**哈尔滨工业大学**

**硕士学位论文开题报告**

**题 目：不平衡样本分类方法在miRNA靶基因预测中应用**

**院 （系） 计算机科学与技术**

**学 科 /专 业 计算机技术**

**导 师 张春慨**

**研 究 生 郭建威**

**学 号 15S151492**

**开题报告日期 2016年9月22日**

**深圳研究生院制**

**二〇一六年九月**

目 录

[1 课题的来源及研究的目的和意义 1](#_Toc462264849)

[1.1课题来源 1](#_Toc462264850)

[1.2研究的目的和意义 1](#_Toc462264851)

[2 国内外研究现状及分析 2](#_Toc462264852)

[2.1国外内研究现状 2](#_Toc462264853)

[2.1.1 miRNA靶基因预测方法 2](#_Toc462264854)

[2.1.2不平衡样本分类的数据层面方法 3](#_Toc462264855)

[2.1.3不平衡样本分类的算法层面方法 4](#_Toc462264856)

[2.2国内外文献综述及简析 5](#_Toc462264857)

[3 主要研究内容及研究方案 5](#_Toc462264858)

[3.1研究内容 5](#_Toc462264859)

[3.1.1miRNA不平衡样本的预测研究 5](#_Toc462264860)

[3.1.2数据层面解决不平衡样本问题 6](#_Toc462264861)

[3.1.3算法层面解决不平衡样本问题 6](#_Toc462264862)

[3.2研究方案 6](#_Toc462264863)

[4 预期目标 7](#_Toc462264864)

[4.1预期目标 7](#_Toc462264865)

[5 已完成的研究工作及进度安排 7](#_Toc462264866)

[5.1已完成的研究工作 7](#_Toc462264867)

[5.2进度安排 7](#_Toc462264868)

[6 已具备的研究条件和所需条件及经费 8](#_Toc462264869)

[6.1实验室条件和经费保障 8](#_Toc462264870)

[6.2所需条件及经费 8](#_Toc462264871)

[7 预计困难及解决方案 8](#_Toc462264872)

[7.1预计困难与技术难点 8](#_Toc462264873)

[7.2解决方案 8](#_Toc462264874)

[参考文献 9](#_Toc462264875)

# 1 课题的来源及研究的目的和意义

## 1.1课题来源

通过查阅与专业相关的资料及文献，根据与导师讨论和自己所学的专业知识确定论文题目。

## 1.2研究的目的和意义

miRNA是一类长约19~25个核苷的内源性单链非编码小分子RNA。miRNA通过可以通过miRNA通过与靶基因完全或者不完全碱基互补匹配的方式，裂解靶基因或抑制其蛋白质表达水平，在后转录时期发挥其基因调控功能,是目前被人们认为最重要的调节基因之一[1]。为了更好地理解miRNA,一个重要的任务就是预测和识别miRNA的靶基因。人类基因组测序结果表明，miRNA基因仅仅占人类生命基因总数的1%左右，每个miRNA基因，生命体大约存在200个靶基因，miRNA调控着人体30%左右的基因[2]，发现的miRNA靶基因只是生命体靶基因集中的一小部分，迫切需要高效的靶基因识别方法。

目前生物学试验方法和计算机生物信息学算法是预测和鉴定miRNA靶基因的主要方法。生物实验方法成本高、耗时长，对实验条件和设备以及实验人员的技能有很高的要求；但生物信息学方法没有这些限制且吞吐量大，越来越多的研究者，利用生物信息学方法解决miRNA靶基因预测问题。生物信息学方法基于机器学习的miRNA靶基因预测算法包括 TargetScan[3]，Pictar[4] RNA22[5]，miTarget [6]等。在[6,7,8,9]的许多研究工作中都存在不平衡样本问题，[8，9]对miRNA靶基因预测中指出了的不平衡样本问题对预测结果的影响，并进行了处理。

