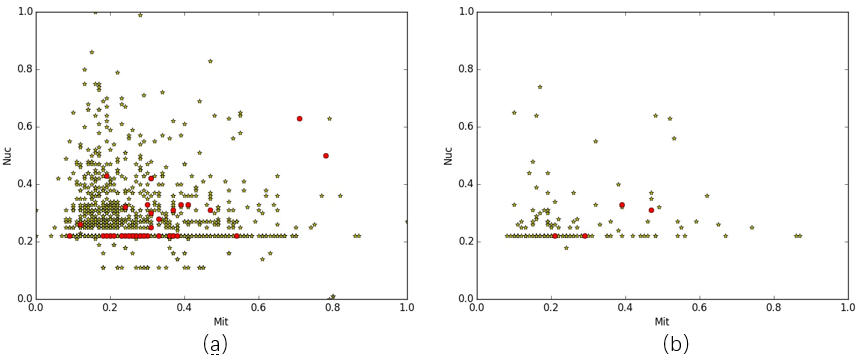
对于那些基于分治策略（divide-and-conquer）的算法来说，小析取项问题将变得更加严重。这类算法都会将原问题分解成子问题，像决策树的生成过程，不断地将原数据集划分成几个子集，而最终每个子集都只有很少的一些代表性实例。如果数据集的IR很高，那么明显的，这种不利因素将更加严重。Weiss 深入地研究了这个问题[[23](#_ENREF_23)]，并列举了几个处理小析取问题的技术：

1. 获取额外的训练数据。数据集不足常常会引起小析取项问题，特别是对于少数类而言。在[[24](#_ENREF_24)]中通过过采样技术，平衡不同类别的样本数目。且保证每个类别内，不同子集的样本数也是相同的。通过这样采样后，分类效果有较好的提升。
2. 使用一个更合适的归纳偏好。如果我们的目标是正确地检测小析取项区域，则必须运用一些复杂的机制，来避免算法偏好大的数据子集。例如，采用代价敏感学习算法，将错分代价结合进分类算法，修正算法对多数类的偏好。
3. 使用更合适的指标。这和前一个提到的技术是相关的。对不平衡数据来说，在数据挖掘过程中建议使用特定的评估指标，使得小析取项中的少数类，在分类模型的构建过程中，能得到正的加权[[25](#_ENREF_25)]。比如，分别使用多数类和少数类的查准率和查全率，可以产生对少数类来说更精确的规则。
4. 避免剪枝。为了对学习得到的规则进行泛化，剪枝倾向于消除大多数的小析取项，所以这类方法是不推荐的。
5. 应用提升方法。提升算法，像AdaBoost算法，是一类迭代算法，在每轮迭代中对训练集的分布设置不同的权重因子，针对被错分的样本，其权重将被增大，而被正确分类的样本，其权重将减小[[26](#_ENREF_26)]。因为在小析取项里的实例是难于识别的，我们有理由相信，提升算法将会提高分类这些实例的性能。在这种想法下，有很多算法被提出，通过修改标准的提升算法的权重更新机制，来提高分类少数类和小析取项的性能[[27](#_ENREF_27), [28](#_ENREF_28)]。

最后，我们必须强调CBO方法的使用[[29](#_ENREF_29)]。CBO是一个重采样策略，会同时消除类间和类内的不平衡。具体的，该方法在第一步使用K-means算法来检测正负类的子簇。在第二步，算法随机的复制每个子簇的样本(除了最大的负类子簇)，来获得一个在类间和类内都平衡的子簇分布。这些子簇可以看成数据集里的小析取项，所以这个预处理机制的目的在于强调这些区域的重要性。

### 密度不足

在分类中可能出现的一个问题是样本规模过小[[30](#_ENREF_30)]，从而导致“密度不足”或者“信息缺乏”等相关问题。这个问题会导致算法没有足够多的数据来概括样本的分布，并且在高维或者数据不平衡的情况下，该问题会变得更加突出。我们选了yeast5数据来做一个展示，以‘Mit’和‘Nuc’两个特征作为X轴和Y轴绘制了两张散点图，如图 2. 2所示。其中图（a）是使用全部的数据绘制的，图（b）是只用其中的10%数据来绘制。可以直观的理解，当没有足够多的数据来描述问题的边界时，学习算法是难以获得一个具有很好的泛化能力的模型的。最重要的是，当少数类样本的密度太低时，它们可能会被简单地看成噪音数据。



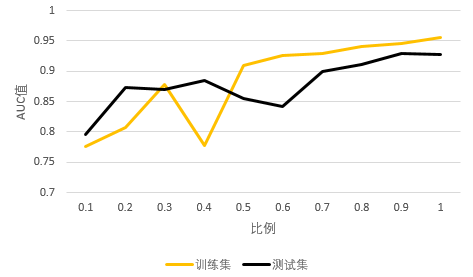
**图 2. 2：yeast5的样本分布图**

不平衡数据和小规模样本问题的结合，给机器学习和数据挖掘领域造成了新的困境。因为在这种场景中，少数类不能很好地被代表，用来学习少数类的知识模型就会变得太过具体，从而导致过拟合。进一步地，正如前面章节所阐述的，训练数据的密度不足可能会引入小析取项问题。所以两个具有同样IR的数据集不能简单地就认为它们代表相同的复杂性，因为训练集中少数类样本是如何代表也是很重要，例如小析取项的存在与否。

在[[31](#_ENREF_31)]中，作者研究了类别分布和训练集规模对分类器性能的影响，并以C4.5作为基本的学习算法。他们分析了不同规模的训练数据集和不同IR的不平衡数据集，并观察在这些情形下的AUC值。

他们的第一个结论是：训练集的数量越多，性能的结果越好，和类别的分布无关。第二个重要事实是不同训练集规模下，产生最好性能的IR有时是不同的，这支持了对于一个学习任务可能存在一个“最好”的类别先验分布的观点。一个有益的建议是：可以通过采样，尝试不同IR比的数据集，找到一个能取得最佳性能的不平衡比。

为了说明样本密度对学习性能的影响，在图 2. 3，我们展示了C4.5在VOWE10的训练集和测试集上的AUC值。训练集的数量从原数据集的10%到整个原数据集。训练集是采用5折交叉验证，而测试集数量不变，图中展示的结果5折交叉验证的平均值。从图可以看出，分类性能基本是跟训练集的实例数呈正相关的，这也反映了先前[[31](#_ENREF_31)]中列举的一些发现。



**图 2. 3：不同样本密度下C4.5分类器的性能图**

### 类别覆盖

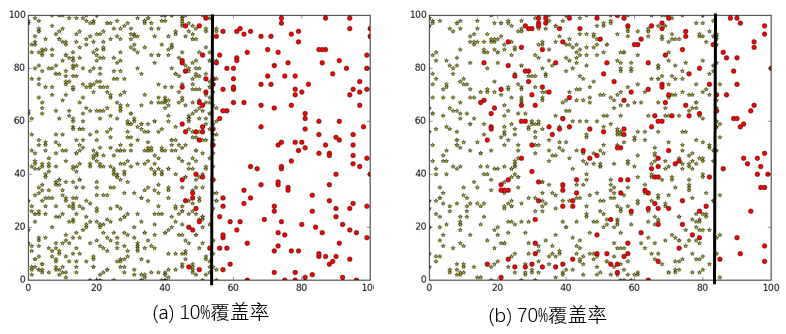
当在数据空间的某个区域中，来自每个类别的训练数据都差不多时，就会出现类别覆盖问题。这种情形会导致在覆盖区域中，每个类别的数据都有几乎相同的先验概率，使得区分该区域数据的类别变得非常困难甚至是不可能。事实上，只要数据集不发生类别覆盖，比如任一线性可分问题，都可以由任一简单的分类器解决，而不管类别分布是怎样的。

有许多工作致力于研究类别覆盖与类别不平衡问题之间的关系，特别的，在[[29](#_ENREF_29)]中，作者采用了几个具有不同IR的，不同类别覆盖程度的人工数据做实验。他们得出结论：类别的先验概率并不是影响分类性能的主要障碍，造成主要障碍的是类别的覆盖程度。

我们创建了一个具有1000个样本，IR率为4的人工数据，即每5个样本中只有一个正例。然后，我们让个别特征值产生不同程度的覆盖，从没有覆盖到100%覆盖。我们使用C4.5分类器来研究在固定IR下类别覆盖对分类的影响。表 2. 展示了实验结果，从中可以观察到随着覆盖程度的增加，性能也急剧变差。另外，图 2. 4也展示了这个问题，从图可以看出当发生类别覆盖时，决策树不仅对两个类别都不能正确的区分，而且是偏向多数类的，导致AUC值很低。

**表 2. 1：不同覆盖率下C4.5的分类性能**

|  |  |  |  |
| --- | --- | --- | --- |
| 覆盖率(%) | TPRate | TNRate | AUC |
| 0 | 0.985 | 1 | 0.988 |
| 10 | 0.805 | 1 | 0.891 |
| 20 | 0.685 | 1 | 0.821 |
| 30 | 0.57 | 1 | 0.757 |
| 40 | 0.46 | 0.989 | 0.782 |
| 50 | 0.28 | 1 | 0.621 |
| 60 | 0.22 | 1 | 0.602 |
| 70 | 0.175 | 1 | 0.578 |
| 80 | 0.165 | 1 | 0.574 |
| 90 | 0.015 | 1 | 0.499 |
| 100 | 0 | 1 | 0.5 |



**图 2. 4：10%覆盖率和70%覆盖率样本分布图**

此外，在[[32](#_ENREF_32)]中也做了类似的工作，他们深入的研究了KNN算法，在不同平衡比、不同覆盖区域时的表现。作者设计了两种不同场景，来研究不平衡比和类别覆盖之间的联系:在第一种，发生类别覆盖的区域的IR和整个数据集的IR相似；而在另一种，类别覆盖区域的IR和整个数据集的IR相反，即在类别覆盖区域，少数类的样本密度要大于多数的样本密度。他们证明：当类别覆盖区的数据是不平衡时，相比于类别覆盖的比例，覆盖区的IR对分类器的性能影响更大；其次，局部的IR对分类器性能的影响要大于全局的IR；最后，使用更全局性的学习过程(如K=9,13)的分类器会取得较好的TP率，而使用更局部性学习过程(如k=1)的模型先比前者会获得更好的TN率。

另外还有作者研究类别覆盖和类别不平衡对模型复杂性的影响，并证明相比于类别不平衡，类别覆盖对模型复杂度的影响更严重[[33](#_ENREF_33)]。而且如果这两个因素一起出现，则造成的困难要比任一单独的因素造成的问题严重得多。为了证明，他们也同样使用了人工数据，并且用SVM来分类。他们测试了不同的不平衡率，类别的覆盖程度，以及不平衡和类别覆盖共同出现的情况。结果表明，当训练集规模小时，高不平衡率会造成性能的急剧下降，这可以归因于小析取项的出现。但当训练集规模达到一定大小时，不平衡问题对SVM性能的影响很小，几乎可以不计。而类别覆盖造成的性能损失较一致，和训练集的大小无关。然而当两者共同出现时，分类的性能将显著地退化。另外，作者还观察了不平衡和类别覆盖对模型复杂度的影响，复杂度以SVM的支持向量的个数来度量。发现不平衡问题对模型复杂度几乎没影响，随着不平衡比的扩大，支持向量的个数甚至会有些微的减少。而类别覆盖率与模型复杂度呈近似线性的关系，随着覆盖率的提升，向量个数也跟着增多。

最后，在[[34](#_ENREF_34)]中，作者提出一个结合了数据预处理和特征选择的框架。在该框架下，数据预处理步骤可以解决类别分布和小析取项的问题，而特征选择在一定程度上减小了重叠程度。从更一般的角度来说，这种框架背后的思想在于：用不同的方法去解决数据的不同复杂性问题，包括类别重叠，特征不相关或特征冗余，噪音数据，类别不平衡，数据规模相对特征维度比率过小等等。

### 噪音数据

众所周知，噪音数据对任一数据挖掘系统的性能都会造成影响[[35](#_ENREF_35)]。而在不平衡数据场景下，噪音数据的存在对少数类的影响要比在正常情况下深远得多[[36](#_ENREF_36)]。因为初始的正例样本较少，只要很少的“噪音”数据就会影响子概念的学习。

