硕士学位论文

基于支持向量机的高维不平衡数据二分类方法的研究

RESEARCH ON CLASSIFICATION METHOD OF HIGH-DIMENSIONAL CLASS-IMBALANCED DATA SETS BASE ON SVM

陆俊儒

哈尔滨工业大学

2016 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

基于支持向量机的高维不平衡数据二分类方法的研究

硕 士 研 究 生：陆俊儒
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：2016 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 CLASSIFICATION METHOD OF HIGH-DIMENSIONAL CLASS-IMBALANCED DATA SETS BASE ON SVM

Candidate: Junru Lu
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: December, 2016
Degree-Conferring-Institution: Harbin Institute of Technology

摘要

近年来，生物信息学、模式识别等不少领域中，出现了同时存在高维问题和不平衡问题的高维不平衡数据。高维问题是指由于数据集特征空间维数过高而存在着一些对分类效果产生负面影响的特征组合,导致分类效果不佳；不平衡问题是指在数据集里不同类别的样本在数量上分布不均匀，导致分类器对少数类关注不够，容易忽略它们蕴含的珍贵信息，从而影响分类效果。在高维不平衡数据中，高维问题和不平衡问题同时存在，互相影响，形成了新的问题。到目前为止，不少研究者单独针对高维问题和不平衡问题进行了研究，并提出了一系列成熟的算法。但是对同时展现出高维特性和不平衡特性的数据，绝大部分现有的方法都是先解决高维问题或不平衡问题，再解决另一个问题,并没有考虑到高维特性和不平衡特性相互影响而形成的新问题。

本文首先对高维问题和不平衡问题分别进行了介绍，并分析了高维问题和不平衡问题相互影响形成的新问题,通过阐述这些问题逐步展开分析研究。然后介绍了支持向量机 SVM,分析它在解决高维问题和不平衡问题中的优势,并对现有的一些解决高维问题和不平衡问题的算法进行总结,分析其优缺点。接着，改进 SVM-RFE 算法以便在考虑不平衡特性的情况下对高维数据进行特征选择,并基于 SVM 自动划分边界样本的特点改进 SMOTE 过采样算法以便在希尔伯特空间下进行边界过采样，同时对边界少数类样本的过采样倍率进行调节,提出了一种针对高维不平衡数据二分类的BRFE-PBKS-SVM算法。随后进行了一系列的实验，并且采用可以有效考察算法效果的多种指标来评价实验结果，证明了该算法的有效性。

关键词：高维不平衡；支持向量机；特征选择；边界样本；过采样

Abstract

In recent years, the problem of classification for high dimensional and class-imbalanced data is found in many fields like bioinformatics and so on. High dimensional problem result in bad classification results because of some combinations of features that influent negative of classification. Class-imbalanced problem means the number of samples of one class is more than another class, which would make the classifier concerns the majority class more but the minority less and influence the result of classification. The two problems are both exist in high dimensional and class-imbalanced data sets. They influence each other and produce a new problem. Many researchers make researches on high dimensional problem and class-imbalanced problem separately and come up with a series of algorithms. However, many people deal with the problems separately instead of consider the new problem when the two problems are both exist in a data set.

This article introduces the two problems and analysis the new problem produce by the influence of the two problems firstly. And then this article introduces SVM, analysis its advantages on dealing high dimensional problem and class-imbalanced problem and analysis the advantages and disadvantages of existed algorithms dealing with the two problems. Next, this article improves SVM-RFE so that the procedure of feature selection could also consider the class-imbalanced problem and improve SMOTE so that the procedure of over-sampling could do in Hilbert space by using the characteristic of dividing boundary automatically of SVM and the over-sampling rates are set adaptably meanwhile. Finally, a classification algorithm aimed at high dimensional and class-imbalanced data sets is come up in this article which named BRFE-PBKS-SVM: Border-Resampling Feature Elimination and PSO Border-Kernel-SMOTE SVM. And a series of experiments were made to prove the effectiveness of this algorithm by using different evaluation indexes.

Keywords: high-dimensional and class-imbalanced，SVM，feature-selection，boundary samples，over-sampling

目录

第 1 章 绪 论

在模式识别和机器学习中，分类是很常见的一项任务。分类的方法研究至今已取得了巨大的成功，比如支持向量机、决策树、神经网络等成熟的分类算法，在许多领域中都发挥了巨大的作用[1]。然而，传统分类算法在处理不平衡数据和高维数据时却遇到了很多问题。不平衡数据是指这样一种类型的数据集：在样本空间中，其中某一类样本在数量上与其他类别的样本存在显著的差异。由于不平衡数据的少数类蕴含着更多珍贵的信息，所以需要我们投入更多的关注，但传统分类器对这种数据的分类往往因为追求全局准确率而使得少数类被忽略，所以需要进行改进[2,3]。高维数据是指样本空间中样本的属性数目庞大的数据集类型，高维数据由于其属性数目过多，会使得训练过程耗时巨大[4]，有研究表明，在涉及向量计算的问题中，随着维数的增加，计算的耗时会以指数倍增加，这就是“维数灾难”；不仅如此，高维数据中还存在着一些冗余度、相关度高的特征组合，这些特征的存在，不仅不会使分类的效果有所提高[5]，还会产生过拟合等一系列问题[6]。

针对不平衡问题和高维问题，有不少学者做了研究并且提出了很多高效的算法。但是近年来，发现不少领域中的数据同时展现出高维度和不平衡两种特性，比如生物信息学[7,8]、图像分类[9]等，不少研究者开始针对高维不平衡数据这种新型数据类型进行深入的探究。

本课题主要研究同时展现出高维特性和不平衡特性的数据集的二分类问题：在考虑不平衡特性的情况下解决高维问题对分类效果带来的负面影响，同时解决不平衡问题和高维问题，并且针对支持向量机的特性，提出了一种基于 SVM 的针对高维不平衡数据分类任务的解决方案。本章首先介绍本课题的研究背景，然后通过介绍国内外的研究现状以及各种相关的经典算法，分析其优劣处以明确本课题的研究意义，最后本章将对本课题的框架进行描述。

1.1 研究背景与研究目的

在二类数据集中，总量较少的一类样本被称为少数类，而总量较多的一类称为多数类；多数类样本的数量与少数类样本的数量之比越大，样本空间的不平衡特性越明显，这就是数据的不平衡特性。在不平衡数据中，由于少数类的稀缺性，所以人们往往对其投入更多的关注，比如医学上的癌细胞检

测[10]、图像识别中的异常检测[11]等；但是一些经典的分类算法对全局准确率的追求使得其在数据展现出不平衡特性的情况下，对多数类的分类展现出了极好的分类效果，但是对少数类的分类效果却十分差，分类效果远远偏离人们的关注点。随着数据的不平衡特性越来越普遍地展现在人类的生活中，国内外学者开始对其投入了广泛的关注。

高维特性作为数据的另一种特性，在数据分类的过程中同样展现出了对分类效果的影响。高维数据是指样本空间中，样本的特征数量庞大，这种特性在样本数目较少的样本空间中表现的尤为明显。在对高维数据的分类过程中，由于特征数目庞大，训练的时间会呈指数级增长，即“维数灾难”；此外，高维数据还存在着特征之间冗余度高、特征之间相关度高的问题，即部分特征对分类效果没有影响，甚至反而会导致组合起来之后的特征使分类效果降低。随着高维数据在图像分类等领域中的普遍出现，高维数据的处理方法也日渐成熟，国内外学者研究出了一系列成熟的降维方法[12,13]和特征选择方法[14-17]。

然而，尽管针对不平衡数据分类方法和高维数据分类方法的研究都分别取得了不错的成果，但是当数据同时展现出不平衡特性和高维特性时，这些方法却有所局限。高维不平衡数据中，属性冗余、相关的问题和不平衡问题相互交织，形成了新的问题[18]。传统解决高维问题的方法是降维和特征选择，若先解决高维问题，由于现有的方法没有在降维或特征选择过程中考虑不平衡问题，有可能会使得新的特征空间中的特征更加不利于少数类的识别，甚至可能会使原始特征空间中的少数类在新特征空间中成为噪声点[19,20]；若先解决不平衡问题，极有可能会使新空间中的方差、样本均值等信息不能较好的反映原始空间中的样本空间的方差和均值关系，还可能会产生数据碎片等新的不平衡问题。例如传统的 SMOTE、Borderline-SMOTE 和各种针对多数类的欠采样方法等，它们算法虽然解决了数据的不平衡问题，但由于其利用在原始样本之间插值来形成新样本的方法使得少数类的密度增加了，导致方差信息有所变化，而且由于新生成的样本引入的是原始样本之间的相关性，所以新生成的样本不一定能还原原始样本空间中特征之间的关系[21]，而有些经典的降维方法如拉普拉斯映射值算法(LE)、局部保持投影算法(LPP)、主成分分析(PCA)等，这些算法追求尽量保留原始样本空间的分布特点[22]，主成分分析 PCA 是其中的典型代表，它是保留原始样本空间中最大方差分布的一种方法，但由于无监督降维方法不考虑类别标签，导致降维后的特征空间不一定有利于分类效果的提升，因此线性判别分析(LDA)等有监督算法又被用于高维数据的处理中，但都无法同时兼顾高维问题和不平衡问题。

综上所述，由于同时存在高维特性和不平衡特性的数据越来越多，其存在的问题也日渐突出，且传统单独针对不平衡问题和高维问题的方法并没有解决这两种特性同时存在所衍生出的新问题，因此需要一种理论方法针对性地处理该问题。所以本课题主要研究同时存在高维特性和不平衡特性的数据的二分类问题，以此为立足点对传统的不平衡数据分类方案和高维数据分类方案进行优化，提出一整种针对高维不平衡数据分类问题的解决方案。

1.2 国内外研究现状

在数据挖掘的分类任务中，目前针对高维不平衡数据的分类方法都是先解决高维问题或者不平衡问题，再解决另外一个问题，并没有考虑高维特性对不平衡数据分类带来的新问题和不平衡特性对高维数据分类造成的影响。

尽管目前针对高维不平衡数据分类的问题仍待继续探索，但是不少国内外学者分别针对不平衡数据和高维数据的分类进行了研究并且取得了不错的进展，并且在各自的应用领域中取得了显著的成效。

1.2.1 不平衡数据的处理

不平衡数据的分类任务主要从两个层面进行：数据层面的采样和算法层面的分类。

数据层面的采样方法是从样本空间中解决数据分布不平衡的重要手段之一，通过欠采样、重采样和混合采样等方法，对类别数目分布不平衡的样本空间进行重构，使原本分布不平衡的数据在数量上趋于平衡，减少数据不平衡对后期数据分类带来影响，防止分类器过多的关注多数类的分类准确率以追求全局准确率而忽略了人们更加关注的少数类的分类准确率[23]。大量实验研究表明，通过采样的方法，能显著提高不平衡数据的分类效果。采样方法发展至今，已经在不平衡样本分类领域中被广泛运用。

欠采样方法是指按照一定的规律删除某些样本，以使分类效果有所提升。1997 年 Kubat 等人提出了一种基于样本点之间的欧氏距离将样本点划分为不同的类型从而进行采样的方法：单边选择算法(one-side selection) [24]。其主要思想是观察与某样本点最近的 K 个样本点的类别，根据这 K 个样本的类别与该样本的类别的差异性，将该样本划分为安全样本、冗余样本、边界样本和噪声样本四种类型。其中安全样本和冗余样本在空间分布上是在它所在的簇较靠内的样本，即使它们是少数类样本，传统分类器对它们的识别程度也能达到较高水平；而边界样本和噪声样本由于其所处位置在空间上多种类别混杂，被称为“不安全样本”，它们往往需要分类器投入更多的关注。单边选择

算法根据样本的空间分布特点，将多数类中的“不安全样本”剔除，保留少数类的边界样本、冗余样本、安全样本，尽量使样本空间获得较好的可分性。

Chawla 等人提出的 SMOTE(synthetic minority over-sampling technique)[25]算法作为一种经典的过采样方法，已经被广泛的运用在不平衡数据的处理中，并且衍生出了不少基于 SMOTE 方法改进的过采样方法。SMOTE 算法的主要思想是在与某个少数类最邻近的 k 个少数类中随机选择一个，然后在这两个少数类的连线之间插值，生成一个仿造的少数类，其公式如下：

$$x^{new} = x_i + rand(0,1) \times (x_j - x_i) \tag{1-1}$$

SMOTE 算法虽然改变了多数类与少数类之间的不平衡比例，但由于其在两个真实少数类之间生成仿造的少数类，所以会改变原始样本空间的方差、协方差、类别密度等信息，对一些追求保留样本空间方差信息的降维方法有所限制，同时也会让 KNN 等基于原始样本空间数据分布特点来进行分类的方法效果大打折扣。但由于 SMOTE 生成的样本具有随机性，使得它能够避免对训练数据过拟合的问题，同时也更好地扩展了少数类的决策空间，不少过采样方法都基于 SMOTE 进行改进，比如 Han 等人提出的针对边界样本进行插值的 Borderline-SMOTE 方法[26]。