不平衡数据集通常是指样本数据集的类别分布表现出不平衡的问题，数据集的不平衡问题会影响分类器的性能。当一些传统的方法面对不平衡问题时，往往不能取得很好的效果[10]，因为现有的成熟的分类器设计都是基于类分布大致均衡这一假设，用于训练的数据集是大致平衡的，以提高数据集的总体分类准确率为目标，忽略了少数类样本的重要性。所以当这些分类器应用到不平衡数据集的分类问题时，对于多数类的有较高的识别率，但少数类的识别率却很低。不平衡样本是生物信息数据的一个常见的问题[11]，实际问题中也存在大量不平衡样本集的分类问题:信用卡欺诈行为检测，网络入侵检测，医学疾病诊断等。实际上不仅是类别数目的不平衡影响分类算法的结果，不同类的数据间的重合程度也会影响分类结果。而且数据集本身的概念复杂性、类重叠与类的不平衡分布的问题，使得不平衡数据的学习受到了广泛的关注和研究。

# 2 国内外研究现状及分析

## 2.1国外内研究现状

### 2.1.1 miRNA靶基因预测方法

随着miRNA知识以及靶基因数据的不断丰富，使用生物信息学算法预测miRNA靶基因日益引起关注。计算方法预测靶基因首先要识别相关特征，进一步的提出计算机模型来预测靶基因[12]。

TargetScan[4]通过在UTR序列上查找miRNA 2~8 nt 的完全匹配序列，再对匹配序列在允许G:U匹配前提下尽量延伸其匹配区间，停止于第一个错配出现，然后对剩余部分计算绑定能:

(2-1)

如果Z大于设定的阈值Zc，且这个基因在人、小鼠、大鼠三个物种的直系同源基因中均被判定为可能的靶基因，则最终即判定为靶基因。TargetScans[13] 基于TargetScan做出三点改进: 为了增加敏感度，种子匹配放宽为2~7 nt 完全匹配；不再考虑绑定能；将鸡和狗加入到直系同源基因判断，更强调遗传保守性。miRanda[14]同样是对序列比对的匹配打分，并考虑自由能和遗传保守性。PITA[15] 提出了热力学模型，使用接入能和种子匹配规则判断miRNA的靶基因。DIANA-micorT[16] 使用38nt的滑动窗口在3’UTR上寻找可能的绑定位点，使用动态规划算法计算出窗口中两个序列的绑定能，使用特定的绑定规则进行过滤，得出满足所有规则的miRNA:UTR即为miRNA:Target对。RNAhybrid利用种子匹配规则找出所有可能绑定位点，并找出自由能最小的绑定结构，并计算其p-value,当MFE及p-value均小于设定的阈值时，被判定为miRNA:Target对。MovingTargets和Nucleus都是利用绑定能和种子匹配规则识别果蝇miRNA靶标的算法。

Miranda等在2006年开发RNA22[17]算法首先获得miRNA数据的公共模式，对数据集按照马尔科夫模型估计模序的统计显著性，保留有显著意义的公共模式，最后包含足够数量的公共模式的UTR被看做一个“靶位点岛”，高于自由能阈值的“靶位点岛”被认为是对应miRNA的靶标。

MiTarget[7]是Kim等2006年开发的第一个基于支持向量机构建的miRNA靶基因预测算法。算法引入基于位置的特征，将支持向量机和径向基函数相结合构建miRNA靶基因分类器。mirTarget[18]是2008年发表的有一个基于SVM的靶基因预测算法，Wang利用miRNA过量表达实验的microarray数据中提取训练靶标，提取5类131个特征并且考虑3’UTR存在多个2~7完全种子匹配的情况，训练SVM得到预测模型。[12]利用CLIPL测序数据作为训练集特定位点的碱基作为新的特征改进了mirTarget算法。

SVMicro[9]是基于序列分析的Site-SVM和UTR- SVM两级结构的miRNA靶基因预测算法。site-SVM模块基于miRNA、可能绑定位点、UTR序列等提取113个绑定位点特征，用SVM训练模型对潜在绑定位点进行打分， UTR-SVM模块基于Site-SVM模块输出和UTR序列提取30个UTR特征，构建UTR-SVM分类器实现miRNA靶基因预测。[5]中针对负类样本存在难分的异常样本，对miRNA靶基因分类器造成影响，改进AdaBoost样本权重调整机制,定义同类样本覆盖集：

(2-2)