根据[[36](#_ENREF_36)],“噪音区域”在一定程度上可以被看成“小析取项”，为了避免错误地对这些样本生成判别函数，必须运用一些过拟合控制技术，如剪枝技术。然而，运用这些技术后，一些正确的少数类也可能被忽略。所以在这种方式下，为了对问题的所有类都能有一个好的表现，应该注意调整学习器的偏差。

例如，Batuwita和Palade设计了FSVM-CIL算法[[37](#_ENREF_37)]，融合了支持向量机和模糊逻辑，意在反映不同训练样本的类内重要性，抑制异常点和噪音的影响。主要思想是对正负样本分别赋予不同的模糊隶属度值，并将这个信息结合进SVM学习算法中，从而减少异常点和噪音的影响。

有作者研究了类别不平衡以及噪音对不同分类算法和数据采样技术的影响[[38](#_ENREF_38)]。在这个研究中，作者提出了3个重要的经验总结：

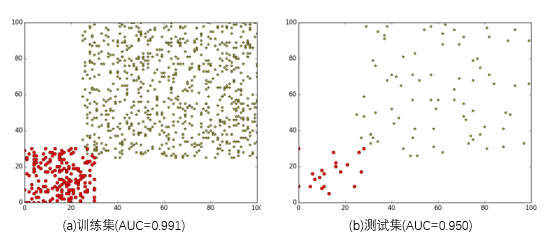
1. 相比于类别不平衡，分类算法对噪音数据更敏感。然而，当不平衡率急剧增大时，它对分类算法的性能以及采样技术的影响更大。
2. 关于数据预处理机制，简单的降采样，比如随机降采样，在所有程度的噪音数据和类别不平衡下，整体上表现最佳。特别的，在有噪音数据存在的情况下，ENN对不平衡程度更具鲁棒性。另外OSS对噪音数据的增加相对不受影响。其他技术，像随机过采样，SMOTE或者Borderline-SMOTE在平均上能取得不错的结果，但没有降采样一样的表现。
3. 最后，在不平衡和噪音数据上的测试算法中，最具鲁棒性的算法是贝叶斯分类器和支持向量机。平均性能要优于规则归纳算法和基于实例的算法。再者，虽然大多数算法在不平衡率增加时，AUC值只有很小的变化。Radial Basis Functions的性能在不平衡率增加时，将严重降低。相比于其他算法，规则学习算法的性能更容易受噪音数据的影响。

另外，在[[39](#_ENREF_39)]中作者提出了一个相似的研究，探索噪音数据和不平衡在Bagging和Boosting技术中的影响。他们的结果显示:一般而言，在处理包含噪音的不平衡数据时，Bagging技术会优于Boosting技术；无放回的Bagging技术性能要优于有放回的方式；最后在采用Boosting技术之前，可以先运用降噪技术处理数据，从而获得更好的效果。

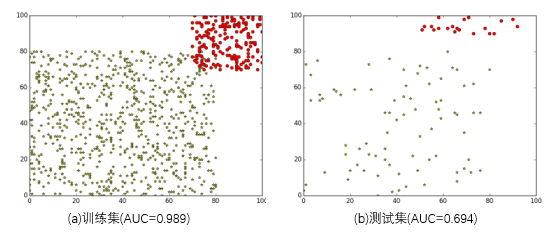
### 数据偏移

当训练集和测试集服从不同的分布时，我们就称数据集发生数据偏移[[40](#_ENREF_40), [41](#_ENREF_41)]，该问题会影响所有类型的分类器的性能。并且因为采样选择偏差的原因，这个问题经常出现。轻度的数据偏移在大多数的现实问题都有出现过，不过一般的分类器都可以在不严重损失性能的前提下解决这个问题。

然而，在处理不平衡分类问题时，数据偏移问题变得特别重要。因为在高度不平衡域中，少数类因为样本特别少，对分类错误特别敏感。



**图 2. 5：类别分布一致数据图**



**图 2. 6 类别分布迁移数据图**

为了有个清晰的理解，我们构造了两个人工数据集，来展示了两个数据偏移影响不平衡分类的例子。在图 2. 5，测试集和训练集来自同一个分布，很容易地看出在训练集的划分很完美地移到了测试集。然而，在图 2. 6，测试集和训练集的分布则不一样。在测试集中，少数类的分布往左上方偏移了，从而导致训练集和测试集在性能上的差异。

因为数据偏移问题和不平衡分类是高度相关的，所以在未来的研究中，关注数据偏移将是一个有趣的切入点。研究在不平衡域里的数据偏移有两种可能的方向:

第一种是关注数据偏移的导致的学习难点，即感兴趣的数据存在一定程度的偏移，导致性能下降时，该如何处理。对于这种情况，一般可以设计技术，来发现数据偏移，衡量偏移程度，特别地要关注少数类样本。再者，我们可以设计能在数据偏移条件下工作的算法，比如可以通过预处理技术或设计特别的算法。

第二种是研究引起数据偏移的原因。目前验证模型的最常采用的技术是分层的交差验证技术，而该技术在学习过程中可能会引入数据偏移问题。所以需要设计一个更合适的验证技术，来避免人工的引入数据偏移问题。

## 解决不平衡数据分类的策略

在这一节，我们将更具体地介绍各种用来处理不平衡数据分类的方法。这些策略被划分为两大类：算法层面方法，包括创建新的分类算法或者修改现有的算法使其能处理类别不平衡问题；数据层面方法，对数据进行预处理，消除类别不平衡的影响。进一步的，代价敏感学习，将数据层面的方法和算法层面的方法都结合了进来。它对小类别样本赋予更大的错分惩罚项，然后尽力的最小化整体的惩罚代价。集成方法在不平衡领域中也经常被采用，这类方法要么在数据层面修改集成学习算法，在使用每个分类器学习之前先对数据预处理；要么在集成学习过程中嵌入一个代价敏感的框架。

### 重采样技术

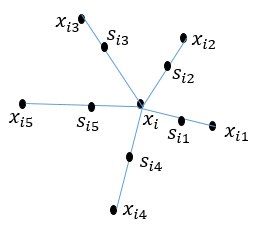
在重采样技术中，一般是采用一些机制，改变原有的类别分布，使不平衡数据集变成平衡数据集。有工作实验证明了:部分基础分类器，在改变后的平衡数据集上，能取得比之前的不平衡数据集更好的性能[[42-44](#_ENREF_42)]。而且，这些技术的主要优点是它们是和底层的分类器无关的。

在所有的采样方法中，最简单的预处理技术是非启发式方法，比如随机降采样和过采样。在随机过采样中，根据需要随机地复制少数类里的部分样本，加到原数据集中。随机降采样会从多数类里随机地挑选部分样本，将这些样本删掉，从而改变数据的不平衡比。虽然随机降/过采样方法简单易实现，但两者都各自引入了不同的学习障碍。这些障碍都可能会造成分类器的性能损失[[45-47](#_ENREF_45)]。对于随机降采样，主要问题是它可能抛弃了潜在的对学习过程来说很重要的有用数据。而对于随机过采样，因为它只是对现有实例的精确复制，所以可能导致过拟合问题[[47](#_ENREF_47)]。

为了处理上面提到的问题，一些更复杂的方法相继被提出。在[[10](#_ENREF_10)]中，作者提出了两种有效的启发式降采样方法:EasyEnsemble和BalanceCascade，它们的目标在于克服随机降采样带来的信息丢失问题。EasyEnsemble是一种集成式的算法，首先从多数类里随机地抽取多个与和少数类一样大小的子集，然后每个子集分别和少数类组合，构成一个个新的训练数据集。用这些训练数据集分别学习得到多个分类器，最后预测时，就集成多个分类器的模型进行预测。而BalanceCascade在学习完第一个假设H(1)后，如果对于一个多数类样本，它能被H(1)正确识别，则认为这个样本存在一定程度的多冗余性。所以在后续处理中，会把这个样本删掉。然后再抽取多数类样本学习新的假设，最后集成所有的假设来做分类。

另一个启发式降采样算法，利用KNN(K-nearest neighbor)分类器来实现降采样过程。根据数据的分布特点，[[48](#_ENREF_48)]提出了4种KNN降采样方法:NearMiss-1、NearMiss-2、NearMiss-3和“most distant”方法。NearMiss-1，首先计算每个多数类样本与距离它最接近的三个少数类样本的平均距离。然后选择平均距离最小的多数类样本；NearMiss-2，则是计算每个多数类样本与距离它最远的三个少数类样本的平均距离，然后选平均距离最小的多数类样本。NearMiss-3，则是为每个少数类样本，选择距离其最近的几个多数类样本，以保证每个少数类样本都是被多数类样本所包围着。最后，“most distant”则是选择那些与最接近它的三个少数类样本的平均距离最远的多数类样本。实验证明NearMiss-2方法可以取得最佳的效果。

在启发式过采样中，SMOTE(Synthetic Minotity Oversampling TEchnique)是这个领域里最著名的一个[[9](#_ENREF_9)]。简单来讲，它的主要思想是，通过在小类别样本之间进行插值的方法产生新的小类别样本，从而实现对训练集的过采样。在该方法下，对于每个小类别样本，在其同类中查找k个最近邻样本，然后根据过采样数量的要求，随机从中选择N个样本，然后在这个小类别样本和被选择的每个邻近样本之间进行随机线性插值，构成新的人工样本。整个过程如图 2. 7所示，其中是选择的样本点，到是一些选择的近邻样本，而到是通过随机插值产生的人工数据点。



**图 2. 7：SMOTE算法示意图**

相比于随机过采样，SMOTE算法能取得较好的性能提升。但是，SMOTE也同样存在不足之处，包括过拟合和不一致性[[49](#_ENREF_49)]。SMOTE对每个原少数类样本都产生相同数目的人工样本，而没有考虑近邻样本，从而加重了不同类别样本的覆盖问题。有鉴于此，许多变形算法被提出来克服这种缺陷。一些代表性的工作包括：Boderline-SMOTE[[50](#_ENREF_50)]，ADASYN(Adaptive Synthetic Sampling)[[51](#_ENREF_51)]，Safe-Level-SMOTE[[52](#_ENREF_52)]以及SPIDER2[[53](#_ENREF_53)]算法。

在采样算法中，还有一大类方法是基于数据清理技术的。在该领域一些代表性工作包括：ENN(edited nearest neighbor)规则[[54](#_ENREF_54)]，如果样本的三个最近邻中有两个是跟自己不一样的，那么这个样本将被移除。Tomke Links[[55](#_ENREF_55)]定义如下:如果有分别来自正负类别的两个样本，它们之间的连线是最短的。意即，不存在第三个样本，使得这两个样本中的任意一个与第三个样本的距离小于它们两者之间的距离，则称两个样本之间的连线是Tomke Links。Tomke Links连线上的两个点要么有一个点是噪音数据，要么两个点都是边界样本。所以一般会移除掉Tomke Links连线上的点，使得类别边界更清晰，从而提高分类性能。还有的作者提出移除那些远离决策边界的多数类样本，认为这些多数类样本对分类的指导意义不大[[56](#_ENREF_56)]。最后，将数据预处理和数据清理技术相结合可以减少因采样而引入的类别重叠问题，比如SMOTE和ENN，或Tomek Links结合[[7](#_ENREF_7)]。

另外还有一些基于聚类的采样算法，这类算法都是先根据一些显著的特征把训练集分成几组，然后再运用降采样或过采样技术。一些重要的例子包括：CBO(Cluster-Based Oversampling)[[29](#_ENREF_29)]，Class Purity Maximization[[57](#_ENREF_57)]，聚类降采样结合SMOTE过采样算法[[58](#_ENREF_58)]以及基于DBSCAN聚类的DBSMOTE算法[[59](#_ENREF_59)]。

### 代价敏感学习

基于采样的方法，都是通过重新构建训练集，获得一个平衡分布的数据集，从而使得传统分类算法在重构的数据集上能正常运行。而代价敏感学习方法，不用改变数据集的原始分布，它是将不同类别样本的错分代价结合进学习过程中，从而获得能处理不平衡数据的学习模型。