还有一类采样方法关注采样倍率的设置，SBC[27]是其中的典型算法。该算法认为样本空间的不同类簇，由于其空间分布不同，重要程度也有所差别，因此不能对同一类样本都设置相同的采样率，应该考虑他们所处的类簇在样本空间中的分布。基于该思想，SBC 算法将不平衡数据中的多数类聚成多个簇，然后按一定的规则设置每个多数类簇的欠采样比例，不同程度的减少每个多数类簇中的样本数目。

算法层面的分类方法在针对不平衡数据的处理方面，主要是寻找对少数类更加敏感的分类方法，使得分类器对少数类投入更多的关注度。近年来，在国内外学者的努力下代价敏感学习、集成学习等一些方法，经改进后，对不平衡数据的分类取得了较理想的效果。

代价敏感学习的核心是误分类代价[28]，针对不平衡数据的特性，代价敏感学习将少数类样本误分类的代价增大并且减小多数类误分的代价。在代价敏感学习中正确分类的代价为 0，而将少数类误分为多数类的惩罚代价大于将多数类误分为少数类的惩罚代价，以此来改变分类器的权值调整策略。代价敏感学习方法结合一些数据层面的处理方法使得不同的样本点误分类代价各不相同，这样针对不平衡数据分类的方法被证明是十分有效的。

集成学习方法则是构建多个基分类器，依据每一个分类器预测的准确度给它们赋予权重，采用加权投票的方法进行多数原则分类[29]。集成学习方法

由于其采用投票原则，最终分类结果不局限于单个分类器，所以能有较好的泛化能力，随机森林就是决策树类算法用于集成学习的一个典型，它不仅对特征进行随机抽样，还对样本进行随机抽样，使得每一个基分类器不局限于相同的样本空间。集成学习方法由于其对处理大规模数据的优良性能、针对不平衡样本分类的特有效果、对数据集的分治处理等，使它在不平衡数据分类领域中被广泛运用。

1.2.2 高维数据的处理

高维数据的处理主要有降维和和特征选择。

线性判别分析(Linear Discriminant Analysis)作为一种经典的有监督降维分类方法，早已在高维数据的处理中被广泛运用。LDA 追求降维之后不同类别之间的样本间隔尽量远、同一类别样本间隔尽量近，按照不同类别间的距离与相同类别间的距离之比最大的方向将原始样本空间进行投影映射[30]。LDA 方法在模式识别、图像处理中是一种被运用的较多的方法，当不同类别的数据之间可区分度较高、数据碎片、边界模糊的问题较少时，该方法能取得十分好的分类效果。但在类别总数是 C 种的情况下，由于其降维后的样本空间最多是 C-1 维，所以当高维数据中存在不平衡特性时，由于数据的特征空间被极度压缩，所以会出现少数类被多数类覆盖、不同类别的样本在降维之后有相同属性的问题[31]。无监督的降维方法不考虑类别信息，它追求在降维过程中，尽量还原原始样本空间中的某些特性。比如经典的 PCA(Principal Component Analysis)降维，就是一种按照原始特征空间中不同方向的方差分布大小来考虑投影方向的方法，使得降维后能尽量保留方差的分布。不少数据实验表明，即便样本空间中有成千上万的特征数，但是真正的方差能量，只用相对于原始特征数不到百分之十的投影方向就能保留大部分的方差能量。PCA 在处理类别信息基本遵循方差分布的数据时能有十分好的效果，比如图像分类等领域。但由于不考虑类别标签，在处理一些方差信息不能反映类别分布情况的数据时，往往会取得极坏的效果。流形学习方法(Manifold Learning)[32]自 2000 年被首次提出以来，已成为信息科学领域的研究重点。其主要思想是：假设高维空间中的数据具有某种特殊的结构，在将高维数据映射到低维后，低维空间中的数据仍能尽量还原原始数据在高维空间中的本质结构特征。

目前的特征选择方法按照特征选择过程与分类器训练过程的关系可以分为过滤式特征选择、包裹式特征选择和嵌入式特征选择方法三大类。过滤式特征选择是一种模型训练与特征选择过程互不相关的特征选择方法，它首先

依据一定的规则进行特征选择，并不限定最终得到的特征子集适合用于训练哪些算法模型。1992 年 Kira 和 Rendell 等人提出的 Relief 方法[33]是一种著名的过滤式特征选择方法，该方法设计了一个“相关统计量”来度量特征的重要性。包裹式特征选择则相反，它将分类器对特征的评价作为特征选择的参考标准,最终选择出的特征子集是最有利提高某种算法模型分类效率的子集。Liu 和 Setiono 等人于 1997 年提出的 LVW(Las Vegas Wrapper)[34]是一种经典的包裹式特征选择方法，它在拉斯维加斯方法(Las Vegas method)框架下使用随机策略来搜索子集，并以最终分类器的误差作为特征子集的评价准则。嵌入式特征选择则是通过对算法模型训练之后，依据模型的某些参数指标，进行一次性特征选择的过程，在模型训练完成的同时，也做出了特征选择。Tibshirani 等人于 1996 年提出的 LASSO(Least Absolute Shrinkage Selection Operator)[35]方法是嵌入式特征选择方法的典型代表，它通过修改损失函数中正则化项的范数，将 L2 范数改为 L1 范数，使特征权重更容易获得“稀疏”解，即它求得的特征权重会有更少的非零分量，通过消除权值中为零的分量对应的特征，达到特征选择的目的。

Corinna Cortes 和 Capnik 等人于 1995 年提出的支持向量机(Support Vector Machine, SVM)[36]在高维数据分类的应用中有着诸多的优势，并且衍生出了不少基于 SVM 改进的特征提取方法,比如 SVM-BFE、SVM-RFE 等[37]，它们都属于包裹式特征选择方法。所以在本课题的研究中，主要研究基于支持向量机的高维不平衡数据的分类方法。

支持向量机迭代特征消除法 SVM-RFE 通过每一轮迭代寻找每种属性的权值，权值的大小代表着 SVM 对该特征的关注程度，通过不断消除特征权重相对较低的特征来达到选取最优特征组合的目的。支持向量机反向特征消除法 SVM-BFE 每次训练消除一个特征，保存将消除某个特征后效果最好的特征组合，继续代入下一轮训练。基于 SVM 的特征选择方法，由于它以分类为目的，消除一些对分类效果有负面影响的特征组合和一些冗余度、相关度较高的特征，从而寻找使分类效果最好的特征组合，在处理高维数据中取得了一系列不错的效果。

由于避免了无监督降维方法不考虑类别标签的问题和有监督降维方法的一些局限性，特征选择方法成为了处理高维数据的不错选择，但由于其容易陷入局部最优解，因此也吸引了许多学者的对其做进一步研究。本文主要研究特征选择方法处理高维不平衡问题。

1.2.3 高维不平衡数据的分类

高维不平衡数据作为同时展现出高维特性和不平衡特性的新兴数据类型，由于其逐渐在诸多领域内频繁出现，开始成为诸多学者的热门研究方向之一。目前针对高维不平衡数据的处理方法，主要有先解决不平衡问题再解决高维问题和先解决高维为题再解决不平衡问题两种，但其最终思想仍然是将高维不平衡数据分解成高维问题和不平衡问题分别解决。Rok Blagus 和 Lara Lusa 等人曾做过一项关于过 SMOTE 方法在高维环境下和低维环境下对不平衡数据分类效果的影响的研究[38]，大量实验结果表明，在低维环境下，通过 SMOTE 及其衍生方法进行过采样，能在保证解决不平衡问题的前提下取得不错的分类效果，然而当数据同时展现出高维特性时，SMOTE 及其衍生方法却不能使大部分分类器对少数类的分类效果有所改善，其原因正是由于数据的高维特性影响不平衡问题而导致的。这就意味着使用 SMOTE 及其衍生方法进行过采样来解决不平衡问题再对高维数据进行分类的方法行不通，而单纯的使用重采样方法在不平衡数据中又存在着采样倍率设置的问题，单纯的欠采样方法则存在着丢失珍贵样本的风险。

通过先处理高维再解决不平衡问题的手段，能够避免传统采样方法对高维数据效果不明显的问题。特征选择方法主要目的是寻找最优的特征组合，使分类效率有所提高。特征选择方法通常被用在先处理高维问题的方案中。针对高维不平衡数据的分类，不少学者的主要研究方向是先通过特征选择选取最优特征组合然后再进行采样等一系列方法解决不平衡问题。Richard Weber 等人在 2014 年研究了在使用不同的特征选择方法处理高维数据之后，用 SMOTE 方法进行采样，以考证不同特征选择方式在与 SMOTE 算法组合的情形下的分类效果，最后用支持向量机进行分类的针对高维不平衡数据的解决策略[39]。通过严密的实验设置，Richard Weber 等人展示了 SVM 解决高维不平衡数据分类问题的优越性，但由于采样在特征选择之后进行并且只对数据不平衡做一次性处理，所以不能保证每一次特征选择都是在考虑数据不平衡特性的情况下的最优选择，从而极易陷入局部最优解，并且没有考虑 SMOTE 算法及其衍生算法在高维空间中的局限性。

由于高维不平衡数据中的高维问题和不平衡问题相互交织，形成了新的问题，使得在处理高维问题的过程中不能一次性解决不平衡问题，同时在解决不平衡问题的过程中也不能忽略高维问题的存在。所以目前针对高维不平衡数据的分类方法，仍需改善。

1.3 主要研究内容

本课题主要研究基于 SVM 的高维不平衡数据的二分类问题，针对已有算法进行改进，提出了一种基于 SVM 的考虑数据不平衡特性的特征选择方法，先解决高维不平衡数据分类任务中的高维问题，再针对 SVM 的特性提出了一种边界过采样方法，解决高维不平衡数据分类任务中的不平衡问题，主要研究内容如下：

首先，通过 SVM 基本理论分析其在解决高维不平衡数据二分类任务中的优越性。

然后，针对传统的特征提取算法框架进行改进，在每一轮特征选择的过程中加入数据采样过程以增加分类器对少数类的关注程度，使得每一次特征选择都能在分类器重点关注少数类的前提下进行。

最后，改进传统 SMOTE 过采样方法，使得 SMOTE 过采样产生的样本点最合适提升 SVM 的分类效果，同时自适应的寻找最有利于提高分类效果的少数类过采样倍率。

1.4 章节结构安排

本文的章节结构安排如下：

（1）介绍本课题的背景和意义，并对本课题的国内外研究现状进行简单的介绍。

（2）高维不平衡数据二分类问题的基础研究。主要介绍了同时存在高维问题和不平衡问题的数据分类的难点，对造成高维不平衡数据分类难的问题进行了本质的研究与分析，同时对 SMOTE、SVM 等算法原理进行简单的介绍，并分析它们在高维不平衡数据二分类的运用中的优点和不足之处。最后，还介绍了评价高维不平衡数据分类效果的指标。

（3）提出对同时存在高维问题和不平衡问题的数据进行二分类的解决方案：提出一种基于 SVM 改进的特征选择方法，同时提出一种适用于 SVM 的边界采样算法，并从理论上论证这两种方法在解决高维不平衡数据二分类问题中的优良性能。

（4）实验验证。对以上提出的高维不平衡数据二分类的解决方案进行实验证，通过不同的试验指标对实验结果进行考察，证明所提出的算法有效性和优越性。

第 2 章 高维不平衡数据二分类方法基础研究

第一章简要介绍了现有的高维数据的分类方法和不平衡数据的分类方法：一系列的采样算法已经被证明能够很好的解决数据的不平衡问题，SMOTE 算法及其衍生方法是其中的典型代表；对于高维问题，如迭代特征选择、包裹式特征选择、嵌入式特征选择等一系列方法也早已投入到工程运用中。