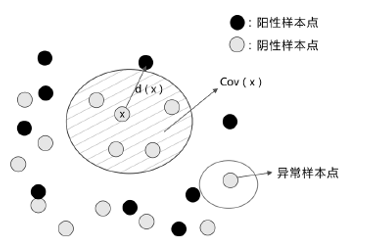
其中D(xi表示x与xi两个样本点之间的欧式距离，表示xi与所有正类样本之间的最小距离。样本的覆盖集Cov（x）的几何意义如图：

图 2-1 覆盖集示意图

覆盖集内包含极少的同类样本点的负类样本点为异常样本，通过减少异常样本的权重，对覆盖集的负类样本的权重进行平滑。

### 2.1.2不平衡样本分类的数据层面方法

数据层面的解决方案试图通过“采样”对不平衡样本集的样本分布进行重构，以使其从不平衡状态转化为均衡分布。采样方法，可以被归纳为“过采样方法”和“欠采样方法”两大类。过采样方法的思想是通过增加少数类样本的数量，欠采样的方法移除一些多数类样本，使多数类与少数类达到平衡。

Tomek等人在1976年提出了Tomek链，这是一种基于样本属性间隔的欠采样策略[19]。此方法的基本思想是递归删除Tomek链上的多数类的样本，直到不平衡样本集转化为平衡状态，或者符合了采样率的要求。基于Tomek链中样本向量之间距离差值的思想，Chawla等人于2002年结合其研究工作，发表了新的SMOTE算法（synthetic minority over-sampling technique[20] ）。SMOTE算法为每个稀有类样本随机选出几个邻近样本, 并且在该样本与这些邻近的样本的连线上随机取点, 生成无重复的新的稀有类样本。还有一些算法也是借鉴了SMOTE的思想或对其进行改进的，例如Borderline-SMOTE[21]和ADASYN[22]等,

Sáez J A等人 在2014年提出了SMOTE–IPF[23]来解决噪声数据和边界样本问题，是基于SMOTE方法的一种改进算法，先使用SMOTE增加负类样本的数量，在迭代的删除噪声点和边界点，SMOTE-IPF在许多情况下能取得较好的结果。 然而，SMOTE仿造生成样本的有效性与可信度一直被部分研究者所诟病。[24] 在2015年对类内不平衡问题进行研究，提出了改进的基于密度的聚类方法先对少数类聚类，在根据不同的过采样率对少数类簇过采样，从而有效的解决了类内不平衡的问题。

### 2.1.3不平衡样本分类的算法层面方法

算法方面地研究大致可以分为代价敏感学习、集成学习以及其他的一些学习方法。

Osuna等人[25]于1997年首次在支持向量机中利用两个不同的正规化参数(C+和C-)来分别控制两类样本的惩罚，由于样本的不平衡容易导致SVM算法对多数类样本的过学习，多数类比较多而导致边界支持向量在正负类样本的边界更加偏向于少数类，即牺牲少数类的正确率提高整体的正确率，代价敏感型的SVM通过调价正规化参数来预防SVM的过学习，[26]将代价敏感思想与Adaboost结合，在权重更新时，对误分的正负类样本采用不同的权重更新策略。

许多研究使用采样方法与集成学习方法结合，例如SMOTEBoost[27]，SMOTEBoost是将SMOTE和BOOST相结合，每次迭代过程中，都是用SMOTE方法合成一些少数类，使用加入了合成少数类的训练集训练基分类器。EasyBOOST[28]与SMOTEBoost相反，将Adaboost和欠采样技术结合起来，在Adaboost算法迭代的每一轮，都对多数类样本进行随机欠采样，抽取的多数类与少数类相同，将这些多数类与样本与所有少数类样本一起组成平衡的数据集训练基分类器。PCboost[29]在每次迭代初始，利用随机采样方式的数据合成算法，添加合成少数类样例，平衡训练信息，删除错分的合成样例。[32]对Adaboost提出改进，过采样样本的权重更新采取不同策略，并提出了使用粒子群优化方法对Adaboost的基分类器的权重系数进行优化。