代价敏感学习考虑给不同类别的样本赋予不同的错分代价，作为一个例子，表 2. 展示了一个代价矩阵，记录了将类别i的样本错分为类别j的惩罚因子。这里我们假定少数类是第0类，多数类是第1类。

**表 2. 2：二分类代价矩阵**

|  |  |  |
| --- | --- | --- |
| 真实类别 | 预测类别 | |
| 第0类 | 第1类 |
| 第0类 | 0 |  |
| 第1类 |  | 0 |

这些错分代价值可以由专家直接给出，也可以通过其他方法学习得到。特别的，在解决不平衡数据分类问题中，我们通常对少数类样本更感兴趣，因此，错分少数类的代价必须比错分多数类的代价要高，即。

给定代价矩阵，根据最小期望代价原则，一个实例应该被分到具有最小期望代价的类别中。把实例分到类别i的期望代价可以表示如下：

(公式 2. 1)

其中是将实例分到类别j的先验概率。也就是说，一个分类器会将一个实例划分为正例，当且仅当：

(公式 2. 2)

等价于：

(公式 2. 3)

所以，对任意给定的代价矩阵，如果假定，则分类器将一个实例划分为正例当且仅当：

(公式 2. 4)

又因为，我们可以得到一个阈值，当时，将实例分为正例，其中：

(公式 2. 5)

根据上面的式子，可以将代价敏感学习分成直接法和元学习法。

直接法：直接的代价敏感学习算法是通过直接引入错分代价到学习算法来实现的。比如在决策树归纳中，建树策略适合于最小化错分代价。代价信息用于：(1)选择最优的属性来划分数据[[60](#_ENREF_60)]；(2)决定某棵子树是否需要被剪枝[[61](#_ENREF_61)]。另一方面，在其他基于遗传算法的方法中，可以将错分代价结合进适应性函数中[[62](#_ENREF_62)]。

元学习：这类的方法学意味着，要么对训练集进行预处理，要么对输出结果进行后期加工处理。在这样的一个方式中，原始的学习算法可以不需要改变。代价敏感元学习算法可以进一步分为两大类：阈值和采样。

阈值法是基于基本的决策理论，即将实例分到具有最小期望代价的类别中。例如在一棵典型的二分类判定树中，将叶结点中实例数最大的类作为叶结点的类别。一个代价敏感算法赋予结点某一类别，使得在该类别下结点的错分代价最小[[63](#_ENREF_63)]。

采样法是基于对训练数据的修改，最流行的技术是根据代价矩阵，对原始训练集的类别分布进行重采样。可通过降/过采样[[64](#_ENREF_64)]，或者赋予实例不同的权重[[65](#_ENREF_65)]。这些修改被证明是有效的，并且可以被应用到任何的非代价敏感学习算法中。

### 集成方法

集成分类器，也称多分类器系统，尝试提高单个分类器的性能，通过建立多个分类器并组合这些分类器来最终得到一个性能优于任一基分类器的集成分类器。因此，基本想法是从原数据集中学习多个分类器，然后对于一个未知的实例，集合这些分类器对该实例的预测结果得到最终结果。

在最近几年来，集成分类器已经成为解决类别不平衡问题的一个可行方案[[10](#_ENREF_10), [66](#_ENREF_66)]。集成方法是基于集成学习算法和前面提到的一些相关技术的组合，这些技术包括数据层面或算法层面的方法，也可以是代价敏感学习方法。在组合数据层面方法和集成学习算法的情形中，新的混合方法通常在训练每个分类器之前都对数据进行预处理。另一方面，代价敏感的集成方法，在学习的过程中，不再是通过修改基分类器，来接受不同的惩罚代价，而是通过集成的学习算法来指导代价最小化过程。这样，可以避免对基分类器的修改，但是仍存在一个主要缺点，即代价成本的定义。

集成学习方法在不平衡数据的应用的分类可以见论文[[67](#_ENREF_67)]，作者将不平衡数据中的集成方法分为4大类。首先，他们划分了一类代价敏感提升方法，这类方法和代价敏感方法很相似，只是代价最小化过程是由一个提升算法来指导的。然后，划分了另外三类方法，这三类方法有一个相同的特征，都是将数据预处理过程和集成学习算法相结合，不同之处在于使用的集成算法不一样，分为bossting，bagging，hybrid ensembles。在[[67](#_ENREF_67)]的研究中，作者总结到集成算法是值得尝试的，通过对数据进行预处理以及训练单个分类器可以提高得到的结果。他们也强调了一些简单的方法具有不错的性能，比如：RUSBoot[[68](#_ENREF_68)]，或者UnderBagging[[69](#_ENREF_69)]。这些方法虽然简单，但相对一些复杂的算法，性能却更好。

## 本章小结

在这一章里，先介绍了5种数据的内在特征。这些特征的存在增加了数据的复杂度，而当这些特征与分布不平衡问题共同出现时，就会使得数据变得非常难以处理，从而极大地降低了传统分类器的性能。两者共同出现对分类器造成的损失，要大于两者在单独出现时所造成的损失的一个简单加和。先对应地，我们也介绍了每种内在特征的处理方法。合理地应用这些处理方法，可以减少数据复杂度，在一定程度上提升分类器的效果。一个合理的建议是，在处理不平衡数据时，可以先用特征筛选算法，选出最重要的两个特征，绘制数据的二维分布图。观察数据是否存在本章提到的内在特征，如果存在，可以采取对应的措施解决。

其次，介绍了用于处理不平衡数据分类的常用策略。从整体来看，这些策略可以分为三大类。数据层面的方法：通过对原始数据集进行采样，改变类别的分布，使数据集变成平衡数据集，从而可以使用传统的分类算法来学习。算法层面的方法：提供修改传统的分类算法，如将错分代价结合进学习过程，使修改后的算法能处理不平衡数据集。最后就是采用集成学习的策略，在数据层面和算法层面上来共同解决不平衡分类问题。

# 衡量不平衡分类性能的指标

分类器评估指标的合理性会影响着学习器的性能，这是因为学习器优化的目标就是使评估指标最大化，可以说衡量指标指导着分类器的建模过程。目前各个领域对分类器性能的评价标准不同，在此综述一下主流的分类器性能评价标准。

## 单一的评估指标

在一个二分类问题中，混淆矩阵(表1)记录着每个类被正确和错误分类的样本数。其中TP(True Positive)表示准确被预测为的正例个数，TN（True Negative）表示准确被预测为的反例个数，FP（False Positive）表示把反例预测为正例的个数，FN（False Negative）表示把正例预测为反例的个数。在本文中，如不做特别说明，则正类指的都是少数类，负类指的都是多数类。

**表 3. 1：二分类问题中的混淆矩阵**

|  |  |  |
| --- | --- | --- |
|  | 预测正类 | 预测负类 |
| 实际正类 | 正分正例(TP) | 错分反例(FN) |
| 实际负类 | 错分正例(FP) | 正分反例(TN) |

### 准确率(Accuracy)与错误率(ErrorRate)

在传统的应用中，最常用的指标是准确率和错误率，分别定义如下：

(公式 3. 1)

(公式 3. 2)

这类指标只是提供了一个对分类器性能的简单衡量，如之前所介绍的，在不平衡的分类问题中，它们容易造成将所有数据都分为负例的不理想结果。有很多代表性的工作，研究了准确率在不平衡分类场景下的缺陷[[36](#_ENREF_36), [70](#_ENREF_70)]。最基本的原因可解释如下：在表 3. 1的混淆矩阵中，第二行是所有的正类样本，第三行是所有的负类样本。即这两行的比例代表了数据集的类别分布，任何同时使用了这两行数据的指标，都不可避免的会受到数据不平衡问题的影响。在准确率的定义公式中，同时用到了正类的值和负类的值。因此当类别分布发生变化时，虽然分类器的实际性能没有改变，但是性能衡量指标却会发生变化。而这种衡量的不一致性，在比较不同算法在不同数据集上的性能时，就会造成问题。所以，在不平衡数据分类场景下，如果指标对数据的分布敏感，那么就难以做一个相对的比较。

### 查准率(Precision)和查全率(Recall)

查准率为在所有预测为正例的数据中，真实正例所占的比例，其定义如下：

(公式 3. 3)

它是对分类精确性的一个衡量，对数据的分布变化敏感。而查全率为在所有真实正例中被正确预测为正例所占的比例，其定义如下：

(公式 3. 4)

它是对分类完备性的一个衡量，对数据的分布变化不敏感。虽然查全率不依赖于数据分布，但单独使用查全率来衡量分类结果是不充分的，因为它没有提供错分正例的信息。同样地，查准率没有提供多少正例被错分的信息。

### F-measure

F-measure是Precision和Recall的加权调和评价，综合反映了Precision和Recall的分类性能，其定义如下：

(公式 3. 5)

其中P和R分别是查准率和查全率，度量了查全率对查准率的相对重要性。是最常用的F-Measure，也称F1值。时，查全率重要度更大；时查准率的重要度更大。F-measure在衡量分类器性能，能比准确率提供更多的信息，但是F-measure也存在对数据分布敏感的问题。

在一些场景中，可能会有很多个二分类混淆矩阵，比如进行多次的训练/测试，每次得到一个混淆矩阵；或是在多个数据集上进行训练/测试，希望估计算法的“全局”性能；又或是在多分类问题中，每个不同类别的组合对，都会产生一个混淆矩阵。在这些情景下，我们会希望在N个二分类混淆矩阵上综合考察查全率和查准率。

一种方法是先在每个混淆矩阵上分别计算出查准率和查全率，可记为，，…，，然后计算平均值，这样就得到“宏查全率（macro-R）”，“宏查准率(macro-P)”，以及对应的“宏F1(macro-F1)”：

(公式 3. 6)

(公式 3. 7)

(公式 3. 8)

另外可以将所有混淆矩阵的对应元素进行平均，得到TP，FP，TN，FN的平均值，分别记为，，，，然后再根据这些平均值计算“微查全率（micro-R）”，“微查准率（micro-P）”和“微F1(micro-F1)”:

(公式 3. 9)

(公式 3. 10)

(公式 3. 11)

### 敏感性(Sensitivity，SN)和特异性(Specificity，SP)

敏感性表示的是正例分类精度的高低，其定义如下：

(公式 3. 12)

根据公式，敏感性和查全率是一样的。特异性表示的是反例分类精度的高低，其定义如下:

(公式 3. 13)

因为敏感性和特异性都只涉及混淆矩阵中某一类的值，所以它们对数据的分布变化不敏感。

### 几何均数(Geometic mean，GM)

几何平均数表示的是正例精度和反例精度的几何平均，其定义如下：

(公式 3. 14)

GM兼顾了正反例的分类精度，能较好地评估学习器在不平衡数据分类问题上的表现。虽然F-measure和GM，在衡量不平衡数据分类性能时，都比准确率有更优的表现。但是在回答关于分类器性能的更一般性的问题:比如如何在一个样本分布范围内，比较两个分类器的性能？这两者也是力有未逮。

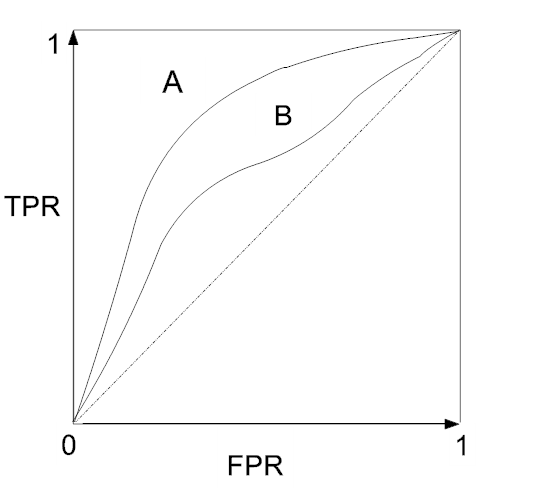
## ROC曲线与AUC值

如之前的章节所介绍的，在不平衡数据分类中，需要特别的指标来衡量分类器的性能。在众多的研究中，ROC（Receiver Operating Characteristic）曲线图是最常用的，用以衡量不平衡数据的分类性能的指标[[71](#_ENREF_71), [72](#_ENREF_72)]。