SVM 作为二分类的经典算法，在高维数据的二分类任务中表现出了良好的性能，出现了许多基于 SVM 的特征选择方法，并且由于 SVM 自动划分边界样本的特性，使得 SVM 对不平衡样本的处理有着良好的可扩展性。

为了寻找能够有效解决二类高维不平衡数据分类问题的方法，本章将会研究以下内容：首先，分析高维不平衡数据分类难的本质原因；然后，论证 SVM 在解决高维不平衡数据二分类问题上的优越性和可提升空间，并总结了现有的一些解决不平数据分类问题的方法和基于 SVM 解决高维数据分类问题的方法；最后，由于评价分类效果的传统指标在用于评价高维不平衡数据的分类效率时有着不少弊端，所以将对此问题的评价方法展开讨论。

2.1 高维不平衡数据的本质

不平衡数据指在类别之间表现出了数量上不相等的数据分布，其中样本总数较少的一类称为少数类，样本总数相对较多的一类称为多数类。在不平衡数据中，人们往往更关注少数类，比如石油检漏数据，正常情况下的样本远远多于漏油情况的样本，不平衡比率有可能达到 100:1 甚至更高，而人们往往更关注的是异常的情况，因为如果有漏油点未被正确识别，将会造成极大的经济损失。

高维数据则是指在样本空间中，样本的特征数量庞大，有些数据集里样本的特征数目甚至超过样本容量。高维数据中往往会存在一些与类别信息无关甚至会降低现有特征组合反映类别信息能力的特征，例如在医学领域中，人们为了寻找某种病症对生命体的影响，往往会收集许多生理指标，而这些生理指标的变化不一定都是某病症引起的，所以准确寻找由该病症引起的生理指标波动，对判断患者是否患有某种疾病，有极大的意义，假如考虑了某种无关的生理指标，将会增加误诊的几率，即过拟合了训练数据。

高维不平衡数据是指同时存在高维特性和不平衡特性数据。在高维不平衡数据中，有些有利于识别多数类的特征组合不一定利于识别少数类，而由

于正常情况下分类器对多数类的关注程度高于少数类，所以在解决高维问题的过程中有可能使得不平衡问题变得更加尖锐。例如在广告点击率 CTR 预测的任务中，提取出来的特征总数可能十分庞大，并且有广告点击行为的正类样本往往十分稀少，而针对庞大的特征组合需要做特征选择，假如剔除了有利于识别少数类的特征组合，将会使整个 CTR 预测任务效果十分差。

2.2 高维不平衡数据分类困难概述

分析高维数据和不平衡数据的本质，考虑两种数据的处理过程，可以发现造成高维不平衡数据分类困难的主要原因有两个：传统解决数据分布不平衡问题的方法不能很好的还原高维数据中的特征之间的关系，导致采样过后的新的特征空间与采样前的原始特征空间差别较大；特征选择过程中由于没有考虑不平衡问题，导致所选择的特征不一定有利于少数类的识别。概括起来就是：高维问题影响采样过程、不平衡问题干扰特征选择。

2.2.1 高维问题影响不平衡问题的解决

由于高维问题的存在，使得传统采样方法无法改变分类器对多数类的倾重，从而使传统采样方法失去意义。文献[21]中的实验研究表明，SMOTE 方法虽然能在低维数据中让分类器增加对少数类的关注程度，但在高维数据中，效果却不明显，究其原因主要有三个点。

第一，虽然 SMOTE 方法没有改变少数类的特征均值，但却改变了少数类的方差分布。根据 SMOTE 在最邻近的 k 个点之间进行随机插值的过采样的方法，过采样之后的少数类均值几乎不变，但方差却有极大的改变，这点在高维实值数据中表现的尤为突出。这会导致样本空间中样本簇的密度不均匀，极有可能会产生数据碎片等问题。

第二，用 SMOTE 方法生成的少数类，会使新样本空间中引入样本之间的相关性，而不是特征之间的相关性。由于 SMOTE 方法利用邻近的两个不同的少数类生成新的少数类，所以新生成的少数类样本会与生成它的两个样本有所关联；而在高维数据中，特征之间往往会有所关联，即特征间存在相关性。SMOTE 方法生成的少数类引入了真实样本之间的相关性，但由于新生成的样本融合了两个真实样本的特征，所以这种特征相关性不一定能还原真实的少数类样本，而我们希望新生成的样本能忠实的还原原始特征空间中的相关性，以便于后续特征选择的过程所选择的特征一定是有利于提高分类效果的，例如对于冗余的特征，新生成的样本的该特征也应该有一定的冗余性，以便在特征选择的过程中将该特征消除，否则将会影响后续的特征选择

过程。

第三，SMOTE 方法改变测试集的样本与训练集的样本之间的欧氏距离。由于 SMOTE 方法通过在两个真实的少数类别之间线性插值生成少数类样本，以达到改变样本空间中数据分布不平衡的目的，因此必然会改变少数类的密度分布。这个问题会导致某些依据密度来进行分类的分类器失效，比如 KNN 分类器。KNN 分类器通过寻找测试样本点在训练集中前 k 个欧氏距离最近的样本点，依据多数表决原则，决定测试样本点的类别；一旦少数类的密度增大，极有可能会使 KNN 分类器失效。在高维空间中，当用欧式距离来衡量样本间的相似度时，测试样本将会更相似于 SMOTE 生成的仿造样本，而不是原始的数据样本。

2.2.2 不平衡问题干扰特征选择

目前的特征选择方法按其搜索方式可分为完全搜索、启发式搜索和随机搜索三大类，完全搜索主要有广度优先搜索(Breadth First Search, BFS)、分支限界搜索(Branch and Bound)、定向搜索(Beam Search)；启发式搜索主要有反向特征消除(Backward Algorithm for Feature Elimination, BFE)、迭代特征消除法(Recursive Feature Elimination, RFE)；随机搜索策略主要运用遗传算法、粒子群优化算法、模拟退火算法，对特征空间进行随机搜索，寻找最优特征组合。虽然已有许多成熟的特征选择方法用于解决高维问题，但当不平衡问题存在时，特征选择的过程却受到了明显的干扰：

由于没有考虑不平衡问题给特征选择带来的影响，在特征选择的过程中，极容易使得特征选择朝着不利于少数类识别的方向进行[40]：一次性完成特征选择的算法（如 LASSO 算法等）则可能直接剔除掉一些对少数类的识别有重要效果的特征组合；迭代消除特征的做法是反向特征消除法的改进，它通过考虑分类器自身的“感受”来进行特征选择，每一轮选择一个分类器判定为对最终结果贡献较低且能使最终结果提升最大的特征进行消除，但同样无法阻止特征选择过程朝着增加多数类识别率的方向进行。

2.3 支持向量机解决高维不平衡问题的优势

支持向量机(support vector machines, SVM)是一种追求间隔最大化的分类模型。SVM 通过引入核技巧来解决非线性分类问题，所以 SVM 的输入空间和训练空间有所不同。输入空间是指数据的原始空间，训练空间是指 SVM 学习数据分布规律时所在的空间，前者是欧几里得空间，后者一般是希尔伯特空间。支持向量机的学习策略就是间隔最大化，即学习一个分离超平面，

使得该超平面尽量分开两类样本，并且使所有样本点到超平面的距离之和最大[41]。

2.3.1 支持向量机基本理论

SVM 的公式如下：

$$f(x)=sign(w^{*}\cdot x+b^{*}) \tag{2-1}$$

其中分离超平面的公式为：

$$y=w^{*}\cdot x+b^{*} \tag{2-2}$$

则 SVM 的目标为求解如下问题：

$$\max_{w,b} \quad \gamma \tag{2-3}$$

$$s.t. \quad y_i\left(\frac{w}{\|w\|}\cdot x_i+\frac{b}{\|w\|}\right)\geq\gamma, \quad i=1,2,\cdots,N \tag{2-4}$$

式(2-3)和式(2-4)表示了 SVM 的求解目标：寻找一个最大的间隔 γ，使得所有样本点到分离超平面的距离都大于等于 γ。由于函数间隔的取值并不影响最优化问题的解，同时考虑到需要寻找一个凸连续可微函数以优化求解，而且现实中的许多数据集由于边界混杂使得它们非线性可分，因此为每个样本点加入一个松弛变量 $\xi_i\geq0$，并且为每个松弛变量支付一个代价 ξ_i，所以由式(2-3)和(2-4)得到下面的最优化问题：

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i \tag{2-5}$$

$$s.t. \quad y_i\left(w\cdot x_i+b\right)\geq1-\xi_i, \quad i=1,2,\cdots,N \tag{2-6}$$

$$\xi_i\geq0, \quad i=1,2,\cdots,N \tag{2-7}$$

其中 C 为惩罚参数，惩罚参数增大时误分类将付出更大的代价。最小化目标函数(2-5)有两层含义：使分隔超平面与分离超平面之间的距离尽量大，同时使所有样本点到分离超平面的距离尽量大于 1 与松弛变量的差值。为求解上述问题，首先构建拉格朗日函数，为第一个约束引进拉格朗日参数 a_i，i=1,2,…,N，为第二个约束引进拉格朗日参数 u，定义拉格朗日函数为：

$$L(w,b,\xi,a,u)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}a_iy_i(w\cdot x_i+b)+\sum_{i=1}^{N}a_i+\sum_{i=1}^{N}(C-a_i-u_i)\xi_i \tag{2-8}$$

由于每个拉格朗日参数 a_i 对应一个样本点(x_i,y_i)，而 KKT 条件[42]则约束了拉格朗日参数与样本点的对应关系，这一性质使得 SVM 可以通过拉格朗

日参数解释训练数据中样本点在特征空间中的几何位置：安全样本、误分样本、边界样本[43]。SVM 的这一特性对解决不平衡数据中的不平衡问题提供了极大的便利性。对拉格朗日函数求导得到 SVM 的解如下：

$$w^* = \sum_{i=1}^{N} a_i^* y_i x_i \tag{2-9}$$

$$b^* = y_j - \sum_{i=1}^{N} y_i a_i^* (x_i \cdot x_j) \tag{2-10}$$

由于现实数据中，不少数据是非线性数据，线性 SVM 不能很好的解决非线性问题，如图 2-1 所示。因此为解决非线性问题，Boser, Guyon 与 Vapnik 又在 SVM 中引入了核技巧[44]：将原始空间的数据映射到特征空间，这个特征空间称为希尔伯特空间，通常希尔伯特空间会比原始空间更高维。

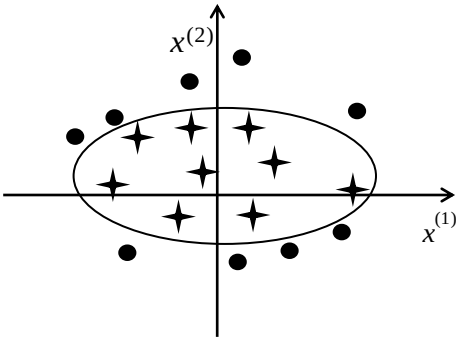

图 2-1 线性不可分数据

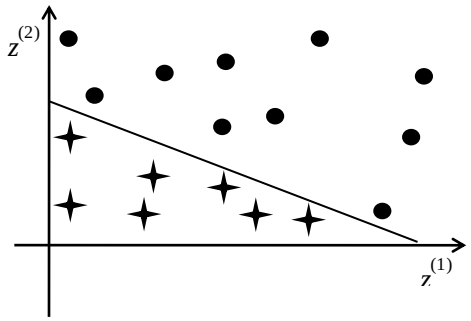

图 2-2 线性可分数据

设原空间为$X \in R^2$，$x = (x^{(1)}, x^{(2)}) \epsilon X$，可以看到图 2-1 中，两类数据的分布是一个非线性的分布。假设通过一个映射函数$z = \varphi(x)$可将原空间 X 映射到新空间 Z 中，则可以从图 2-2 中看到，原本的非线性分类问题转化为了线性分类问题，其中$Z \in R^2$，$z = (z^{(1)}, z^{(2)}) \epsilon Z$。然而通常需要找到这样一个映射函数$\varphi$是困难的，所以核技巧不显式地定义映射函数$\varphi$，而是通过核函数 $K(x,z)$ 隐式的将输入空间映射到希尔伯特空间：

$$K(x, z) = \varphi(x) \bullet \varphi(z) \tag{2-11}$$

因此将式(2-9)、(2-10)代入拉格朗日函数(2-13)，可得到求解极大值的问题为一个凸二次规划问题：

$$\min_{a} \quad \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} a_i \tag{2-12}$$