## 2.2国内外文献综述及简析

基于规则的miRNA靶基因预测算法，根据已知miRNA：mRNA相互作用的生物学信息，从miRNA与靶位点互补性，miRNA：mRNA双链的稳定性以及靶基因位点的保守性等方面设定识别规则。根据已知发现miRNA和靶基因之间的信息而提出的靶基因识别规则，可能只是靶基因识别规则的一部分，导致不能很好的诠释miRNA：mRNA作用机制，从而造成预测准确率低、特异性差和假阳性率高。此外实验还证明了并不是所有的靶基因都具有保守性。

基于机器学习的方法从训练数据集中选取一定数量miRNA:mRNA的特征训练机器学习模型去预测miRNA:Target对。基于机器学习的miRNA靶基因预测算法突破了物种间保守性，如miTarget等引入了机器学习方法来提取特征参数，以弥补保守性这个限制而丢失了的靶基因，从而提高靶基因预测的准确率、降低假阳性率，是计算方法的研究重点。而对于不平衡样本的研究，从数据层面上，欠采样移除了训练集数据，可能造成部分重要信息地遗失，SMOTE增加了少数类样本的数量，并且一定程度上避免了随机过采样的过拟合问题[6]，而Borderline-SMOTE强调了边界样本的重要性。基于数据层面的解决方案但无论过采样还是欠采样都有一定的缺点。代价敏感类学习是通过“少数类样本的错分代价远远大于多数类样本错分”和“最小化条件风险”这两条基本原则，来使得分类器算法在进行分类的过程中，有意识地向少数类样本方向进行偏斜，代价矩阵的建立往往仅是凭借于经验主义，少有具体的科学依据。集成学习框架的优势是将会把分治之后所有分类器边界进行结合，以形成单一分类器算法难以准确得出的复杂边界，集成学习方法常常和欠采样方法结合从而既能改变样本的分布，又能发挥了集成学习方法泛化能力好的优点。

# 3 主要研究内容及研究方案

## 3.1研究内容

### 3.1.1miRNA不平衡样本的预测研究

miRNA靶基因预测问题的正类样本来自于生物实验验证的miRNA和靶基因的杂合体，虽然对于miRNA靶基因识别机制尚未了解清楚，通过对已发现的miRNA:target对分析，研究者发现了一些较明显的统计规律[9]。这些统计学特征包括：绑定点位序列有明显的遗传保守性，即绑定点位序列往往会存在于不同物种间同源基因的相同位置；绑定区内特征，miRNA：Site绑定后形成的杂合体的紧密程度是miRNA与mRNA能否绑定的重要参考，这一绑定的紧密程度可以由“绑定能”这一概念来表征；绑定区上下文特征；绑定点的开放性 ，mRNA可以在自身化学键作用下，通过自身回折形成碱基配对的复杂二级结构，从而影响其开放性；绑序列定点位置特征 ，偏好绑定在3’UTR的两端；多绑定点特征，如果存在多个绑定点，会增加3'UTR与miRNA的绑定机率。通过这些统计学规律我们可以构造出基于机器学习的miRNA靶基因预测方法，而在miRNA靶基因预测中的训练样本出现不平衡样本问题，又可以根据样本的分布特点，通过分析miRNA及靶基因特征，使用不同的不平衡样本分类方法解决靶基因预测问题，降低其假阳性率。miRNA靶基因预测算法通常使用热力学特征、结构特征和基于位置的特征等来训练样本，如何根据样本的特点发现数据新的特征也是需要研究的方面。

### 3.1.2数据层面解决不平衡样本问题

数据层面研究，由于不平衡样本存在正负类样本边界重叠、噪音数据和类内不平衡问题，所以从数据层面解决，主要是解决数据的分布问题，强调边界样本研究的重要性，寻求方法解决类内不平衡问题。边界重合对不平衡样本的分类结果影响很大，有研究表明当正负类样本边界重合时，少数类样本的比例和在边界的分布对分类的正确率影响很大，所以考虑对数据集的重采样对少数类在边界的分布调整使得分类效果最佳；另外当数据类内不平衡问题时少数类由于存在数据碎片，导致分类器难以找到少数类的全部边界。对于这些问题，应该考虑研究合适的采样方法以期真实的还原少数类样本的分布情况。