在二分类问题中，很多学习器的输出是一个实值或概率预测，如实例属于正类的概率。分类时可以划定一个分类阈值，把大于该阈值的分为正类，其他的分为负类。这个实值或概率预测结果的好坏，直接决定了学习器的泛化性能。一般地，我们会根据这个实值或概率预测值，将样本进行降序排序。然后设定一个划分阈值，将样本分为两部分，前一部分分为正类，后一部分分为负类。在不同的应用中，可以根据需要设定不同的划分阈值。假设我们更关心查全率，则可以将阈值设小一点；若更重视关心查准率，则可以将阈值设大一点。这个排序本身的质量好坏，体现了综合考虑学习器在不同任务下的“期望泛化性能”。或者说，“一般情况下”泛化性能的好坏，ROC曲线正是研究这种排序性能好坏的有利工具。

ROC全称是“受试着工作特征曲线”，它源于“二战”中用于敌机检测的雷达信号分析技术，二十世纪六七十年代开始被用于一些心理学、医学检查应用中，后面被引入机器学习领域[[11](#_ENREF_11)]。我们根据学习器的预测结果对样本进行排序，然后设定一个阈值进行划分。分完后可以得到一个TPR（敏感度）和TNR（特异度）。将分类阈值从逐渐递增到，就可以得到一系列的（TPR,TNR）点对。以FPR为横轴，TPR为纵轴，将这些点连起来就可以得到ROC曲线。当正负类样本在测试集上的比例发生变化时，ROC曲线不会发生改变，所以ROC曲线对类别分布不敏感，在评价不平衡数据集下分类器的性能具有优势。

ROC曲线具有良好的可视性，可以直观地体现分类器的好坏，比如图 3. 1中，曲线A在曲线B的左上方，完全包住了曲线B。则可以断言，曲线A对应的分类器的性能要优于曲线B对应的分类器的性能。但ROC曲线不能定量地反应分类器的性能，通常使用ROC线下曲线的面积，简称AUC（Area under ROC）值来定量衡量分类器的性能。根据（fawcett,2006），AUC值有一个重要的统计意义：当随机选取一个正样本和一个负样本，正样本被预测为正类的概率值大于负样本的概率值的概率，这和Wilcoxon-Mann-Witney Test[[73](#_ENREF_73)]是等价的。因为直接去计算曲线下的面积是很难的，所以很多研究是根据这个统计性质来计算AUC值得。此外，AUC值和基尼系数[[74](#_ENREF_74)]也有密切的关联，当基尼系数用在ROC曲线时，其是表示ROC曲线和对角线围成的面积与对角线之上的面积的比。Hand和Till[[75](#_ENREF_75)]指出:。AUC值介于0到1之间，而且值越大，对应分类器的性能越好。由于随机分类器的ROC曲线对应于ROC上的对角线，所以一个有效的分类器的AUC应大于0.5。



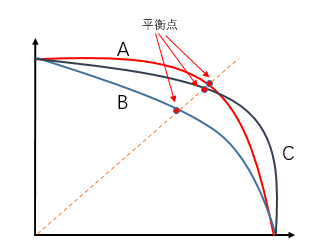
**图 3. 1：ROC曲线示例图**

## Precision-Recall曲线

虽然ROC曲线是可视化分类器性能的一个有利工具，但它也存在一定缺陷。在数据分布非常不平衡的情况下，ROC曲线会过高地评价分类器的性能。具体的原因，会在4.1节解释。在这种情形下，PR曲线会比ROC曲线提供更多的信息[[12](#_ENREF_12)]。

和ROC曲线类似地，我们根据学习器输出的结果对样本进行降序排序。设定不同的划分阈值，将样本分为正负类，计算得到一系列的(查全率，查准率)点对。然后以查全率为横轴，查准率为纵轴作图，就得到了查准率-查全率曲线，简称PR曲线。PR曲线和ROC曲线有很强的相关性:在ROC图中，一条曲线A覆盖另一条曲线B，当且仅当曲线A在PR图中覆盖曲线B[[12](#_ENREF_12)]。所以同样地，学习器A的PR曲线覆盖学习器B的PR曲线，则学习器A的性能优于学习器B。当发生交叉时，如曲线A和曲线C，可以比较两条曲线下的面积。它在一定程度上表征了学习器在查准率和查全率上取得相对双高的比例。但这个值一般不好计算，所以，大家设计了一些综合考虑查全率和查准率的性能指标。其中平衡点（Break-Even Point，简称BEP）就是其中一个，它是查全率等于查准率时的取值。如图 3. 2中学习器A的BEP大于学习器C的BEP，则可以认为学习器A优于学习器C。

为了说明PR曲线能提供比ROC曲线，更多的信息。让我们考虑在一个正负样本分布差别非常大的数据分类场景中，假设学习器发生改变，导致很多错分正例产生。在ROC中，因为负样本总数非常大，所以FPRate并不会发生很大的改变。而另一方面，PR曲线中的P会发生较大的变化。所以当错分正例发生大改变时，PR曲线能捕抓到学习器的性能变化。鉴于此，当数据分布非常不平衡时，PR曲线是一个更好的性能衡量指标。已经有很多的研究工作使用PR曲线来衡量学习器的分类性能[[76](#_ENREF_76), [77](#_ENREF_77)]。



**图 3. 2： Precision-Recall曲线示例图**

## 代价敏感错误率与代价曲线

ROC曲线的另外一个缺陷是它们无法提供对分类器性能评估的置信区间，而且也无法给出不同分类器性能的统计意义[[13](#_ENREF_13), [62](#_ENREF_62)]。此外，当数据的类别分布发生变化，或者错分代价发生变化时，ROC曲线也没办法追踪分类器性能的变化[[13](#_ENREF_13), [62](#_ENREF_62)]。为了解决这些问题，有研究者提出了另外一个更综合性的衡量指标，即代价曲线。

在实际的很多应用中，不同类型的错误造成的后果是不一样的。比如在癌症诊断中，如果把正常人误判为患癌，则后果是多做了进一步检查。而把癌症患者误判为健康，虽然也是患了一次错误，但是后果却可能是失去了拯救生命的最佳时机；再如，门禁检测系统:错误地把合法的人员拒在门外，会减低用户体验。但是错误地把入侵者放进门内，则会造成严重的安全事故。对于这些错误后果不一样的应用，可以为不同错误赋予不同的代价。

在二分类问题中，可以根据领域知识设定一个代价矩阵，如表 3. 2所示，其中表示将第i类样本预测为第j类样本的代价。一般来说，；如果将0类误判为1类所造成的损失大于将1类误判为0类的损失，则；

**表 3. 2： 二分类代价矩阵**

|  |  |  |
| --- | --- | --- |
| 真实类别 | 预测类别 | |
| 第0类 | 第1类 |
| 第0类 | 0 |  |
| 第1类 |  | 0 |

在前几节介绍的衡量指标中，大都隐式地假设了均等代价，没有考虑不同错误会带来不同的后果。在非均等代价下，我们所希望的不再是简单地最小化错误次数，而是想最小化总体代价。如果将代价矩阵中的第0类作为正类、第1类作为负类，令和分别代表正样本子集和负样本子集，则代价敏感错误率为:

(公式 3. 15)

类似地，可以给出基于分布定义的代价敏感错误率，以及其他一些性能衡量指标如正确率的代价敏感版本。另外，如果中的i，j取值不限于0、1，则可以定义出多分类问题的代价敏感性能衡量指标。

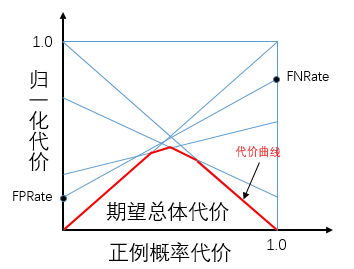
在非均等代价下，ROC曲线不能直接反映出分类器的期望总体代价，而代价曲线则可以实现这个目标。代价曲线图的横轴是取值为[0,1]的正例概率代价

(公式 3. 16)

其中p是样本为正例的概率；纵轴是取值为[0,1]的归一化代价

(公式 3. 17)

其中FPRate是公式3.2定义的假正例率，FNRate=1-TPRate是假反例率。代价曲线的绘制方法如下：ROC曲线上每一点对应了代价平面上的一条线段，设ROC曲线上点的坐标为（TPRate,FPRate），则可对应计算出FNRate，然后在代价平面上绘制一条从(0,FPRate)到（1，FNRate）的线段，线段下的面积即表示了该条件下的期望总体代价；如此将ROC曲线上的每个点转化为代价平面上的一条线段，然后取所有线段的下界，围成的面积即为在所有条件下分类器的期望总体代价，在图 3. 3中，粗线与横轴围成的面积即期望总体代价。



**图 3. 3：代价曲线示例图**

## 不平衡多分类评价指标

虽然到目前为止讨论的评价指标都是二分类情况下使用的，部分指标通过简单的修改就能适用到多分类的情况。例如，[[71](#_ENREF_71), [72](#_ENREF_72)]讨论了多分类情况下的ROC曲线图。对于一个n分类问题，分类结果会产生一个的混淆矩阵，包括n个正确分类的单元，和个错误分类的单元。不像在二分类，只有一对TP和FP，现在我们一共有n个收益(benefits)和个代价(cost)需要处理。一种直接的做法是生成n张的ROC图，每个类别一张。在第i张里把第i类当成正类，其他所有类别当成负类。这种方法虽然简单直接，但是会造成我们在3.3节讨论的ROC的缺陷。因为在这种设定下，负类是由其他的所有n-1类组成的，所以类别分布严重倾斜。

类似地，在多分类问题中，会产生多个AUC值。Provost 和Domingos[[78](#_ENREF_78)]提出了一种基于概率估计的方法。首先画出每个类别对应的ROC图，并计算AUC值。然后计算每个类别的分布概率，将这个概率作为每个类别的一个因子，将所有的ROC值累加起来得到一个最终的ROC值。虽然这种计算方式简单，但同样面临前面提到的类别分布倾斜问题。为了解决这种局限性，Hand和Till提出了一个M评估指标，它是对AUC的推广，很好地继承了AUC的内在特性，并且能用于多分类问题中。M指标的主要优势是对类别分布和错分代价不敏感。

除了多类别的ROC分析，也有多种其他评估指标被提出，用以衡量多不平衡分类的效果。例如，在代价敏感学习中，可以将错分代价作为性能评价指标[[79](#_ENREF_79)]。另外，Sun等人[[80](#_ENREF_80)]扩展了G-mean的定义，将其改成多个类别的查全率的几何平均。

## 本章小结

在这一章里，我们主要讨论用来评估分类器性能的一些主流评估指标。评估指标在模型训练过程中有着非常重要的作用，针对不一样的应用场景，往往需要不同的指标来评估。首先介绍了在分类中常用的单一指标，如准确率、查全率、查准率等。这些指标比较简单，导致要么不适用不平衡分类场景，要么表现的信息不足。

其次，介绍了在不平衡分类中最常用的ROC曲线和AUC值。它们对类别分布不敏感，适合不平衡分类场景，而且具有良好的可视性。但在类别分布倾斜太严重时，ROC曲线会存在过高评价分类性能，以及对分类器性能变化不敏感的问题。

基于此，介绍了PR曲线和代价曲线。这两种曲线更适合在类别分布倾斜严重时使用，而且能提供比ROC曲线更多的信息。最后简单的介绍不平衡多分类场景下的性能评价指标。

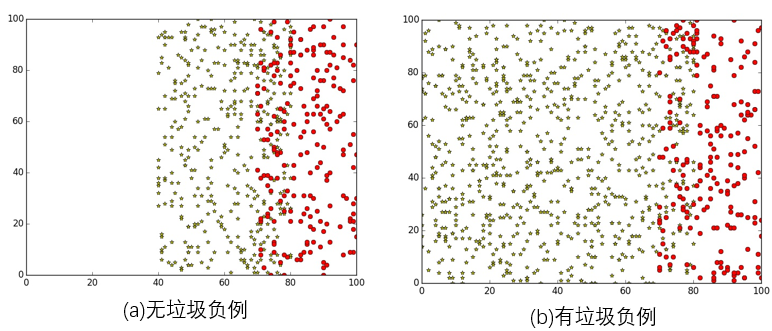
# 最优分类阈值

在第三章中，介绍了用来评估不平衡数据分类结果的各种指标。不平衡数据的分类因其特有的类别分布倾斜问题，导致以往常用的指标，如准确率等，不能再正确的评估分类器的分类效果。