2.3.2 支持向量机对解决不平衡问题的优势

针对不平衡数据类别分布不均匀的问题，现有的方法一般是通过欠采样、

重采样、SMOTE 过采样等方法来解决。欠采样是依据一定规则，删除一些多数类样本，使两类样本在数量上达到平衡，由于无法事先确定样本的价值，所以欠采样有可能会使得一些珍贵的样本损失；而重采样则通过复制少数类样本，使得两类样本在数量上均衡，尽管它克服了欠采样方法的缺陷，但由于重采样方法一味的复制少数类，使其对噪声敏感，极有可能会使得模型过拟合训练集中的少数类分布的情况；SMOTE 过采样方法则依据一定的规则在两个少数类样本之间插值，SMOTE 算法则同时克服了欠采样方法的缺点和重采样方法的缺点，因此已被广泛运用到不平衡数据的处理中。

过采样、重采样的方法虽然可以解决不平数据分类中的样本稀缺的问题，但样本边界重叠作为不平衡数据分类的另一个难题，不能直接运用采样方法来解决，如图 2-3 所示，需要先将多数类和少数类分别聚簇以确定重叠的边界样本，再运用一系列的采样方法才能解决该问题[45]，而由于 SVM 具有自动划分训练集在特征空间中的边界的特点，所以对于解决边界样本重叠的问题有着天然的优势。

式(2-5)所示的几何间隔是 SVM 需要求解的最优化目标函数，它表示所有样本点到分离超平面的距离的最小值，而对于距离分离超平面的距离小于几何间隔的样本点，被称为支持向量，即训练集特征空间中的边界样本。考虑 KKT 对偶互补条件式，可以得到如下拉格朗日参数 a 与样本点在样本空间中的分布关系的结论[46]：

第一：若 $a_i=0$，样本点被正确分类却位于间隔平面外，属于安全样本。

第二：若 $0<a_i<C$，样本点被正确分类但落在间隔边界上，属于边界样本。

第三：若 $a_i=C$，样本点距离分离超平面的距离小于几何间隔，样本点落在两个间隔边界之间且有可能被误分类，属与边界样本。

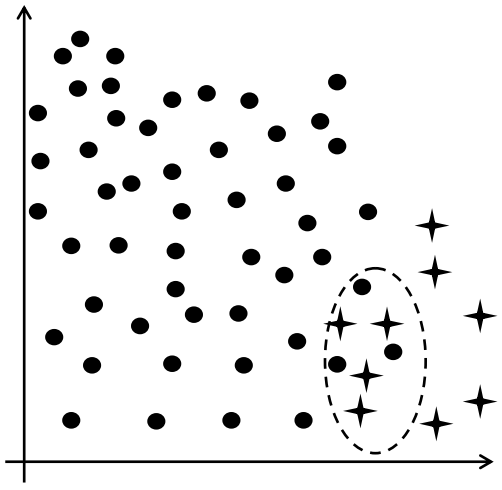

图 2-3 边界样本重叠

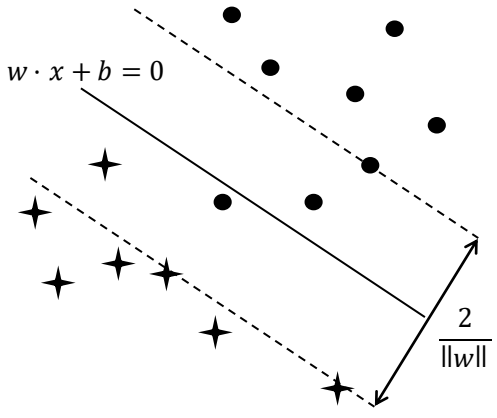

图 2-4 支持向量与边界样本

如图 2-4 所示，实线表示分离超平面，虚线表示间隔平面，两条虚线与实线的间隔一致，即几何间隔。两个间隔超平面之间的样本点即支持向量（边界样本）。

观察分离超平面的两个未知变量 w 和 b 的解，可以发现由于安全样本对应的 a 值为 0，所以虽然所有的训练样本都参与了 SVM 的求解，但最终对 w 和 b 的值有所贡献的只是边界样本，因此在支持向量机中，边界样本重叠问题是不平衡问题的主要问题。由于 SVM 的特性，训练集的原始特征空间是欧几里得空间，为求解最大几何间隔，如果引入了核函数进行非线性变换解决非线性问题，则欧几里得空间变换成希尔伯特空间，在希尔伯特空间下，难以使用传统的聚簇方法确定样本边界，但根据求解 SVM 得到的 a 值可以快速的判断 a_i 对应的样本点 (x_i,y_i) 是边界样本还是安全样本，因此十分有利于解决不平衡问题。

2.3.3 支持向量机对解决高维问题的优势

支持向量机以统计学习理论中的 VC 维理论为基础，追求结构风险最小化，这使得它在高维模式识别这一领域中被广泛运用。由 Vapnik 和 Chervonenkis 提出的 VC 维[47]是对 SVM 假设空间的表达能力的一种描述。VC 维表达了 SVM 能正确分类的样本总数与特征空间中样本特征总数的关系：对于特征总数为 n 的样本空间，其 VC 维等于 n+1，支持向量机总能把样本容量小于等于 VC 维的数据集的任意分布完全正确分类。例如 2 维欧几里得空间的 VC 维是 3，那么在数据为二类情况下，对于 2 维空间的 3 个不共线样本的任意分布组合（共 8 种），总存在一条直线能将它们正确分类，如图 2-5 所示。

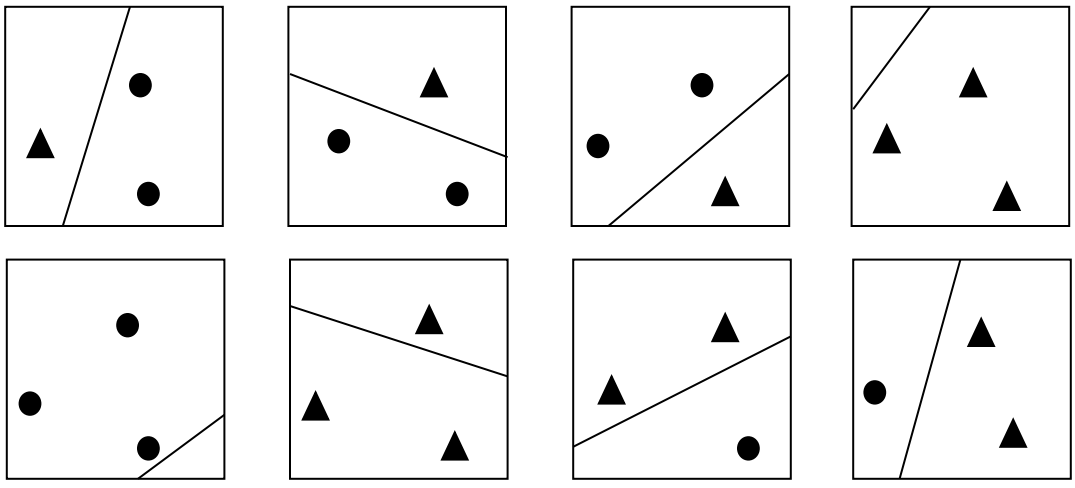

图 2-5 2 维欧氏空间下的 2^3 种二类样本组合方式

图中的两种不同标号分别表示两种不同的类别，该图说明在 VC 维等于 3 的情况下，SVM 最多能将 3 个不共线的二类样本在 2 维欧几里得空间中的 2^3 种组合分布完全正确分类；若样本点共线，则 2 维空间退化成 1 维，同样 VC 维也为 2。

因此特征空间的特征越多，其 VC 维就越高，SVM 的表达能力就越强；若特征空间中的特征数远远大于样本总数或者是一个无穷组合时，则有 VC(F)=∞。所以 SVM 在解决高维数据的分类问题、小样本数据的分类问题有着独特的优势。SVM 序列极小化方法[48]是利用该特性和追求结构风险最小化的一种特征选择方法，它的基本思想是将求解 SVM 分离超平面得到的向量 w 作为衡量样本特征重要性的标准，在经验损失不增加的情况下，每一步都去掉一个权向量 w 的分量对应的特征，使得去掉该特征后经验损失不增加，从而最大限度地减少特征数，该特征应该是 w 的绝对值尽量小的分量所对应的特征，其具体过程如表 2-1 所示。

由于 SVM 特征选择序列极小化算法采用贪婪策略的思路是删除特征权重较小的那一个特征，使得经验损失不增加，以此来保证它的泛化能力，因此所删除的特征往往连局部最优特征都不是。

表 2-1　SVM 特征选择序列极小化算法伪代码

输入：原始数据集
输出：经过特征选择后的数据集
(1) 求解 SVM 的分类超平面，得到 w 和 b，并计算出相应的经验损失 $R_{\text{emp}}[f]$;
(2) 重新排列权向量 $w=(w^1,w^2,\cdots,w^n)$ 的分量，使得：$\lvert w^{i1}\rvert \le \lvert w^{i2}\rvert \le \cdots \le \lvert w^{in}\rvert$
(3) 如果 n=1，则转(8)；否则，令 k=1，转(4)；
(4) 去掉原始训练集中与 w^{ik} 对应的特征 x^{ik}；
(5) 用去掉特征 x^{ik} 的数据集训练 SVM，得到新的 w 和 b，记为 w^{new}，b^{new}，并计算相应的经验损失 $R_{\text{emp}}[f^{\text{new}}]$；
(6) 如果 $R_{\text{emp}}[f^{\text{new}}] \leqslant R_{\text{emp}}[f]$，则置 $w=w^{new}$，$R_{\text{emp}}[f]= R_{\text{emp}}[f^{\text{new}}]$，$n=n$-1，转(2)；
(7) 如果 $k=n$，则转(8)；否则置 $k=k$+1，转(4)；
(8) 返回经过特征选择后的数据集。

除此之外，由于 SVM 在解决高维问题、小样本问题中的优越性，一系列基于 SVM 的包裹式特征选择方法也被运用到实际中，如 SVM-BFE 和 SVM-RFE 等。其中 SVM-BFE 参考的指标是去除某一个特征之后的全局分类准确率或经验损失等，它与 SVM 特征选择序列极小化方法不同的是，SVM 特征选择序列极小化方法是在保证每一轮删除的特征对应的 w 分量的绝对值

尽量小的情况小，使得删除掉特征之后经验损失不增加，而 SVM-BFE 则不考虑 w 分量的绝对值，单纯的追求删除的特征使经验损失减少程度$\nabla R_{emp}[f]$最多，从而导致删除掉的特征有可能是特征权值较大的特征，容易导致最终的特征组合对训练集拟合不够。SVM-RFE 是 SVM-BFE 的改进版本，它受 SVM 特征选择序列极小化算法的启发，考虑使用一个特征评价体系来对特征的重要程度进行排序，即 w 分量的绝对值，在特征选择的过程中参考特征的排序顺序进行迭代搜索，每一轮迭代首先考虑选择能使经验损失的减少程度$\nabla R_{emp}[f]$最多的特征，然后再考虑该特征在特征排序中的位置，当选择每一个特征造成的$\nabla R_{emp}[f]$都是相同时，尽量选择特征排序比较靠后的特征，以保正特征每一轮被删除掉的特征是分类器投入较少关注的特征。

2.4 评价标准

高维不平衡数据中同时存在高维问题和不平衡问题，而在高维不平衡数据分类任务中，处理高维问题的方法主要是特征选择，因此需要一个评价指标来衡量每一步特征选择的效果，但由于样本空间中存在着数据不平衡的问题，所以用传统评价标准进行评价的时候就会造成下面的问题：传统分类器追求全局分类准确率 ACC，所以即使将少数类样本全部判别为多数类仍得到一个较高的全局准确率。在这种情况下，传统的单一的评价体系将不再适用于不平衡样本分类的评价体系中。因此，我们需要一些新的指标来评价高维不平衡样本特征选择的过程以及最终的分类效果。这些标准主要有两类，一类称为“原子标准”，另外一类则称为“复合标准”，它是一种经大量研究之后所提出的原子标准和数学理论复合而成的复杂并且能够很好适应不平衡样本分类问题评价体系。此外，受试者曲线(ROC)作为一种衡量分类效果与模型稳定程度的评价标准，由于其能较直观的反应模型的特性，所以被广泛的应用于不平衡样本分类的评价工作中。