### 3.1.3算法层面解决不平衡样本问题

算法层面研究，主要从对不平衡样本的分类结果上出发，由于对不平衡样本分类结果好坏的衡量标准不再是准确率，而是综合考虑了正负样本的分类错误的重要性，本质是对不同样本的分类错误的重视程度不一样，代价敏感型分类算法是对正负样本分类采取不同的惩罚，从而使算法更加注重少数类的分类正确性，代价敏感型的分类算法主要是研究代价矩阵的取值。集成算法可以把几个弱分类器集成起来，成为一个强分类器，其中adaboost也是强调误分样本的重要性，对错分样本的权重进行调整，所以对集成分类器的研究结合少数类样本的特点提高少数类的分类正确率，提高其泛化能力。Smote-Boost结合SMOTE和集成学习算法，集成分类算法的研究应该注重和结合采样技术，实际上是改变样本的分布并且集成。

## 3.2研究方案

分析不平衡样本数据分布特点，对数据分布特征的研究是前期研究的重点，如何对数据进行预处理，需要对数据源进行详细的分析，特别是在miRNA靶基因的预测当中，由于并无现有的包含相关特征的训练数据集，所以如何收集miRNA靶标数据，mRNA基因数据，如何提取特征。然后构建实验框架，以确保后续的研究效果能够得到验证和对比。从数据层面针对影响不平衡样本分类结果的数据分布特点，研究采样方法以改进现有采样方法的弊端，重点研究边界数据和类内不平衡分布时，采样方法对数据分布的改变和对最终分类效果的影响，参考现有研究SMOTE的思路或者是聚类和采样技术结合的思想提出改进的方法以模拟少数类的分布情况。算法层面主要研究集成算法，由于集成算法在分类问题的优势是把弱分类器集成强分类器，并提高分类的正确率，对于不平衡样本主要是提高少数类的分类正确率，同时不过多的牺牲多数类的正确率。具体的是探讨集成算法的基分类器的投票权重设置策略和样本权重的分配策略。结合采样方法和集成方法来解决不平衡问题也是一个很好的思路，通过样本分布的改变强调少数类，并且利用集成算法良好的分类表现以期取得更好的效果。对于miRNA的靶基因预测，应用现有的算法来降低结果的假阳率，或者是从序列数据特点出发，分析发现对于预测有意义的特征以提高预测正确率。

# 4 预期目标

## 4.1预期目标

实现一个在传统的不平衡样本分类算法基础上改进的方法。使用了新的处理方法之后，使miRNA靶基因预测效果较使用之前有所改善。

# 5 已完成的研究工作及进度安排

## 5.1已完成的研究工作

已经完成了有关miRNA靶基因预测相关论文的阅读，基本掌握了相关背景知识，从相关数据库中下载了验证的miRNA靶基因对和miRNA测序数据，阅读了一些不平衡样本相关论问，基本了解了不平衡样本的问题。

## 5.2进度安排

2016年08月-2016年09月，阅读大量文献资料，对将要研究的项目做整体初步的了解。

2017年10月-2017年01月，对一些基本的分类方法进行分析与记录，建立基本的集成分类器。

2017年02月-2017年03月，提出数据处理的方案与分类算法。

2017年04月-2017年06月，完成所设计的算法框架。

2017年07月-2017年08月，对各个不平衡样本分类算法用不同的数据进行实验，对比实验效果。

2017年09月-2017年10月，撰写、修改论文。

2017年11月-2017年12月，研究总结，完成论文答辩。

# 6 已具备的研究条件和所需条件及经费

## 6.1实验室条件和经费保障

实验室有多台工作站、CPU和个人电脑；拥有多种编程语言开发环境。实验室经费充足。

## 6.2所需条件及经费

打印参考资料和购买相关参考书籍可能需要经费。

# 7 预计困难及解决方案

## 7.1预计困难与技术难点

课题进行需要涉及的生物信息学知识较多，涉及生物学实验数据特点的研究、miRNA及其靶基因序列的结合特点、将生物信息学特征进行向量化，在数据层面分析这些特征的分布，数据的理解等。