在不平衡数据领域里，最常用的指标是ROC曲线和其对应的AUC值。ROC曲线具有对类别的分布不敏感，而且可以很好地可视化分类结果等优点。然而，在一些分布严重倾斜的问题中，ROC曲线在评估分类器性能时，会出现问题。在第四章中，我们是介绍其他指标，如PR曲线、代价曲线来替换ROC曲线。而在本章中，我们将从另外一个角度来解决ROC曲线的缺陷问题。

## ROC 的缺陷

ROC曲线常被用来指导不平衡数据的分类，然而当数据集非常不平衡时，实际的分类性能可能没有AUC值所反映的那么好。因为在不平衡数据集中可能存在很多‘垃圾’负样本，即那些很容易就能区分为负例的样本。这些‘垃圾’负样本在分类时提供的信息少，但却会拔高AUC值。仅有很少的一部分负样本是和正样本混合在一起的，这些负样本很难被区分。分类器的一些主要指标，如查准率，查全率就是受这少部分负样本影响的，然而这些负样本对AUC值的影响并不大。为了有个直观的理解，我们构造了两个数据集。第一个数据集中有400个负例和200个正例，在第二个数据集中，添加了200个远离分界面的负例，如图 4. 1所示。然后用同一个分类器对两个数据集进行分类，基本指标列于表 4. 1。



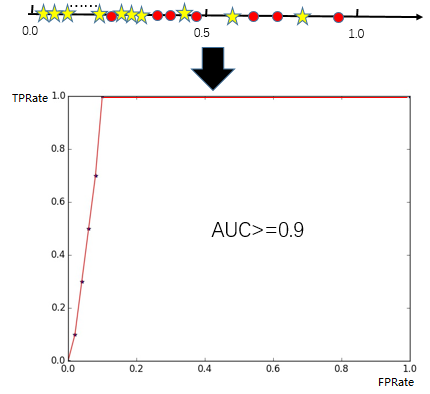
**图 4. 1：垃圾负例示例图**

**表 4. 1：垃圾负例分类结果表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 查准率 | 查全率 | F1值 | AUC值 |
| 无垃圾负例 | 1 | 0.62 | 0.765 | 0.787 |
| 垃圾负例 | 0.896 | 0.690 | 0.780 | 0.975 |

在表 4. 1中，两个数据集在查准率、查全率和F1值都相差不大，但是有垃圾负例的数据集其AUC值要远大于无垃圾负例数据集的AUC值。AUC50正是为了解决这个问题，而被提出的一种AUC改进指标。AUC50是只计算累计到前50个错分正例时的AUC值。虽然AUC50在某种程度上减轻了‘垃圾’负样本的影响，但是不同数据集的分布差别很大，统一定前50个错分正例这个数，显得太过随意。一种情况是所有正例都在排在前50个错分正例之前，则AUC50和AUC计算是一样的，当然这是较好的情况，说明排序结果很好。但在另一方面，如果数据集非常大，则前50个错分正例只考虑了很少的一部分样本，不具备衡量分类器性能的意义。所以虽然AUC50在衡量分类器性能会比单纯的AUC好一点，但并没有从根本上消除问题。

实际上AUC值仅是衡量了分类器对样本的排序能力。进一步的，高AUC值不意味着高准确率。比如，在一个人工数据集中，只有1%的正例样本。当这所有正例样本的得分，都在前10%的时候，AUC值的得分就可以达到0.9以上。如果分类阈值设置的不合理，比如最常见的把分类阈值设为0.5时，分类准确率将会非常低，不到0.5。



**图 4. 2：正例得分排在前10%时AUC曲线图**

以往大多数的研究都把精力放在提高AUC值上，而对于如何在测试集或未知的数据集中寻找一个最优的划分阈值，却鲜有人关注。然而如前面所述，即使能取得很高的AUC值，当分类阈值设置不合理时，其他的指标如查准率，查全率或者F-measure都可能很不理想。因此本文主要研究了如何寻找最优的分类阈值，与我们工作最相似的研究是[[81](#_ENREF_81), [82](#_ENREF_82)]。[[81](#_ENREF_81)]提出了一个ROC convex hull(ROCCH)方法，试图在一系列不同环境下，选择最优的分类器。在[[82](#_ENREF_82)]他们详细地分析了AUC和错误率之间的关系，然而他们的工作缺乏普遍性，难以推广应用。

## 寻找最优分类阈值

在本文中，我们将研究如何在训练集寻找一个合适的划分阈值，并将这个划分阈值应用到测试数据集或未知数据集中，使得分类结果在其他非AUC指标，如查全率、查准率和F-measure上也能取得较好的表现。具体地，我们将首先研究在正负类的错分代价相同时，如何寻找最优分类阈值。然后将延伸至正负类错分代价不同场景下，最优分类阈值的确定。

在正负类错分代价相同的场景下，我们考虑使用正类的F-measure做为指标，来指导最优阈值的查找。在不平衡数据分类问题中，我们更关心的是少数类（一般设为正类）的分类情况。设定不同分类阈值，对AUC值没有影响，只会影响查准率和查全率。F-measure是对查准率和查全率的一个权衡，因此在寻找最优阈值时，我们考虑使用F--mseaure来指导。基本思想是，设定不同的阈值，计算该阈值下的F值。寻找取得最大F值时的阈值，作为最优分类阈值。这里涉及到两个问题，如何寻找训练集的最优阈值，以及如何将这个最优阈值应用到测试集或者未知数据集中。

### 错分代价相同的最优分类阈值

如果训练集的数据量不大，则可以将所有样本的得分从高到低排序。本文如不做特别说明，则所说的得分均是指样本被预测为正类的概率值。取相邻两个得分的均值做为一个候选阈值。遍历所有的候选阈值，找到F值最大时的阈值做为最优的，如算法1；然而对于数据量大的，我们设计了一个算法，在不遍历所有的候选阈值的情况下，找出最优阈值。在数据量大时，我们直接把某个样本的得分设为阈值。下面我们先介绍4个结论，然后给出具体算法。为方便证明，我们先假设每个样本的预测得分都是不一样的，后面再说明当有样本得分相同情况下的处理方法。

|  |
| --- |
| 算法1：遍历寻找最优分类阈值  输入：所有样本的预测得分（,…）  输出：最优阈值T\* |
| 1.将预测得分进行降序排序 |
| 2.; |
| 3. for i=1 to n-1 |
| 4. |
| 5.根据进行分类，计算 |
| 6.if |
| 7.; |
| 8.end if |
| 9.end for |
| 10.return |

首先定义是一个样本，是所有正类样本的集合，是所有负类样本的集合；是按降序排序的得分序列； 定义为得分不小于的正样本的个数，即。类似地，定义为得分不小于的负样本的个数；当选取作为划分阈值时，可计算F值为：

(公式 3. 18)

结论1：最优阈值只可能位于正类样本中；

证明：假设最优阈值为 并且 是一个负类样本，那么有：

(公式 3. 19)

现在选择一个得分最接近且小于的正例，不妨设为，容易得到，并且有：

(公式 3. 20)

这和之前设定的是最优分类阈值矛盾，所以最优分类阈值只可能产生于正例中。

结论2：如果 和 都是正例，则有。

证明：因为 和 都是正例，有，则有：

(公式 3. 21)

另外定义是X序列中第j个正例的得分，是把设为阈值时的F值，则从结论1中，可以得到，并且有结论3如下。

结论3：如果第j个正样本的得分是，并且其后的K个样本都是负样本，那么最大的是：.

证明：取得最大值时，仅当第个样本是正样本。假设得分在和之间有y个负样本，则有

(公式 3. 22)

当y=0时，取得最大值。同理，的最大值是，以此类推，最后一个正样本的最大F值为。又根据结论2，我们有 <<...<，因此结论3得证。

结论4：如果是目前找到的最大F值，并且其后的K个样本都是负类样本，则就是我们要寻找的最优阈值，其中.

证明：根据结论3，在之后的最优F值是：

(公式 3. 23)

令，并求解不等式

(公式 3. 24)

解得，结合，可以得到：

(公式 3. 25)

上面的证明都是基于样本得分都不一样的情况，其实对于出现正负样本得分一样的情况，结论也是成立的。假设在得分下有多个正负样本，因为如果把该得分设为划分阈值，则该得分下的所有正负样本都是分为正类。那么可以给负样本的得分加上一个假想的很小的正数，使负样本排在前面，容易得知，这样处理后不影响划分结果，而且也不影响相关结论的成立。给定一个训练集，结合上面的结论，我们设计了一个时间复杂度为的高效求解最优阈值的算法，列于算法2。

|  |
| --- |
| 算法2: 启发式寻找最优阈值 |
| 输入:降序排序的得分序列； 值；正样本个数 |
| 输出:最优阈值 |
| 1.begin |
| 2. Init |
| 3. while is negative do |
| 4. |
| 5. |
| 6. end while |
| 7. ; |
| 8. |
| 9. for to n do |
| 10. if is negative |
| 11. |
| 12. else |
| 13. if |
| 14. break; |
| 15. else |
| 16. |
| 17. |
| 18. |
| 19. |
| 20. if |
| 21. |
| 22. ; |
| 23. end if |
| 24. end if |
| 25. end if |
| 26. end for |
| 27. end |

### 错分代价不同的最优分类阈值

前面介绍的是在正负类的错分代价相同的场景下，最优分类阈值的寻找。然而在现实应用中，正类的错分代价往往是远大于负类的错分代价的。在错分代价不同时，就不适合用F1作为分类性能的评价指标，在这里我们采用最小错分代价为指标来指导最优分类阈值的寻找。

**表 4. 2：二分类错分代价矩阵**

|  |  |  |
| --- | --- | --- |
| 真实类别 | 预测类别 | |
| 第0类 | 第1类 |
| 第0类 |  |  |
| 第1类 |  |  |

假设有如上的错分代价矩阵，一般地有。沿用4.2.1节中的定义，当选取作为划分阈值时，定义此时整体的错分代价为：

(公式 3. 26)

则我们的目标就是找到最优的划分阈值，使得整体的错分代价最小。容易得到，将上述式子除以一个正数，不影响最优阈值的寻找，所以我们将优化目标除以，且令，则整体错分代价变为：

(公式 3. 27)

与4.2.1节类似，在数据量比较少的时候，可以遍历所有的数据点，从而找到整体错分代价最小的划分阈值。在数据量比较大的时候，同样可以推导出如下的4个结论:

（1）最优阈值只可能位于正类样本中。

（2）如果 和 都是正例，则有。

（3）如果第j个正样本的得分是，并且其后的K个样本都是负样本，那么最小的是：。

（4）如果是目前找到的最小的全局错分代价，并且其后的K个样本都是负类样本，则就是我们要寻找的最优阈值，其中。

证明过程和4.2.1节类似，这里就不再赘述。下面是算法：

|  |
| --- |
| 算法3: 错分代价不同的最优阈值 |
| 输入:降序排序的得分序列； 值；正样本个数 |
| 输出:最优阈值 |
| 1.begin |
| 2. Init |
| 3. while is negative do |
| 4. |
| 5. |
| 6. end while |
| 7. ; |
| 8. |
| 9. for to n do |
| 10. if is negative |
| 11. |
| 12. else |
| 13. if |
| 14. break; |
| 15. else |
| 16. |
| 17. |
| 18. |
| 19. |
| 20. if |
| 21. |
| 22. ; |
| 23. end if |
| 24. end if |
| 25. end if |
| 26. end for |
| 27. end |

### 最优分类阈值运用

在寻找到训练集的最优阈值后，要如何应用这个阈值，来指导测试集或未知数据集的划分是我们研究的另外一个问题。最初我们是直接将这个阈值应用到测试集里面，但实验发现，这样做得到的结果不理想。仔细分析后发现，测试集和训练集的得分分布可能差异很大。在我们实验的一个数据集中，训练集的得分区间在0.55到0.98之间，而测试集的得分区间是0.25到0.81。