如下表 2-2 所示，通过统计混淆矩阵的各个指标以及这些指标的复合指标，我们可以分别考察不同类别的分类效果，不单纯的追求准确率而是同时考虑少数类和多数类的分类准确率。

表 2-2 混淆矩阵

	分类为正类	分类为负类
正类样例	正确正例 TP	错误负例 FN
负类样例	错误正例 FP	正确负例 TN

2.4.1 原子标准

传统分类评价准则如下公式(2-13)所示，它用于评价不平衡样本分类表现的全局准确率，这种传统分类准则主要有以下弊端：在以全局准确率为最优化目标的前提下，分类器很有可能会将数量非常稀少的少数类样本直接全部归为多数类样本空间中,这样造成的后果是少数类样本的识别率极其低下。所以，面对不平衡样本分类问题的时候，用全局准确率对评价不平衡数据的分类效果是不公平的。因此又产生了原子评价标准。

$$Overall-Accuracy=(TP+TN)/(TP+FP+FN+TN) \tag{2-13}$$

公式(2-14)到公式(2-18)列出了一些不平衡样本分类中被经常使用的原子评价标准。其中 *TPRate* 又称查全率（公式 2-14），与查准率 *FPRate*（公式 2-16）是最经常被使用的原子评价标准[49]，不平衡样本分类所追求的目标就是希望能够同时得到较高的查全率和查准率。计算公式如下：

真阳率（*TPR*），又名召回率（*Recall*）：

$$TPR=Recall=TP/(TP+FN) \tag{2-14}$$

$$Precision=TP/(TP+FP) \tag{2-15}$$

伪阳率（*FPR*）：

$$FPR=FP/(TN+FP) \tag{2-16}$$

真阴率（*TNR*）：

$$TNR=TN/(TN+FP) \tag{2-17}$$

伪阴率（*FNR*）：

$$FNR=FN/(TP+FN) \tag{2-18}$$

虽然上述原子评价标准相对全局准确率这一评价标准来说，已经对不平衡样本分类的评价工作有了相当大的改善,但在现实不平衡样本分类问题中，如果只考虑查全率或查准率的话也会导致一些新的问题发生，所以还需要更为复杂的标准或者复合标准进行评估分类效果。

2.4.2 复合标准

为了克服原子标准的一些弊端，一些“复合标准”被提出，最具代表性的是 *F-Measure* 和 *G-Means* 这两种复合评价标准。

其中 *F-Measure* 被频繁的应用到不平衡样本分类的评价工作中，如公式(2-19)所示。*F-Measure* 由查全率、查准率以及平衡因子复合得到，当 *Recall*

和 *Precision* 都较大时，*F-Measure* 才能取得较高的值[50]。

$$F-Measure=\frac{(1+\beta^2)\times Recall\times Precision}{\beta^2\times Recall+Precision} \tag{2-19}$$

式中β通常β设为 1，此时即 F1 值。在本章后续的实验中，主要用 F1 值来评价。

2.4.3 受试者工作特征曲线

受试者工作特征曲线（ROC）于 1988 年被提出[51]，一经提出便在诸多领域中得到了广泛地应用。ROC 以 Specificity（即伪阳率 *FPRate*）为 X 轴、Sensitivity（即真阳率 *TPRate*）为 Y 轴来搭建的空间。通过将伪阳率和真阳率值组成的分散的点连接起来形成 ROC 曲线。在 ROC 曲线中，用曲线所覆盖的面积值 AUC（Area under the ROC curve）来衡量算法的分类效果，其值越大越好，如图 2-6 所示，其中对角线为随机猜测的结果，其 AUC 值为 0.5，当算法的 AUC 值小于 0.5 时，该算法的效果将比不上随机猜测的结果。

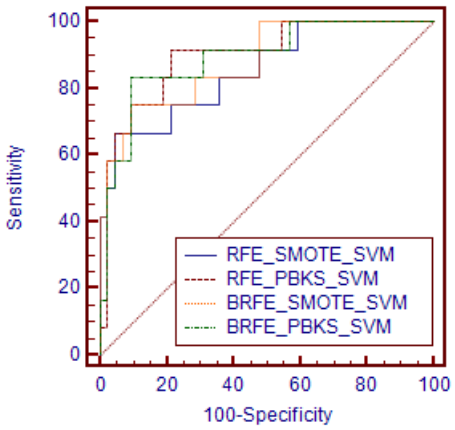

图 2-6 ROC 空间内的五条曲线

2.5 本章小结

本章首先分析了高维不平衡数据分类难的原因，究其本质是由于高维问题和不平衡问题同时存在，彼此交织，导致难以运用现有的方法先解决其中一个问题再解决另一个问题；然后对 SVM 的基础理论进行了详细的推导，从理论出发分析了 SVM 的特性以及其对于解决高维问题和不平衡问题的优越性，为下一章针对高维不平衡问题改进特征选择算法以及过采样算法提供理论依据；最后针对高维不平衡数据分类问题应采用的评价标准进行了详细的介绍，对比了各自的优势和劣势，为后续试验将要采用的评价方式以及可靠性做了详细的分析。

第 3 章 改进 BRFE-PBKS-SVM 算法

如上文所述，高维不平衡数据同时展现出高维特性和不平衡特性，并且两种特性相互影响又形成了新的问题。由于传统解决不平衡问题的方法在高维空间中收效甚微，因而现有的研究方法主要是先解决高维问题再解决不平衡问题；由于这种处理思路十分合理并且具有针对性，所以本课题也沿用先解决高维问题，再解决不平衡问题的思路，并针对高维问题和不平衡问题相互影响形成的新问题，对现有的方法进行研究和改进。

BRFE 即边界重采样特征选择(Border-Resampling-Feature-Elimination)，PBKS(PSO-Border-Kernel-SMOTE)是用 PSO 优化过采样倍率的在希尔伯特空间下的边界 SMOTE 过采样。由于现有的解决方法在解决高维问题的过程中，没有考虑不平衡问题带来的影响，所以 BRFE-PBKS-SVM 算法首先考虑解决受到不平衡问题干扰的特征选择问题：利用 SVM 自动划分边界样本的特点，对边界中被误分的少数类进行重采样，从而影响特征选择的过程，使最终得到的特征子集给予少数类更多的关注。然后，改进 SMOTE 算法，在希尔伯特空间下对少数类支持向量进行过采样并找到其对应的欧几里得空间中的近似原像，同时剔除被误分的多数类，使得过采样得到的样本点尽量符合希尔伯特空间中的边界分布并有利于提升 SVM 的分类效果。

3.1 考虑不平衡问题的特征选择算法

第 2 章已经介绍过现有的特征选择框架，基于 SVM 的特征选择方法都是属于包裹式特征选择方法，其特点是所选择的特征是最有利于提高某种分类器性能的方法。而基于 SVM 的特征选择方法按其选择方式不同又分为前向特征选择和后向特征消除两种方法，由于前向特征选择方法所选择出的特征组合有可能会忽略特征之间的组合关系并且无法考虑不平衡问题，所以针对高维不平衡问题的特征选择，本文主要采取后向特征选择法。

在高维不平衡数据中，高维问题与不平衡问题相互影响形成的新问题主要体现在特征选择的过程中：已有的特征选择算法没有考虑选择的特征是否有利于提高少数类的识别率，可能导致最终得到的特征子集不仅不利于提高分类器对少数类的关注程度，反而使得不平衡问题更加尖锐。而现有的解决方案在考虑解决高维问题的过程中，很少有考虑到不平衡问题带来的影响，所以需要在特征选择的过程中加入数据采样过程，特征选择和数据采样同时

进行。并且在高维不平衡数据中，由于高维问题的存在使得 SMOTE 过采样方法收效甚微，所以只能采用重采样的方法来提高分类器对少数类的关注。

针对上面提到的问题，基于 SVM 的特性和现有的特征选择方法，本文提出了一种解决高维不平衡数据分类任务中的高维问题的新思路：利用 SVM 中不平衡问题主要集中表现为边界样本不平衡和 SVM 自动划分边界的特点，每一轮迭代特征选择过程都对 SVM 的少数类边界样本进行重采样，以此来修正权值 w 并把它作为 SVM-RFE 中的特征评分参考来对特征进行排序。这样做的好处是，在每一步的特征选择过程中都考虑了数据不平衡带来的问题，使得权值 w 在更加关注少数类的前提下对每一个特征做出了评价，从而使得 SVM-RFE 特征选择过程朝着有利于识别少数类的方向进行，使得最终选择出的特征子集更有利于分类器关注少数类。

3.1.1 SVM-RFE 算法描述

第 2 章已经介绍了 SVM 序列极小化特征选择算法，基于 SVM 特征选择序列极小化特征选择算法和 SVM-BFE 的缺陷，SVM-RFE 将这两种方法组合，克服了彼此的缺点：SVM-RFE 不再单纯的使用原始 w 的分量的绝对值大小或者经验损失的减少程度$\nabla R_{emp}[f]$来考虑最应该删除的特征，而是将两者结合，用权向量 w 来评价特征空间中每一个特征的重要程度，按 w 的分量的绝对值大小作为迭代的搜索路径，每一轮从绝对值最小的开始迭代，直到搜索到一个使得经验损失函数减少最多并且特征权重最小的一个特征为止，然后删除该特征，其算法具体过程如表 3-1 所示。

通过这两种方法的结合，SVM-RFE 能够充分地利用 SVM 在高维数据分类问题中的优越性进行特征选择，并能保持 SVM 的表达能力不变，甚至提升。由于 SVM-RFE 在特征选择的过程中按照一定的规则对特征进行评分，依据评分来决定进行迭代后向特征选择的顺序。这种方法提供了一种兼顾不平衡问题的特征选择思路：可以通过一定的手段来改变 SVM-RFE 的特征评分体系，使得特征的评分是基于提高了分类器对少数类的关注程度来进行的。在进行迭代后向特征选择的过程中，每一轮都对剩下的特征重新评分，以考察现有特征空间中的每一个特征的重要程度。由于 SVM-RFE 是一种包裹式的特征选择方法，其衡量每一个特征重要程度的方法就是在消除某个特征后 SVM 取得的分类效果，而其消除特征的路径是依据特征权重 w 的绝对值排序的，这就提供了在 SVM-RFE 特征消除过程中考虑不平衡因素的思路：利用重采样增加 SVM 对边界少数类的关注度，改变 w 的分量值及其绝对值大小顺序，从而改变 SVM 对每一个特征的评分，影响特征选择的路径，使特

征选择的方向朝着有利增加分类器对少数类的关注程度的方向进行。

表 3-1　SVM-RFE 算法伪代码

输入：原始数据集

输出：经过特征选择后的数据集

(1) 求解SVM的分类超平面，得到 w 和 b，并计算出相应的经验风险 $R_{emp}[f]$;

(2) 按一定规则重新排列权向量 $w=(w^1,w^2,\cdots,w^n)$的分量来评价每一个特征：

$$w^{i1},w^{i2},\ldots,w^{in}$$

对应的原始数据集中的特征的重要程度是：

$$x^{i1} \le x^{i2} \le \cdots \le x^{in}$$

(3) 如果 n=1，则转(10)；否则，k=1，转(4)；

(4) 去掉 train_set 中与 w^{ik} 对应的特征 x^{ik}，并设置标志位 tag=0；

(5) 用去掉特征 x^{ik} 的数据集训练 SVM，得到新的 w 和 b，记为 w^{new}，b^{new}，计算相应的经验风险的增量为$\nabla R_{emp}[f^{new}]$，并且设置初始经验风险的增量为$\nabla R_{emp}[f]$=0；

(6) 如果$\nabla R_{emp}[f^{new}]<\nabla R_{emp}[f]$，则置$\nabla R_{emp}[f]=\nabla R_{emp}[f^{new}]$,tag=k；

(7) 如果$\nabla R_{emp}[f^{new}]==\nabla R_{emp}[f]$且 tag==0，则 tag=k；

(8) 如果$\nabla R_{emp}[f^{new}]>\nabla R_{emp}[f]$且 k=n，则转(10)；

(9) 如果 k==n，消除第 tag 个特征得到 new_train_set 并令 train_set=new_train_set，n=n-1，转(3)；否则转(3)；