## 7.2解决方案

我会广泛查阅资料、全面深入地了解相关知识，设计比较完善的实验，在老师的指导、同学的帮助下，来克服困难。

# 参考文献

1. MicroRNAs: Genomics, biogenesis, mechanism, and function (Reprinted from Cell, vol 116, pg 281–297, 2004).
2. Min H, Yoon S. Got target? Computational methods for microRNA target prediction and their extension.[J]. Experimental & Molecular Medicine, 2010, 42(4):233-44.
3. Lewis B P, Shih I H, Jones-Rhoades M W, et al. Prediction of Mammalian MicroRNA Targets[J]. Cell, 2004, 115(7):págs. 787-798.
4. Grün D, Wang Y L, Langenberger D, et al. microRNA target predictions across seven Drosophila species and comparison to mammalian targets. [C]// Cell. 2010.
5. Huynh B T. et al (2006) A pattern-based method for the identification of microRNA-target sites and their corresponding RNA/RNA complexes
6. Kim S K, Nam J W, Rhee J K, et al. miTarget: microRNA target gene prediction using a support vector machine[J]. Bmc Bioinformatics, 2006, 7(18):1-12.
7. Ghoshal A, Shankar R, Bagchi S, et al. MicroRNA target prediction using thermodynamic and sequence curves.[J]. Bmc Genomics, 2015, 16(1):1-21.
8. 陈志茹. 基于SVM集成学习的miRNA靶基因预测研究[D]. 燕山大学, 2015.
9. 刘辉. miRNA靶标预测的系统生物学方法研究[D]. 中国矿业大学, 2009.
10. Japkowicz N, Stephen S. The Class Imbalance Problem: A Systematic Study[J]. Intelligent Data Analysis, 2002, 6(5):429-449.
11. Dittman D J, Khoshgoftaar T M, Wald R, et al. Comparison of data sampling approaches for imbalanced bioinformatics data[J]. Brain Research, 2014, 878(1-2):163-173.
12. Wang X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-Ligation studies.[J]. Bioinformatics, 2016, 32(9).
13. Lewis B P, Burge C B, Bartel D P. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets[J]. Cell, 2005, 120(1):15-20.
14. Enright A J, John B, Gaul U, et al. MicroRNA targets in Drosophila[J]. Genome Biology, 2003, 5(1):: R1.
15. Kertesz M, Iovino N, Unnerstall U, et al. The role of site accessibility in microRNA target recognition.[J]. Nature Genetics, 2007, 39(10):1278-84.
16. Kiriakidou M, Nelson P T, Kouranov A, et al. A combined computational-experimental approach predicts human microRNA targets.[J]. Genes & Development, 2004, 18(10):1165-78.
17. Huynh B T. et al (2006) A pattern-based method for the identification of microRNA-target sites and their corresponding RNA/RNA complexes[C]// Cell. 2010.
18. Wang X. miRDB: a microRNA target prediction and functional annotation database with a wiki interface.[J]. RNA, 2008, 14(6):1012-1017.
19. Tomek, I. (1976) Two Modifications of CNN. IEEE Transactions on Systems Man and Communications, 6, 769-772.
20. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1):321-357.
21. H. Han, W.Y. Wang, B.H. Mao, Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. The 2005 International Conference on Advances inIntelligent Computing (ICIC'05), 2005: 878-887.
22. Haibo He, Yang Bai, Garcia, E.A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[J]. 2008:1322-1328.
23. Sáez J A, Luengo J, Stefanowski J, et al. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering[J]. Information Sciences, 2015, 291(5):184-203.
24. 石凤兴.针对类内不平衡样本分类方法的研究.[D] 深圳：哈尔滨工业大学深圳研究生院.2015:12
25. Osuna E E. Support vector machines: Training and applications[C]// 1970:1308-16.
26. 陶新民. 不均衡数据SVM分类算法及其应用[M]. 黑龙江科学技术出版社, 2011.
27. N.V. Chawla, A. Lazarevic, L.O. Hall, et al., SMOTEBoost: Improving prediction of the minority class in boosting, The 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2003: 107-119.
28. Liu X Y, Wu J, Zhou Z H. Exploratory undersampling for class-imbalance learning.[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 2006, 39(2):539-50.
29. 李雄飞, 李军, 董元方,等. 一种新的不平衡数据学习算法PCBoost[J]. 计算机学报, 2012, 35(2):202-209.

**导师意见：**

**导师签字：**

**签字日期：**