后面我们尝试了另外一种方法，即使用训练集的最优阈值的位置信息，来指导测试集的划分。在求得训练集的最优分类阈值后，统计该阈值在得分序列排前百分之几。然后去测试集找同样占比的阈值作为测试集的最优阈值。首先定义训练集中最大的得分为Maxtrain，同样的定义Mintrain，Maxtest，Mintest。其次定义训练集的最优阈值为Thresholdtrain，则测试集的最优阈值Thresholdtest为：

(公式 3. 28)

实验发现，第二种使用方法的结果要优于第一种，两种方法的展示如图 4. 3所示：



**图 4. 3：最优阈值使用方式**

最后，为了避免过拟合，在实验中，会将训练集切分成多份。每份数据分别计算得到一个最优阈值的位置信息。然后对这些位置信息取均值，用均值去指导测试集的划分，整个算法的流程展示于图 4. 4中：



**图 4. 4：寻找最优阈值流程图**

## 本章小结

在本章中，我们从另外一个角度，来探讨解决ROC曲线作为评估指标时的缺陷问题。ROC曲线及其对应的AUC值，在评估学习器的分类结果时，主要是衡量了学习器对样本的一个排序能力。高AUC值，并不一定保证分类结果一定是也好的，需要设定一个合理的划分阈值。

为此，本文提出了一个寻找最优分类阈值的框架。首先，在正负类错分代价相同的场景下，采用了F1值作为评估指标，衡量划分的结果。然后采用了一个时间复杂度为线性的算法，查找训练集中的最优分类阈值。而在正负类错分代价不同的时候，则是以最小化整体的错分代价为目标，寻找最优分类阈值。最后尝试了两种将训练集的最优阈值，应用到测试集或未知数据集的策略。第一种直接采用值的方式，第二种是利用位置信息的方式。在下一章中，我们将采用一个远程蛋白质同源检测的数据集，来验证本文方法的有效性。

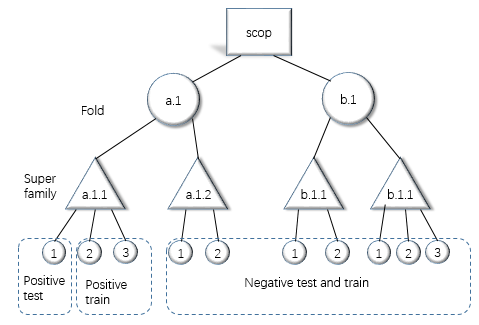
# 蛋白质远程同源检测

在前面的章节中，研究了ROC曲线和其对应的AUC值，在评估不平衡数据分类时的不足之处。针对这些不足，提出了一个寻找最优分类阈值的框架。在本章中，我们将采用一个蛋白质远程同源检测的数据集，并做了多组对比实验，来证明本文所提方法的有效性。

## 数据集

蛋白质远程同源性检测是蛋白质结构和功能研究中一个重要的问题。蛋白质远程同源性检测是指利用蛋白质序列间的相似性检测一个新的蛋白质的家族，之后凭借现有的蛋白质家族的结构与功能信息来推测新的蛋白质的结构、功能。该领域中的计算方法利用蛋白质的序列信息、理化性质和进化信息提取特征，使用机器学习算法来预测蛋白质远程同源性。

为了验证方法的有效性，使用了Li et al.2003的一个基准数据集[[83](#_ENREF_83)]来做实验。该数据集是用来做蛋白质的远程同源检测的基准数据，被许多研究[[84](#_ENREF_84), [85](#_ENREF_85)]拿来做比较。这个数据集是从1.53版的SCOP数据库导出的，共有4352条蛋白质序列数据。这些序列分别被分成54个家族，每对序列之间的相似性是小于10-25。SCOP数据库的层次结构如图 5. 1所示。其中折叠子(fold)层次描述空间几何结构的关系，超家族(superfamily)层次描述远源的进化关系，家族(family)层次描述相近的进化关系。我们按如下方式构建了54个数据集：每个家族的序列作为测试集的正样本，不属于同一家族但属于同一父族的序列作为训练集的正样本。另外不属于同一父族的序列，构成训练集和测试集里的负样本。数据集的基本信息列于下表 5. 1中。



**图 5. 1：SCOP数据库的层次结构**

**表 5. 1：蛋白质远程同源检测数据集**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Index** | **Train** | **Test** | **Index** | **Train** | **Test** |
| 1 | 12/2330 | 9/1746 | 28 | 11/423 | 95/3653 |
| 2 | 11/2509 | 8/1824 | 29 | 46/3880 | 5/421 |
| 3 | 19/3002 | 8/1263 | 30 | 43/3627 | 8/674 |
| 4 | 17/2686 | 10/1579 | 31 | 33/3203 | 9/873 |
| 5 | 12/2890 | 6/1444 | 32 | 16/1553 | 26/2523 |
| 6 | 10/2408 | 8/1926 | 33 | 37/3591 | 5/485 |
| 7 | 29/3208 | 10/1105 | 34 | 30/3682 | 5/613 |
| 8 | 33/3650 | 6/663 | 35 | 90/3102 | 31/1068 |
| 9 | 26/2256 | 23/1994 | 36 | 99/3412 | 22/758 |
| 10 | 17/2370 | 14/1951 | 37 | 42/3542 | 9/759 |
| 11 | 41/3557 | 8/693 | 38 | 24/2946 | 11/1349 |
| 12 | 26/3625 | 5/696 | 39 | 113/3895 | 8/275 |
| 13 | 40/3470 | 9/780 | 40 | 88/3033 | 33/1137 |
| 14 | 11/307 | 140/3894 | 41 | 26/3191 | 9/1104 |
| 15 | 21/2928 | 10/1393 | 42 | 94/3240 | 27/930 |
| 16 | 26/2876 | 13/1437 | 43 | 20/3204 | 7/1121 |
| 17 | 37/3002 | 16/1297 | 44 | 12/3060 | 5/1275 |
| 18 | 34/3761 | 5/552 | 45 | 13/2083 | 14/2242 |
| 19 | 44/3569 | 9/730 | 46 | 29/3477 | 7/839 |
| 20 | 46/3732 | 7/567 | 47 | 40/3374 | 11/927 |
| 21 | 46/3732 | 7/567 | 48 | 10/1199 | 26/3117 |
| 22 | 48/3894 | 5/405 | 49 | 10/2241 | 9/2016 |
| 23 | 18/1246 | 44/3044 | 50 | 10/2241 | 9/2016 |
| 24 | 22/3280 | 7/1043 | 51 | 36/3692 | 6/615 |
| 25 | 48/3894 | 5/405 | 52 | 13/2345 | 11/1983 |
| 26 | 56/3875 | 6/415 | 53 | 14/2525 | 10/1803 |
| 27 | 13/1938 | 16/2385 | 54 | 17/1744 | 25/2563 |

从表 5. 1可以看出，每个家族都是不平衡的数据集。其中最高的不平衡比达到241：1，一些家族的训练集里只有10条正样本，因此这个数据集很适合本文做研究。

## 特征提取

在实验中，膜蛋白序列的特征提取方法，采用的是一个基于氨基酸的组分-理化性质而构建的188维特征向量[[86](#_ENREF_86)]。其中的组分是指氨基酸序列的组成成分，它表征了各类氨基酸在某一个氨基酸序列中的含量，共20维特征。理化性质是氨基酸的8种物理化学性质，具体包括：疏水性，标准范德华力体积，极性，极化性，电荷，表面张力，二级结构以及可溶性。不同种氨基酸具有各自特殊的理化性质，因此，可以通过提取氨基酸序列的理化性质作为特征来预测蛋白质的类型。

在生物信息里，氨基酸的每种理化特性都有三种不同状态，例如疏水性可以分为：疏水型，亲水型和中性型，可以简记为。在组合这些特征时，我们用三个表示器来描述特征的构成：表示组成，表示转换，D表示分布。具体地，对于特性，C表示具有状态的氨基酸占全部氨基酸的比重，表示具有状态的氨基酸后面跟具有状态的氨基酸的频率（和记作一种转换），表示具有状态的第一个，第25%个，第50%个，第75%个以及最后一个氨基酸分别处在序列中的位置。由此，对于每种特性都有3+3+15=21维特征。8种特性共构成了168维理化特征，加上前面的20维组分特征，共有188维特征。氨基酸组分-理化特征提取速度快，而且已经被很多研究拿来应用，并取得很好的效果[[87](#_ENREF_87), [88](#_ENREF_88)]，因此在本文中采用氨基酸组分-理化特性来提取特征。

分类算法采用的是随机森林。每个家族得到的AUC值列于表 5. 2，从表里可以知道，大多数家族的AUC值都比较大，在0.8以上。

**表 5. 2：各家族AUC值**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Index | AUC | Index | AUC | Index | AUC | Index | AUC |
| 1 | 0.973 | 15 | 0.982 | 29 | 0.717 | 43 | 0.823 |
| 2 | 0.809 | 16 | 0.690 | 30 | 0.883 | 44 | 0.766 |
| 3 | 0.983 | 17 | 0.871 | 31 | 0.972 | 45 | 0.848 |
| 4 | 0.963 | 18 | 0.768 | 32 | 0.971 | 46 | 0.820 |
| 5 | 0.785 | 19 | 0.860 | 33 | 0.998 | 47 | 0.925 |
| 6 | 0.964 | 20 | 0.948 | 34 | 0.587 | 48 | 0.901 |
| 7 | 0.791 | 21 | 0.982 | 35 | 0.936 | 49 | 0.833 |
| 8 | 0.894 | 22 | 0.986 | 36 | 0.983 | 50 | 0.981 |
| 9 | 0.949 | 23 | 0.666 | 37 | 0.933 | 51 | 0.913 |
| 10 | 0.950 | 24 | 0.667 | 38 | 0.569 | 52 | 0.847 |
| 11 | 0.961 | 25 | 0.895 | 39 | 0.908 | 53 | 0.819 |
| 12 | 0.986 | 26 | 0.644 | 40 | 0.952 | 54 | 0.943 |
| 13 | 0.998 | 27 | 0.683 | 41 | 0.757 | average | 0.861 |
| 14 | 0.517 | 28 | 0.981 | 42 | 0.790 |  |  |

## 实验比较

### 统一阈值

因为之前很少有学者对这方面进行研究，所以没找到其他研究结果来和本文方法进行比较。在本文中，我们设计了几个对比实验来验证方法的有效性。在实验中采用F1值来衡量分类的效果。F1是F-measure在时的计算公式，定义为：， 其中P和R分别是正类的查准率和查全率。

首先，对所有家族都设定一个统一的划分阈值。这里分别取了0.1、0.2、0.3和0.5坐为阈值，然后计算每个阈值下所有家族的查准率、查全率以及F1值，然后取平均值，结果列于表 5. 3中。在表 5. 2中，大多数家族的AUC都在0.8以上。但从表 5. 3看，平均的查全率、查准率以及F1值都不是很理想。特别地，大多数应用默认设定的0.5为划分阈值，而其结果要远劣于其他的阈值。最好的分类结果是在0.1取得，此时平均查准率为0.169，平均查全率为0.223，平均F1值是0.18。

**表 5. 3：统一阈值的分类结果**

|  |  |  |  |
| --- | --- | --- | --- |
| Threshold | MeanPrecision | MeanRecall | MeanF1 |
| 0.5 | 0.01852 | 0.00013 | 0.00026 |
| 0.3 | 0.03063 | 0.00251 | 0.00463 |
| 0.2 | **0.17411** | 0.04963 | 0.07171 |
| 0.1 | 0.16850 | **0.22259** | **0.18001** |

### 最优阈值的使用方式

然后，我们采用本文提出的方法，对每个家族寻找一个最优阈值。5.2.2节提到，有两种方法来使用训练集的最优阈值，指导测试集的划分。这里我们设计了一个对比实验，来对比这两种方法的效果，结果如**表 5. 4**所示：