(10) 返回经过特征选择后的数据集。

针对上文提到的问题以及现有方法改进思路，如何得到一个好的特征评分体系权值 w,是基于 SVM 解决高维不平衡数据中的高维问题的关键所在。

3.1.2 基于 SVM 边界样本重采样方法的特征评分体系

在 SVM-RFE 中，主要用求解 SVM 得到的权向量 w 作为特征重要性的评价指标，w 的每一个分量对应一个特征，其分量的绝对值越大，说明该分量对分类结果影响越大。第 2 章对 SVM 的基础理论作了详细的推导，在此再次列出 SVM 的分离超平面公式以及 SVM 的解 w、b，其中 x 为样本点，a_i 为样本点对应的拉格朗日参数，y_i 为对应的类别标签，K 为核函数，k 为支持向量的个数：

$$y = w \cdot x + b \tag{3-1}$$

$$w = \sum_{i=1}^{N} a_i y_i x_i \tag{3-2}$$

$$b = \frac{1}{k}\sum_{j=1}^{N}\left[y_j - \sum_{i=1}^{N} y_i a_i K(x_i, x_j) \right] \tag{3-3}$$

由于训练支持向量机的数据，原始空间是欧几里得空间，但训练空间却是希尔伯特空间。欧几里得空间中的直线在希尔伯特空间中有可能是一条更高维的曲线，并且由于通常只通过定义核函数来将训练空间映射到希尔伯特空间而不显示的定义映射函数，因此传统确定边界样本的方法行不通。由于根据每个样本点(x_i,y_i)对应的拉格朗日参数a_i的值，能够定位样本点在希尔伯特空间中的相对位置：$a_i=0$的样本点被正确分类，属于安全样本；$0<a_i<C$的样本点能被正确分类，但位于两个间隔平面上；$a_i=C$的样本点位于间隔平面的另一侧，被误分类，属于边界样本。

利用上述性质，可以直接依据拉格朗日参数a_i的值判断每个样本点(x_i,y_i)在希尔伯特空间中的几何位置。在第 2.3.2 中已经详细分析了 SVM 对完成不平衡分类任务的优势：根据求解 w 的公式(3-2)，由于安全样本对应的拉格朗日参数每个样本点$a_i=0$，因此安全样本对最终的分类超平面以及间隔平面的确定没有贡献，不平衡问题也主要体现为边界样本不平衡问题，如图 3-1 所示。

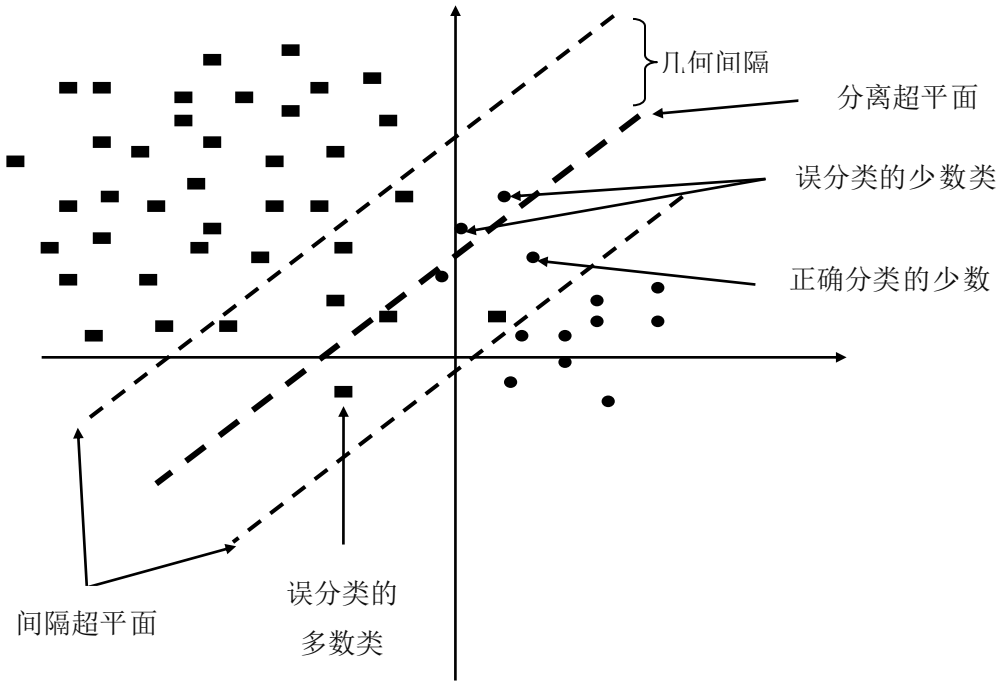

图 3-1 SVM 几何空间示意图

图中的矩形为多数类，圆形为少数类，两个间隔超平面之间的样本点为支持向量，间隔超平面两侧的样本点为安全样本，其对应的拉格朗日参数为 0；被误分的样本点处于两个间隔超平面之间。由于对 SVM 的权值向量 w 有所贡献的样本点为两个间隔超平面之间的支持向量，所以只要对间隔平面内被误分的少数类边界样本进行重采样，就能提升 SVM 对少数类的关注程度。

由于 SMOTE 过采样方法在高维空间中的失效，所以只能通过重采样来解决特征选择过程中伴随着的不平衡问题。对被误分的少数类进行单倍率重采样是指这样一个过程：复制一次数据中被误分的少数类，使得分类器提高对少数类的关注程度。

假设图 3-1 中的 2 个在边界中被误分的少数类为 x_p 和 x_q，若对它们进行单倍率的重采样，则式(3-4)和式(3-5)的结果都会改变。单倍率的重采样，即对这两个样本点，进行一次复制。具体地，假设对 x_p 和 x_q 进行单倍率的重采样之后，x_p 和 x_q 仍然没有被正确分类，则有：

$$w_{new}=\sum_{i=1}^{N}a_i y_i x_i+a_p y_p x_p+a_q y_q x_q=\sum_{i=1}^{N}a_i y_i x_i+C y_p x_p+C y_q x_q \tag{3-4}$$

$$b_{new}=\frac{1}{k+2}\sum_{j=1}^{N+2}\left[y_j-\sum_{i=1}^{N}y_i a_i K(x_i,x_j)-y_p CK(x_p,x_j)-y_q CK(x_q,x_j)\right] \tag{3-5}$$

其中 C 为惩罚参数，当样本点被误分时，其对应的拉格朗日参数的解等于 C。假设图(3-1)中的圆形类别标签为-1，矩形为+1，而 w 实际上是分离超平面的法向量，则式(3-4)相比较于式(3-3)，可以写成式(3-6)，其中 $\boldsymbol{op}$、$\boldsymbol{oq}$ 分别代表原点到点 x_p 和 x_q 的向量。式(3-6)的几何意义是：分离超平面的法向量朝着 $\boldsymbol{op}+\boldsymbol{oq}$ 的方向移动，具体移动幅度的大小与惩罚参数 C 的设置有关。

$$w_{new}=w+y_p C\boldsymbol{op}+y_q C\boldsymbol{oq} \tag{3-6}$$

接下来再对式(3-5)做一个形变，分离出所有与 x_p 和 x_q 有关的项，如式(3-7)所示：

$$\begin{aligned}b_{new}=&\frac{1}{k+2}\sum_{j=1}^{N}\left(y_j-\sum_{i=1}^{N}y_i a_i K_{ij}\right)-\frac{C y_p}{k+2}\sum_{j=1}^{N}\left(CK_{pj}+CK_{qj}\right)-\\&\frac{y_p}{k+2}\left(2CK_{pq}+CK_{pp}+CK_{qq}+\sum_{i=1}^{N}y_p a_p K_{ip}+\sum_{i=1}^{N}y_q a_q K_{iq}-y_p y_p-y_q y_p\right)\\=&\frac{1}{k+2}\sum_{j=1}^{N}\left(y_j-\sum_{i=1}^{N}y_i a_i K_{ij}\right)-\frac{C y_p}{k+2}\left(2K_{pq}+K_{qq}+K_{pp}-\frac{y_p y_p}{C}-\frac{y_p y_q}{C}\right)\end{aligned} \tag{3-7}$$

事实上，在求解 SVM 的过程中，每一个支持向量都对应一个 b 的值，而最终求得的 b 值是所有支持向量所对应的 b 的均值，于是又由式(3-5)得到最终 b 的值与每一个支持向量的 b 值的关系，如式(3-8)所示：

$$b_{new}=\frac{1}{k+2}\sum_{j=1}^{N+2}b_j=\frac{1}{k+2}\sum_{j=1}^{N}b_j+\frac{b_p+b_q}{k+2}=\frac{1}{k+2}(b+b_p+b_q) \tag{3-8}$$

因此，式(3-7)中的第二项与式(3-8)中的第二项相等，即：

$$\begin{aligned} y_p + y_q - \sum_{i=1}^{N+2} y_i a_i K_{ip} - \sum_{j=1}^{N+2} y_j a_j K_{jq} &= C y_p (2K_{pq} + K_{pp} + K_{qq} - \frac{y_p y_q}{C} - \frac{y_p y_p}{C}) \\ &= y_p + y_q - \sum_{i=p,q} y_i a_i K_{ip} - \sum_{j=p,q} y_j a_j K_{jq} \end{aligned} \tag{3-9}$$

又假设现有一个分离超平面，x_p 和 x_q 作为仅有的支持向量位于该几何空间中的分离长平面上或者分离超平面被误分的一侧，则它们对应的拉格朗日参数 a_p 和 a_q 均等于 C，则该分离超平面如式(3-10)所示：

$$y = w_{pq} \cdot x + b_{pq} \tag{3-10}$$

其中 w_{pq} 和 b_{pq} 的值如式(3-11)和式(3-12)所示：

$$w_{pq} = \sum_{i=p,q} a_i y_i x_i = y_p C \boldsymbol{op} + y_q C \boldsymbol{oq} \tag{3-11}$$

$$b_{pq} = \frac{1}{2} \sum_{j=p,q} \left[y_j - \sum_{i=p,q} y_i a_i K_{ij} \right] = \frac{1}{2}(b_p + b_q) \tag{3-12}$$

又由式(3-9)整理得：

$$\sum_{i=1}^{N} y_i a_i K_{ip} + \sum_{j=1}^{N} y_j a_j K_{jq} = 0 \tag{3-13}$$

观察式(3-11)与式(3-6)的关系以及式(3-12)与式(3-8)的关系，再结合式(3-13)的结果，可以得到如下结论：对边界中被误分的少数类进行单倍率的重采样，实际上相当于用一个这样的分离超平面来修正现有的分离超平面：以被误分的样本点到空间中的原点的向量之和作为法向量、以当前分离超平面中被误分的少数类支持向量所对应的偏移量 b 的均值作为偏移量的 SVM 分离超平面。如图 3-2 所示。

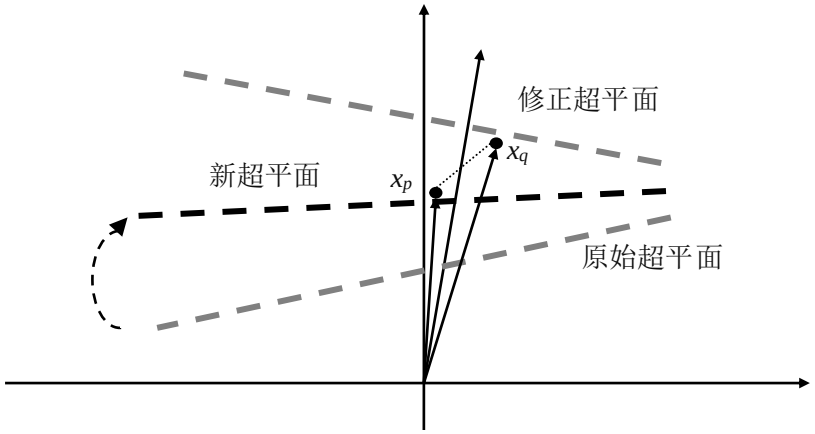

图 3-2 超平面修正示意图

从图 3-2 中可以看到，在对被误分的少数类进行单倍率的重采样之后，即使 x_p 和 x_q 没有被正确分类，但他们到新超平面的距离也缩小了；因此随着重采样的进行，一定能使得分离超平面朝着能正确划分所有少数类的一侧移

动，但其代价是有可能会将部分边界多数类误判成少数类。

对比图 3-3 与图 3-1，发现经过对误分少数类边界样本的重采样，图 3-3 中被误分的少数类从图 3-1 中的 2 个变成了 1 个，而分离超平面也朝着能将更多少数类正确划分的方向移动，同时几何间隔缩小，两个间隔超平面朝着分离超平面收缩。

在运用重采样方法来解决不平衡问题的思路中，采样倍率的设置是一个很关键的环节。在 SVM 的希尔伯特空间中，由于每一次重采样都会使分离超平面产生变化，从而支持向量也随之动态的改变，假如边界中的每一个少数类支持向量，其采样倍率都是固定的，有可能会出现这样的情况：采样的幅度过大使得分类器牺牲了过多的多数类的识别正确率来换取不必要的分离超平面的移动，这种不必要的分离超平面移动并没有真正的提高少数类的识别率。

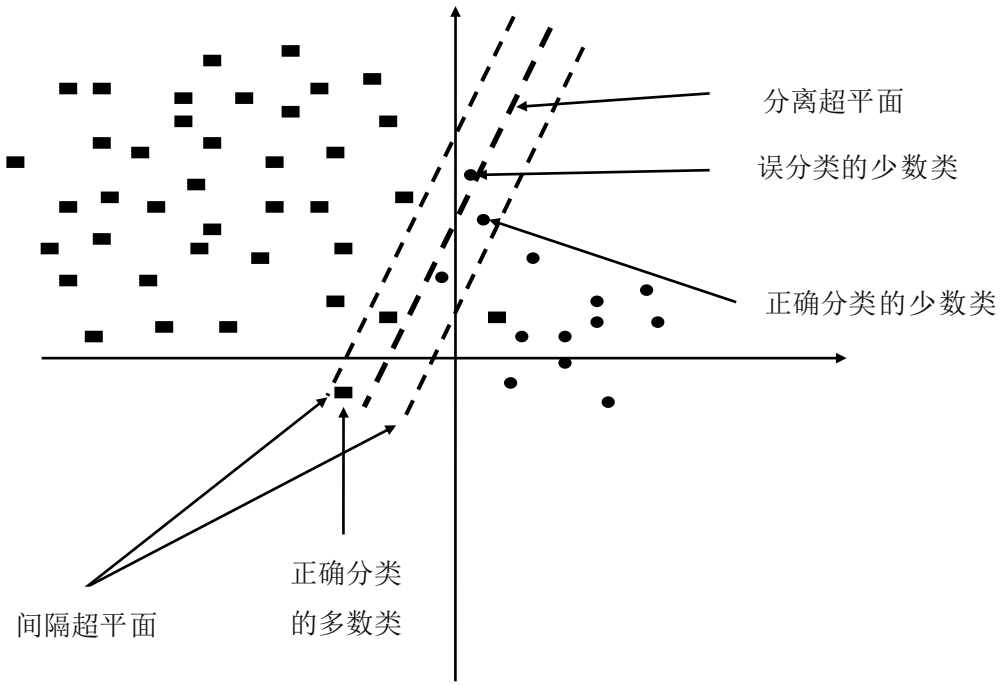

图 3-3 重采样后的 SVM 几何空间示意图

分析在 SVM 中运用重采样会造成上述问题的原因：考虑公式(3-2)，发现每当对少数类边界样本进行一轮单倍率的重采样之后，分离超平面及间隔超平面都会发生变化，导致边界样本有所变化，而在 SVM 中，边界样本作为支持向量，是必然存在的一类样本点。因此不能设置单一的重采样倍率，单纯的对每个少数类边界样本进行 n 倍的复制来使分离超平面移动。

针对这个问题，本文运用 F1 值作为评价指标，提出了一个重采样思路：

每一轮都对边界中被误分的少数类样本进行单倍率的重采样，即对被误分的少数类边界样本复制一次，再重新在希尔伯特空间下训练 SVM，以使得分离超平面在尽量少牺牲多数类的识别正确率的情况下，朝着最有利于正确识别少数类的方向移动，直到不再有被误分的少数类为止，然后取其中最大的 F1 值对应的 w 作为当前一轮特征选择的特征评分，算法流程图如图 3-4 所示：

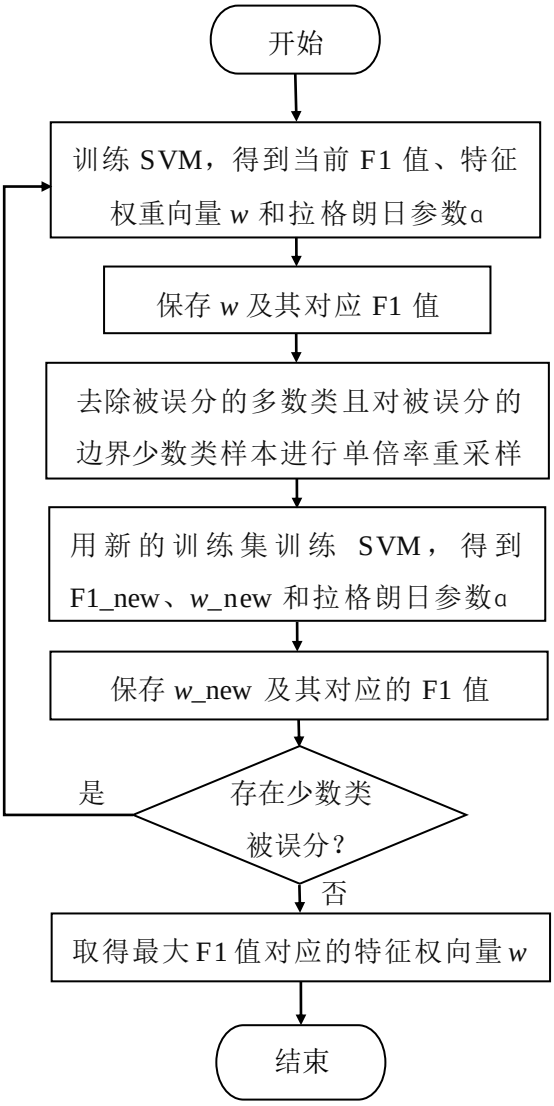

图 3-4 SVM 边界重采样算法流程图

在图中可看到，由于边界的不断变化，对被误分的少数类边界样本的重采样不是一蹴而就的，为取得较好的效果，需要通过反复对新的被误分少数类进行重采样以逐步修正 SVM 的分离超平面的几何位置，直到不存在被误分的少数类为止。这样做的好处是使权向量 w 在 SVM 提高对少数类的关注程度的前提下，更加公平的反映训练集中特征的重要程度。经过图 3-4 中的流程后，训练 SVM 得到的权向量 w，能够在更加关注少数类的情况下更好

的反应特征的重要性，使得有利于提升少数类分类效率的特征权重得以增加。

3.1.3 改进的 SVM-BRFE 特征选择算法

通过 3.1.1 节介绍的 SVM-RFE 特征选择过程，发现可以在特征迭代选择的过程中，通过改进包裹式特征选择过程的特征评价体系来兼顾不平衡问题，于是 3.1.2 节利用了 SVM 自动划分边界的特点来对希尔伯特空间下的样本点进行重采样来使支持向量机模型的 F1 值有所提高，并用此时 SVM 的特征权向量 w 作为特征的评价标准。下一步便是将这两者结合起来，在考虑不平衡问题的情况下对高维不平衡数据进行特征选择，解决高维问题。该算法的时间复杂度为 $O(d^2)$，d 为特征的总数，主要过程如表 3-2 所示：

表 3-2　SVM-BRFE 算法伪代码

输入：原始数据集 train_set

输出：经过特征选择后的数据集

(1) 求解 SVM 的分类超平面，得到 w 和 b 和拉格朗日参数α，并计算出相应的 F1 值；令 new_train_set=train_set；

(2) 对 new_train_set 中 $a==C$ 的少数类进行单倍率重采样，更新 new_train_set 并用它来训练 SVM，得到并保存 F1_new 和 w^{new}，得到 w^{new} 按绝对值大小排序：

$$|w^{new,i1}| \leq |w^{new,i2}| \leq, \ldots, \leq |w^{new,in}|$$

对应的原始数据集中的特征的重要程度是：

$$x^{i1} \leq x^{i2} \leq \cdots \leq x^{in}$$

(3) 还存在被误分的少数类，转(2)，否则转(4)；

(4) 令当前 w 为历史最大 F1 值对应的 w，当前 F1 值为历史最大的 F1 值；

(5) 消除重采样所复制产生的样本点；

(6) 如果 n=1，则转(3)；否则，k=1，转(7)；

(7) 去掉 train_set 中与 w^{ik} 对应的特征 x^{ik}，并设置标志位 tag=0；

(8) 用去掉特征 x^{ik} 的数据集训练 SVM，得到新的 w 和 b，记为 w^{new}，b^{new}，计算相应的 F1 值的增量为▽F1_new，并且设置当前 F1 值增量▽F1_now=0；

(9) 如果▽F1_now<▽F1_new，则置▽F1_now =▽F1_new，tag=k；

(10) 如果▽F1_now==▽F1_new 且 tag==0，则 tag=k；

(11) 如果▽F1_now>▽F1_new 且 k==n，则转(13)；

(12) 如果 k==n，消除第 tag 个特征得到 new_train_set 并令 train_set=new_train_set，n=n-1，转(2)；否则转(7)；

(13) 返回经过特征选择后的数据集。

首先，训练 SVM，得到最初的特征权重向量 w、拉格朗日参数 a和 F1 值，

记录下这 3 个值以便后续对比使用。

然后，对$a=C$ 的少数类进行单倍率重采样，并用重采样后的数据训练 SVM，使 SVM 的分离超平面朝着 F1 值增大的方向移动；由于分离超平面的每一次变化都会伴随着分隔超平面的同时变化，边界样本也会有所改变，因此需要不断重复该过程，每一次都对新的少数类样本边界进行单倍率的重采样，直到找到使 F1 值最大的分离超平面为止，用这个 w 值作为一轮特征选择的特征评分。

最后，按照特征的重要程度从小到大排列进行迭代特征消除，每轮消除一个特征使得 F1 值提高最多；由于每一轮消除了一个特征之后 SVM 的分离超平面同样也会改变，边界样本也随之发生改变，因此也同样需要对剩下的特征重新评分以产生新的特征权重 w 来评价每一个特征在新的特征空间下的重要程度。

在此，值得注意的是，特征选择部分的重采样过程并不参与训练集的更新：对少数类边界样本进行重采样只是为了得到一个相对于多数类和少数类比较公平的特征权重 w，以更好的衡量在高维不平衡数据中，每一个特征的重要程度，而不是为了直接改变 SVM 对少数类的关注程度以提高直接分类效果和 F1 值，也就是说每一轮特征选择前的重采样过程只是为了解决收到不平衡问题影响的高维问题，而不是为了解决不平衡问题。因此，当得到最大的 F1 值时，当前一轮的重采样过程结束，保存 SVM 在取得最大 F1 值时的权重向量 w，用它来衡量特征的重要程度并对特征排序，接着去除掉重采样复制的少数类样本点，只保留原始的少数类样本点，然后进入特征选择过程。每当选择出一个特征之后，又重复上述过程，直到选择出最优的特征子集为止。从表 3-2 中可以看到，重采样过程并不更改 train_set，只有在特征选择的过程中才在每选择一个特征之后更新 train_set。

通过以上的几个步骤：对边界进行重采样以寻找最优特征权重以衡量特征重要程度、特征选择、更新训练集并重复以上过程，最终保留最有利于提升 F1 值的特征，其他特征将被剔除，使得后续的训练过程在一个特征冗余、无关特征组合尽量少和维数尽量低的情况下进行，减少了高维问题对不平衡问题的影响和对 SMOTE 过采样算法的束缚，有利于在后续过程中改进传统过采样算法来解决不平衡问题，提升分类效果。

3.2 希尔伯特空间下的过采样算法

上文已经提出了解决高维不平衡数据中高维问题的方法，由于在特征选

择的过程中进行的样本边界重采样只是解决了不平衡问题对特征选择过程的影响，并没有真正完全的解决高维不平衡数据分类中的不平衡问题，所以在进行了特征选择之后，仍然需要考虑解决不平衡问题。由于高维不平衡数据中的高维问题得到了解决，使得特征之间相互矛盾、冲突以及特征冗余等问题大大减小，训练集的维数有所减少，因而大幅度减少高维问题对不平衡问题的影响，使得一些经典的过采样方法可以用来解决高维不平衡数据分类问题中的不平衡问题。虽然在 SVM 中，不平衡问题主要体现在边界样本中的不平衡的这一点特性有利于不平衡问题的解决，但由于 SVM 训练空间是希尔伯特空间，而原始训练集的特征空间是希尔伯特空间，导致 SMOTE 方法不能直接用于使用 SVM 分类的特征空间中。