**表 5. 4：最优阈值使用方式比较**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 折数 | 值 | | | | | 位置信息 | | |
| Precision | | Recall | | F1 | Precision | Recall | F1 |
| 2 | 0.118 | 0.156 | | 0.105 | | 0.165 | 0.280 | 0.159 |
| 3 | 0.139 | 0.127 | | 0.110 | | **0.205** | 0.330 | **0.204** |
| 5 | 0.139 | 0.147 | | 0.123 | | 0.160 | 0.343 | 0.187 |
| 10 | 0.132 | 0.146 | | 0.121 | | 0.156 | **0.388** | 0.190 |

对比表 5. 3，直接使用值的方式，其分类结果要远优于把阈值统一设为0.5时的分类结果。不过要稍劣于统一阈值为0.1时的分类结果。而采用位置信息的方式，分类性能要优于直接使用值的方式。当把训练集分成3份时，位置信息方式取得最佳效果。平均的查准率与平均的查全率分别达到0.205和0.33，F1值达到0.204。

### 调节因子

此外在寻找训练集的最优阈值时，采用F-measure来衡量每个阈值下的分类性能。其中有个参数是用来调节查准率和查全率，我们设计了一组实验，观察在不同下，算法的效果，结果列于**表 5. 5**中:

**表 5. 5：不同值下的分类性能**

|  |  |  |  |
| --- | --- | --- | --- |
|  | MeanPrecision | MeanRecall | MeanF1 |
| 0.5 | **0.2452** | 0.2633 | 0.1884 |
| 0.8 | 0.2284 | 0.3091 | **0.2066** |
| 1.0 | 0.2051 | 0.3303 | 0.2044 |
| 1.2 | 0.1948 | 0.3364 | 0.2036 |
| 1.5 | 0.1725 | **0.3729** | 0.1996 |

如表 5. 5所展示的，首先平均查全率随着值的增加而增加，而平均查准率则是在递减。这与可以权衡查全率和查准率相符合，在实际应用中，可以根据要求调节如在论坛的反作弊任务中（打击黄反帖子），作弊帖子是少数类。如果把非作弊贴误判为作弊帖而删掉，会影响论坛的用户体验。所以在这个应用中，反作弊者要求查准率要非常高。而在医院癌症诊断中，患病是少数类。而在这种情况下，医生更关注查全率。

其次，在远程同源蛋白质检测这个数据集中，当把值设为0.8时，取得最好的分类效果。平均的查准率和平均的查全率达到0.228和0.309，F1达到0.207。最佳不同的正负比导致的，测试集的正负比更高，提高查全率有利于整体的性能。

### 最优分类结果上界

上面所有的实验结果中，虽然相对的平均查准率和查全率，可以比较出不同设定下的分类性能，但是绝对值看起来似乎并不高。而且如之前提到的，很少有研究这方面内容的相似工作，所以也没能找到其他人的实验数据来做对比。为此设计了一个实验来证明我们方法的有效性。首先把寻找训练集最优阈值的算法直接应用到测试集，计算在最优阈值下测试集的F1值。这个F1值是理论上所能达到的最优结果。经过实验，测试集最优的平均查准率和平均查全率是0.284与0.494，平均最优F1值是0.292。而用算法找到的最优阈值，平均F1值是0.207，和最优F1值比较接近。

为了更详细的分析，我们把每个家族单独列出来。在**图 5. 2**中，横轴是我们算法所得到的F1值，纵轴是理论最优的F1值，每个小三角代表一个家族的F1值对。在**图 5. 2**里，大多数点是位于对角线附近的，说明算法找到的最优阈值下的F1值，是接近理论最优的F1值。特别地，有几个家族正好位于对角线之上，表示这几个家族取得理论上最优的分类效果。通过这个实验，证明本文所提的寻找最优分类阈值是有效的。



**图 5. 2：各家族的最优分类边界**

### AUC值与F1值

最后，我们把每个家族的AUC值和F1都列出来，如**图 5. 3**所示。从图可以看出，AUC值和F1值的分布趋势基本是一致的，高AUC的家族同样会有高的F1值。特别地，对那些AUC值大于0.9的家族，其F1值都高于0.06，说明本文算法对AUC值高的数据更有效果。



**图 5. 3：ROC与F1分布图**

## 本章小结

本章主要是实验分析部分。因为前人相似的工作比较少，很少有数据能拿来直接和本文的工作做比较。因此在本章中，设计了较多的对比实验，来验证本文方法的有效性。首先证明了传统的把划分阈值默认设为0.5时，分类的效果很不理想。其次比对了两种应用最优阈值方法的效果，实验证明，采用位置信息的方式，其结果要优于直接使用值的方式。另外还设计实验，比较了本文方法所得到的结果与理论上能实现的最优结果，验证了本文方法的有效性。

# 总结与展望

在这章中总结了本文的主要工作，然后对未来的研究做了展望。

## 本文总结

不平衡数据分类，是机器学习和数据挖掘领域里，一个热门的研究难点。一方面，该问题广泛存在于现实生活中，解决好这个问题，可以给生产，生活带来益处。另一方面，传统的分类算法，在不平衡的数据上，性能可能会大大的下降。本文在广泛学习前人工作的基础上，总结了目前解决不平衡数据分类的方法。并深入地研究，如何寻找数据集的最优分类阈值，主要完成了以下工作：

1. 总结了国内外在解决不平衡数据分类时，所采用的策略。主要包括两方面:一个是在数据层面上的处理，使用采样算法，将数据集修改成平衡数据。另外一个是在算法层面上的处理，通过修改代价函数等手段，使传统分类算法能作用在不平衡数据上。
2. 深入地介绍了5种数据内在特征。包括小析取项、数据密度稀疏、类别覆盖、噪音数据以及分布迁移。阐述这些特征与数据分布不平衡问题共同出现时，导致的困难，并介绍对应的解决方案。
3. 介绍分类中常用的衡量指标，并分析解释，为何常用指标不适合不平衡数据分类场景；重点介绍了AUC指标，说明其在衡量不平衡数据分类时的优势，以及不足之处。
4. 针对AUC的缺陷，提出了一个寻找最优分类阈值的框架。具体地，根据错分代价是否相同，分别以F1值和总体最小错分代价为目标，设计了寻找训练集的最优分类阈值的算法。另外研究了，如何用训练集的最优分类阈值来指导测试集或未知数据集的分类。最后做了对比实验，证明本文提出的算法的有效性。

## 未来展望

在不平衡分类问题中，以往的大多数工作都把注意力放在，如何尽可能地提高AUC值。而高AUC值，只代表了正负样本排序较好，为取得好的分类结果，还需确定一个合理的分类阈值。本文正是从这点出发，探讨如何从训练集寻找一个最优的分类阈值，去指导测试集或未知数据集的分类，使得分类结果在精确度，F值都能取得好的结果。在未来的工作中，还有很多问题可以研究：

1. 多分类情况。虽然在实际应用中，大多数是二分类问题，本文也只讨论了二分类情况，不过多分类的问题也是存在的。在多分类场景下，问题变得更加复杂，不能通过简单设定一个阈值来解决。可以考虑的解决方案有：将多分类转换成二分类问题，包括one vs all的形式，或分层次组合的形式；采用逐层划分的形式，先根据样本属于第一类的概率及一个阈值判断是否将样本归为第一类，如果不属于第一类，再根据第二类的概率和另外一个阈值判断是否属于第二类，依次类推。
2. 数据偏移问题。在第二章介绍过数据的分布偏移对分类问题造成的影响。在本文中，我们是假设训练集和测试集的分布是大致一样的，从而利用位置信息，将训练集最优阈值映射到测试集的最优阈值。当分布发生偏移时，我们的方法将失效，此时需要研究新的映射形式。
3. 低AUC值情况。低AUC值，说明模型对样本的排序不是很好。此时可以尝试多段分类的方法来划分数据，即不是像之前那样简单地设定一个阈值，来划分数据，而是设定多个区间段，建位于某几个区间段的样本分为正类，其他的分为负类。然后再将训练集的区间段映射到测试集或未知数据集。

参考文献

[1].Chan, P.K. and S.J. Stolfo. *Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection*. in *KDD*. 1998.

[2].He, H. and E.A. Garcia, *Learning from imbalanced data.* Knowledge and Data Engineering, IEEE Transactions on, 2009. **21**(9): p. 1263-1284.

[3].Fernández, A., et al., *Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches.* Knowledge-based systems, 2013. **42**: p. 97-110.

[4]. Lin, M., K. Tang, and X. Yao, *Dynamic sampling approach to training neural networks for multiclass imbalance classification.* Neural Networks and Learning Systems, IEEE Transactions on, 2013. **24**(4): p. 647-660.

[5]. López, V., et al., *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics.* Information Sciences, 2013. **250**: p. 113-141.

[6]. Japkowicz, N. and S. Stephen, *The class imbalance problem: A systematic study.* Intelligent data analysis, 2002. **6**(5): p. 429-449.

[7]. Batista, G.E., R.C. Prati, and M.C. Monard, *A study of the behavior of several methods for balancing machine learning training data.* ACM Sigkdd Explorations Newsletter, 2004. **6**(1): p. 20-29.

[8]. Zhou, Z.-H. and M.-L. Zhang, *Ensembles of multi-instance learners*, in *Machine Learning: ECML 2003*. 2003, Springer. p. 492-502.

[9]. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002: p. 321-357.

[10].Liu, X.-Y., J. Wu, and Z.-H. Zhou, *Exploratory undersampling for class-imbalance learning.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2009. **39**(2): p. 539-550.

[11].Spackman, K.A. *Signal detection theory: Valuable tools for evaluating inductive learning*. in *Proceedings of the sixth international workshop on Machine learning*. 1989. Morgan Kaufmann Publishers Inc.

[12].Davis, J. and M. Goadrich. *The relationship between Precision-Recall and ROC curves*. in *Proceedings of the 23rd international conference on Machine learning*. 2006. ACM.

[13].Drummond, C. and R.C. Holte, *Cost curves: An improved method for visualizing classifier performance.* 2006.

[14].Kubat, M., R.C. Holte, and S. Matwin, *Machine learning for the detection of oil spills in satellite radar images.* Machine learning, 1998. **30**(2-3): p. 195-215.

[15].Cohen, G., et al., *Data imbalance in surveillance of nosocomial infections*, in *Medical Data Analysis*. 2003, Springer. p. 109-117.

[16].Phua, C., D. Alahakoon, and V. Lee, *Minority report in fraud detection: classification of skewed data.* ACM SIGKDD Explorations Newsletter, 2004. **6**(1): p. 50-59.

[17].Zheng, Z., X. Wu, and R. Srihari, *Feature selection for text categorization on imbalanced data.* ACM SIGKDD Explorations Newsletter, 2004. **6**(1): p. 80-89.

[18].Peng, Y., et al., *Cost-sensitive ensemble of support vector machines for effective detection of microcalcification in breast cancer diagnosis*, in *Fuzzy Systems and Knowledge Discovery*. 2005, Springer. p. 483-493.

[19].边婧, 彭新光, and 闫建红, *入侵检测大数据集代价敏感重平衡分类策略.* 小型微型计算机系统, 2012. **11**: p. 044.

[20].Sun, Y., A.K. Wong, and M.S. Kamel, *Classification of imbalanced data: A review.* International Journal of Pattern Recognition and Artificial Intelligence, 2009. **23**(04): p. 687-719.

[21].Alcala-Fdez, J., et al., *KEEL: a software tool to assess evolutionary algorithms for data mining problems.* Soft Computing, 2009. **13**(3): p. 307-318.

[22].Jo, T. and N. Japkowicz, *Class imbalances versus small disjuncts.* ACM Sigkdd Explorations Newsletter, 2004. **6**(1): p. 40-49.

[23].Weiss, G.M., *Mining with rare cases*, in *Data Mining and Knowledge Discovery Handbook*. 2005, Springer. p. 765-776.

[24].Japkowicz, N., *Concept-learning in the presence of between-class and within-class imbalances*, in *Advances in artificial intelligence*. 2001, Springer. p. 67-77.

[25].Weiss, G. *Timeweaver: A genetic algorithm for identifying predictive patterns in sequences of events*. in *Proceedings of the Genetic and Evolutionary Computation Conference*. 1999. Citeseer.

[26].Schapire, R.E. *A brief introduction to boosting*. in *Ijcai*. 1999.

[27].Chawla, N.V., et al., *SMOTEBoost: Improving prediction of the minority class in boosting*, in *Knowledge Discovery in Databases: PKDD 2003*. 2003, Springer. p. 107-119.

[28].Fan, W., et al. *AdaCost: misclassification cost-sensitive boosting*. in *ICML*. 1999. Citeseer.

[29].Prati, R.C., G.E. Batista, and M.C. Monard, *Class imbalances versus class overlapping: an analysis of a learning system behavior*, in *MICAI 2004: Advances in Artificial Intelligence*. 2004, Springer. p. 312-321.

[30].Raudys, S.J. and A.K. Jain, *Small sample size effects in statistical pattern recognition: recommendations for practitioners.* IEEE Transactions on Pattern Analysis & Machine Intelligence, 1991(3): p. 252-264.

[31].Weiss, G.M. and F. Provost, *Learning when training data are costly: the effect of class distribution on tree induction.* Journal of Artificial Intelligence Research, 2003: p. 315-354.

[32].García, V., R.A. Mollineda, and J.S. Sánchez, *On the k-NN performance in a challenging scenario of imbalance and overlapping.* Pattern Analysis and Applications, 2008. **11**(3-4): p. 269-280.

[33].Denil, M. and T. Trappenberg, *Overlap versus imbalance*, in *Advances in Artificial Intelligence*. 2010, Springer. p. 220-231.

[34].Martín-Félez, R. and R.A. Mollineda, *On the suitability of combining feature selection and resampling to manage data complexity*, in *Current Topics in Artificial Intelligence*. 2009, Springer. p. 141-150.

[35].Brodley, C.E. and M.A. Friedl, *Identifying mislabeled training data.* Journal of Artificial Intelligence Research, 1999: p. 131-167.

[36].Weiss, G.M., *Mining with rarity: a unifying framework.* ACM SIGKDD Explorations Newsletter, 2004. **6**(1): p. 7-19.

[37].Batuwita, R. and V. Palade, *FSVM-CIL: fuzzy support vector machines for class imbalance learning.* Fuzzy Systems, IEEE Transactions on, 2010. **18**(3): p. 558-571.

[38].Seiffert, C., et al., *An empirical study of the classification performance of learners on imbalanced and noisy software quality data.* Information Sciences, 2014. **259**: p. 571-595.

[39].Khoshgoftaar, T.M., J. Van Hulse, and A. Napolitano, *Comparing boosting and bagging techniques with noisy and imbalanced data.* Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2011. **41**(3): p. 552-568.

[40].Quionero-Candela, J., et al., *Dataset shift in machine learning*. 2009: The MIT Press.

[41].Shimodaira, H., *Improving predictive inference under covariate shift by weighting the log-likelihood function.* Journal of statistical planning and inference, 2000. **90**(2): p. 227-244.

[42].Weiss, G.M. and F. Provost, *The effect of class distribution on classifier learning: an empirical study.* Rutgers Univ, 2001.

[43].Laurikkala, J., *Improving identification of difficult small classes by balancing class distribution*. 2001: Springer.

[44].Estabrooks, A., T. Jo, and N. Japkowicz, *A multiple resampling method for learning from imbalanced data sets.* Computational intelligence, 2004. **20**(1): p. 18-36.

[45].Holte, R.C., L. Acker, and B.W. Porter. *Concept Learning and the Problem of Small Disjuncts*. in *IJCAI*. 1989. Citeseer.

[46].Drummond, C. and R.C. Holte. *C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling*. in *Workshop on learning from imbalanced datasets II*. 2003. Citeseer.

[47].Mease, D., A.J. Wyner, and A. Buja, *Boosted classification trees and class probability/quantile estimation.* The Journal of Machine Learning Research, 2007. **8**: p. 409-439.

[48].Mani, I. and I. Zhang. *kNN approach to unbalanced data distributions: a case study involving information extraction*. in *Proceedings of workshop on learning from imbalanced datasets*. 2003.

[49].Wang, B. and N. Japkowicz. *Imbalanced data set learning with synthetic samples*. in *Proc. IRIS Machine Learning Workshop*. 2004.

[50].Han, H., W.-Y. Wang, and B.-H. Mao, *Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning*, in *Advances in intelligent computing*. 2005, Springer. p. 878-887.

[51].He, H., et al. *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. 2008. IEEE.

[52].Bunkhumpornpat, C., K. Sinapiromsaran, and C. Lursinsap, *Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem*, in *Advances in knowledge discovery and data mining*. 2009, Springer. p. 475-482.

[53].Napierała, K., J. Stefanowski, and S. Wilk. *Learning from imbalanced data in presence of noisy and borderline examples*. in *Rough Sets and Current Trends in Computing*. 2010. Springer.

[54].Wilson, D.L., *Asymptotic properties of nearest neighbor rules using edited data.* Systems, Man and Cybernetics, IEEE Transactions on, 1972(3): p. 408-421.

[55].Tomek, I., *Two modifications of CNN.* IEEE Trans. Syst. Man Cybern., 1976. **6**: p. 769-772.

[56].Anand, A., et al., *An approach for classification of highly imbalanced data using weighting and undersampling.* Amino Acids, 2010. **39**(5): p. 1385-91.

[57].Yoon, K. and S. Kwek. *An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics*. in *Hybrid Intelligent Systems, 2005. HIS'05. Fifth International Conference on*. 2005. IEEE.

[58].林舒杨, et al., *不平衡数据的降采样方法研究.* 计算机研究与发展, 2011(S3): p. 47-53.

[59].Bunkhumpornpat, C., K. Sinapiromsaran, and C. Lursinsap, *DBSMOTE: density-based synthetic minority over-sampling technique.* Applied Intelligence, 2012. **36**(3): p. 664-684.

[60].Ling, C.X., et al. *Decision trees with minimal costs*. in *Proceedings of the twenty-first international conference on Machine learning*. 2004. ACM.

[61].Bradford, J.P., et al., *Pruning decision trees with misclassification costs*, in *Machine Learning: ECML-98*. 1998, Springer. p. 131-136.

[62].Turney, P.D., *Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm.* Journal of artificial intelligence research, 1995: p. 369-409.

[63].Domingos, P. *Metacost: A general method for making classifiers cost-sensitive*. in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. 1999. ACM.

[64].Zadrozny, B., J. Langford, and N. Abe. *Cost-sensitive learning by cost-proportionate example weighting*. in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. 2003. IEEE.

[65].Ting, K.M., *An instance-weighting method to induce cost-sensitive trees.* Knowledge and Data Engineering, IEEE Transactions on, 2002. **14**(3): p. 659-665.

[66].Kuncheva, L.I. and J.J. Rodríguez, *A weighted voting framework for classifiers ensembles.* Knowledge and Information Systems, 2014. **38**(2): p. 259-275.

[67].Galar, M., et al., *A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches.* Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2012. **42**(4): p. 463-484.

[68].Seiffert, C., et al., *RUSBoost: A hybrid approach to alleviating class imbalance.* Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2010. **40**(1): p. 185-197.

[69].Barandela, R., R.M. Valdovinos, and J.S. Sánchez, *New applications of ensembles of classifiers.* Pattern Analysis & Applications, 2003. **6**(3): p. 245-256.

[70].Guo, H. and H.L. Viktor, *Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach.* ACM SIGKDD Explorations Newsletter, 2004. **6**(1): p. 30-39.

[71].Fawcett, T., *ROC graphs: Notes and practical considerations for researchers.* Machine learning, 2004. **31**(1): p. 1-38.

[72].Fawcett, T., *An introduction to ROC analysis.* Pattern recognition letters, 2006. **27**(8): p. 861-874.

[73].Hanley, J.A. and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 1982. **143**(1): p. 29-36.

[74].Breiman, L., et al., *Classification and regression trees*. 1984: CRC press.

[75].Hanley, J. and B. McNeal, *A simple generalization of the area under the ROC curve to multiple class classification problems.* Radiology, 1982. **143**: p. 29-36.

[76].Bunescu, R., et al., *Comparative experiments on learning information extractors for proteins and their interactions.* Artificial intelligence in medicine, 2005. **33**(2): p. 139-155.

[77].Davis, J., et al. *View Learning for Statistical Relational Learning: With an Application to Mammography*. in *IJCAI*. 2005. Citeseer.

[78]Provost, F. and P. Domingos, *Well-trained PETs: Improving probability estimation trees.* 2000.

[79].Abe, N., B. Zadrozny, and J. Langford. *An iterative method for multi-class cost-sensitive learning*. in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004. ACM.

[80].Sun, Y., M.S. Kamel, and Y. Wang. *Boosting for learning multiple classes with imbalanced class distribution*. in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. 2006. IEEE.

[81].Provost, F. and T. Fawcett, *Robust classification for imprecise environments.* Machine learning, 2001. **42**(3): p. 203-231.

[82].Cortes, C. and M. Mohri, *AUC optimization vs. error rate minimization.* Advances in neural information processing systems, 2004. **16**(16): p. 313-320.

[83].Liao, L. and W.S. Noble, *Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships.* Journal of computational biology, 2003. **10**(6): p. 857-868.

[84].Dong, Q.-w., X.-l. Wang, and L. Lin, *Application of latent semantic analysis to protein remote homology detection.* Bioinformatics, 2006. **22**(3): p. 285-290.

[85].Saigo, H., et al., *Protein homology detection using string alignment kernels.* Bioinformatics, 2004. **20**(11): p. 1682-1689.

[86].Zou, Q., et al., *BinMemPredict: a web server and software for predicting membrane protein types.* Current Proteomics, 2013. **10**(1): p. 2-9.

[87].Cai, C., et al., *SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence.* Nucleic acids research, 2003. **31**(13): p. 3692-3697.

[88].郭颖婕, *基于随机森林的植物抗性基因识别方法研究*. 2011, 哈尔滨工业大学.

攻读硕士学位期间发表的学术论文

[1] 谢思发,林琛,苏旋,江弋.《Hadoop平台下热点事件检测》[J]小型微型计算机系统. Vol.35 No.4,2014.

[2] Si-fa Xie, Wei Weng, Ke Chen, Xiangrong Liu. Identifying Reliable Posts And Users In Online Social Networks.International.[J]Journal of Pattern Recognition and Artificial Intelligence (IJPRAI),Vol.28 No.6, 2014.

[3] Si-fa Xie, Pei Li, Yi Jiang, Y Zhao. A Discriminative Method For Protein Remote Homology Detection Based On N-Gram.[J]Genetics and Molecular Research(GMR).2015.

[4] Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, Ying Ju. Finding the Best Classifiaction Thresholdin Imbalanced Classification.[J]Big Data Research.2016.

致 谢

首先感谢我的导师邹权老师，当初在我困惑是否要读研时，是邹权老师和林琛老师力劝我读研，从而能有我现在的成绩。三年研究生期间，邹老师对我的学习、工作、生活各方面一直给予细心的指导，同时给了我宽松的学习成长环境。曾多次资助我参加学术会议，熬夜帮我修改论文，帮我联系公司实习。同时他的勤奋进取精神，也深深地影响着我，是我以后工作学习的一笔财富。在此向邹权老师致以我最诚挚的敬意和衷心的感谢！

同时，感谢吴梅红老师和吴云峰老师。他们在百忙之中仍拨冗对我进行指导，为我修改论文。感谢李旭斌师兄，陈文强师兄，唐振坤师兄曾经在学业上的帮助。感谢同门的王菁菁，李金金，李培以及官新均等师弟师妹，在生活学习上的互帮互助，让我感受到了集体的温暖，我不会忘记我们一起熬夜奋战的日子。感谢舍友魏陈强，杨志伟，7年的舍友之情是彼时的一种缘分，也将在以后成为一种羁绊。感谢所有关心、帮助和支持过我的同学、朋友们。感谢厦门大学给了我学习、深造的机会，感谢各位老师在学习期间给我的指导。

最后，我要深深地感谢我的父母、家人，感谢他们一直以来对我学业上的支持、精神上的鼓励、生活上的关心，他们对我无私的爱一直是我最坚强的后盾、不断进取的动力源泉！