如第二章所述，高维不平衡数据分类问题的难点，主要由高维问题和不平衡问题两部分组成，上文已经给出了高维问题的解决方法。在解决了高维问题之后，高维不平衡数据中高维问题对不平衡问题的干扰已经大大消减，因此可以考虑运用 SMOTE 等过采样方法来解决接下来的问题。但由于本文采用的包裹式特征选择算法来解决高维不平衡数据分类问题中的高维问题的思路，所选择出的特征空间是最有利于提高 SVM 的分类效率的特征子空间，因此后续问题的解决过程不宜采用其他分类器。

SVM 是在希尔伯特空间下求解拉格朗日极小极大值的对偶问题，在 SVM 的训练过程中，如公式(2-31)所示，通过引入核函数使欧几里得空间变换为希尔伯特空间从而解决非线性分类问题；一个核函数唯一对应一个希尔伯特空间[52]。由于希尔伯特空间引入的目的就是为了解决非线性分类问题，因此在欧几里得空间中的超平面或连接两点形成的直线，在映射到希尔伯特空间后有可能被扭曲成超曲面或者曲线，同样希尔伯特空间中的超平面或者连接两点形成的直线在被逆映射到欧几里得空间中有可能被扭曲成超曲面或者曲线。

在利用 SVM 对不平衡数据进行分类时，由于不平衡问题主要体现为边界样本不平衡，因此利用 SMOTE 方法对 SVM 自动划分的少数类边界样本进行线性插值，增加边界中少数类的数量，才能增加 SVM 对少数类的关注程度，使分离超平面沿着有利于 F1 值提高的方向移动。而考虑到上面所提到的欧几里得空间中的直线在希尔伯特空间中有可能被扭曲成曲线的问题，用 SMOTE 方法在两个少数类边界样本之间合成新的少数类的过程必须在希尔伯特空间下进行，才能保证所产生的新的少数类，在希尔伯特空间下仍然是边界样本，从而使分离超平面沿着有利于提高少数类识别率的方向移动。

这样一来就产生了一个新问题：由于过采样之后，训练数据集必然发生

变化，需要用新的训练数据训练 SVM；而在 SVM 中，输入空间是欧几里得空间，特征空间为希尔伯特空间，它们是两种不同的空间，因此在希尔伯特空间下进行 SMOTE 过采样产生的新的少数类边界样本点需要找到它在输入空间中对应的原像，才能更新训练集，这就涉及到了一个空间转换的问题。假设在希尔伯特空间中参与合成新样本点的两个样本点为双亲样本点，则所生成的新的样本点在欧几里得空间中对应的原像，极有可能不在双亲样本点在欧几里得空间中对应的原像的连线之中，反之亦然，如图 3-5 和图 3-6 所示。

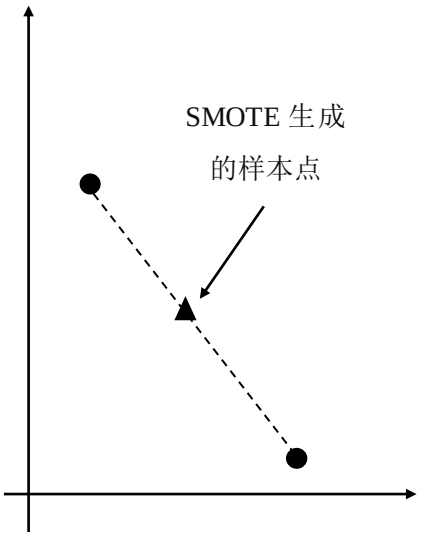

图 3-5 欧几里得空间下的 SMOTE 采样

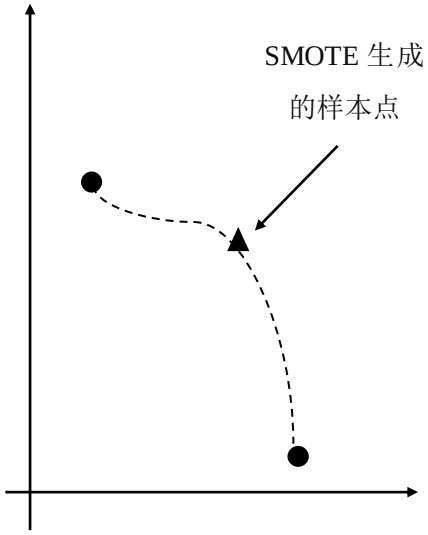

图 3-6 希尔伯特空间下的 SMOTE 采样

3.2.1 PSO 算法描述

在运用重采样、过采样等方法来解决不平衡问题时，由于每个样本点的重要程度不同，提高某些样本点的采样倍率会使分类器的分类效果提升，相反，对某些样本设置过高的采样倍率则有可能会是分类效果下降，因此采样倍率的设置往往是一个关键的环节。本文在解决高维不平衡数据分类任务中的高维问题时，也针对不同的样本点进行了多次的单倍率重采样。然而，由于人为设定的重采样倍率是根据先验知识来设置的，不一定是最优的重采样倍率，因此针对高维不平衡数据分类任务中的不平衡问题，本文引入了粒子群优化算法（Particle Swarm Optimization, PSO）[54]来对边界少数类的重采样倍率进行自适应的设置。

PSO 算法是一种计算方法，它主要通过迭代优化问题和一个特定的适应度函数来试图得到最佳候选解决方案。PSO 算法中的“粒子”代表整个群体

搜索空间中的每一个个体，PSO 算法优化的具体过程如下：初始化每个粒子的位置和更新速度，这个速度包含粒子在下一步搜索过程所要运行的轨迹，然后搜索当前区域的最优解。PSO 算法每次的更新中，有两个重要的值需要保留，其中一个是群体中每个粒子的个体最优解，记为 $pbest$；还有一个解为全局最优解，代表粒子群整个搜索空间中最优解，记为 $gbest$。粒子的速度和位置时刻根据这两个值进行更新，每一次的迭代都是根据以下两个公式进行更新的：

$$v_i^{t+1} = w \times v_i^t + c_1 \times r_1 \times (pbest_i - x_i^t) + c_2 \times r_2 \times (gbest - x_i^t) \tag{3-14}$$

$$x_i^{t+1} = x^t + v_i^{t+1} \tag{3-15}$$

其中v_i^t表示第 i 个粒子在第 t 轮迭代时的速度，x_i^t表示第 i 个粒子在第 t 轮迭代时的位置。这里的 w 是惯性因子，与 SVM 中的 w 不同，它用于控制步长，权衡局部最优解和全局最优解；c_1 和 c_2 是学习因子，分别代表粒子本身的思考能力和行为能力，r_1 和 r_2 随机取开区间(0,1)中的值。

在本文中，将 PSO 运用到自适应寻找过采样倍率的过程时，公式的各项分别有各自不同的意义，x_i^t表示第 i 个粒子在第 t 轮迭代时对每个样本点设置的过采样倍率；$pbest_i$代表第 i 个粒子在它过去所曾被设置过的每一个过采样倍率时，SVM 所取得的分类效果中最好的分类效果所对应的过采样倍率；$gbest$表示所有粒子在对每个样本点设置的不同的采样倍率中，历史上 SVM 所取得效果最好的一组过采样倍率。

除此之外，PSO 算法还将被运用到利用网格法寻找希尔伯特空间下过采样产生的样本点在欧几里得空间中的对应原像中。此时x_i^t表示第 i 个粒子在第 t 轮迭代时所处网格的位置；$pbest_i$代表第 i 个粒子在它过去所经历的所有网格中与 SMOTE 在希尔伯特空间产生的样本点最近的网格；$gbest$表示所有$pbest_i$中，位置最优的网格。

通过将 PSO 算法运用到希尔伯特空间对应欧几里得空间的近似原像的寻找和边界过采样算法 Border-Kernel-SMOTE 中，能够让分类器自适应的选择最能还原希尔伯特空间下的少数类边界分布的样本点和最有利于提高 SVM 分类效率的少数类过采样倍率。

3.2.2 希尔伯特空间下的 PSO-Border-Kernel-SMOTE 算法

在第二章中已经提过，SVM 通过公式(2-30)隐式的将欧几里得空间映射到希尔伯特空间，直接定义核函数 K，而不显示的定义空间映射函数φ(x)，

这样一来，直接求解映射函数φ(x)的反函数来得到希尔伯特空间中的点在欧几里得空间中的原像的思路无法进行。

考虑到上述问题，精确的原像无法求出，只能求解近似原像来代替，因此利用希尔伯特空间与对应的欧几里得空间之间的距离关系来寻找希尔伯特空间中 SMOTE 算法合成的样本 z_{ij} 在对应的欧几里得空间下的近似原像 x_{ij}[53]。

在解决该问题之前，首先提出希尔伯特空间下的距离度量方式：

$$z_i = \varphi(x_i) \tag{3-16}$$

$$d^2(\varphi(x_i),\varphi(x_j)) = K_{ii} + K_{jj} - 2K_{ij} \tag{3-17}$$

设欧几里得空间到希尔伯特空间的隐式映射如式(3-16)所示，并假设显示定义的核函数如式(2-30)所示。在以后的书写中，都用 K_{ij} 代替 $K(x_i,x_j)$，它表示欧几里得空间中的两个点 x_i 和 x_j 在被映射到希尔伯特空间后的内积。则希尔伯特空间下的距离的平方如式(3-17)所示。

当核函数是高斯核时，如式(2-33)所示，欧几里得空间下的距离平方与希尔伯特空间下的距离平方的关系如式(3-18)和式(3-19)所示，D^2 表示欧几里得空间下的距离的平方，d^2 表示希尔伯特空间下的距离的平方。

$$D^2(x_i,x_j) = -2\sigma^2 In(-\frac{1}{2}d^2(\varphi(x_i),\varphi(x_j)) + 1) \tag{3-18}$$

$$d^2(\varphi(x_i),\varphi(x_j)) = 1 - \exp(\frac{D^2(x_i,x_j)}{-2\sigma^2}) \tag{3-19}$$

SMOTE 算法寻找与样本点 x_i 最邻近的前 k 个样本，然后在这个 k 个样本中随机选择一个样本点 x_j，在样本点 x_i 与样本点 x_j 之间进行线性插值。由于本文主要考虑少数类边界样本的过采样，因此将在希尔伯特空间下，对于每个处于边界中的少数类样本点，随机选择边界中的另一个少数类样本点作为 SMOTE 算法的输入，则希尔伯特空间下的 SMOTE 过采样公式如式(3-20)所示，其中 λ_{ij} 是一个在开区间(0,1)之间的随机数。

$$z_{ij} = \varphi(x_i) + \lambda_{ij} \times (\varphi(x_j) - \varphi(x_i)) \tag{3-20}$$

要寻找 z_{ij} 在希尔伯特空间下的近似原像，样本点之间的距离约束对确定原像的近似位置十分重要：

假设希尔伯特空间下用 SMOTE，过采样生成的样本点 z_{ij} 与 SVM 中每个少数类边界样本之间的距离平方向量 $\boldsymbol{d}_{z_{ij}}$ 如式(3-21)所示，假设边界中少数类样本的总数是 k 个：

$$d_{z_{ij}} = [d^2(z_{ij},\varphi(x_1)), d^2(z_{ij},\varphi(x_2)), \cdots, d^2(z_{ij},\varphi(x_k))] \tag{3-21}$$

又假设在训练集原来的欧几里得空间中有一个未知样本点为 x_{ij}，则 x_{ij}

与式(3-21)中这 k 个样本点的距离平方向量 $\boldsymbol{D}_{x_{ij}}$ 如式(3-22)所示。在式(3-21)和式(3-22)中，下标 1,2,⋯,k 所对应的样本点必须一致。

$$D_{x_{ij}}=[D^2(x_{ij},x_1),D^2(x_{ij},x_2),\cdots,D^2(x_{ij},x_k)] \tag{3-22}$$

当核函数为高斯核函数时，结合式(3-19)和式(3-22)，将欧几里得空间下的向量 $\boldsymbol{D}_{x_{ij}}$ 映射到对应的希尔伯特下，如式(3-23)所示。

$$D_{x_{ij}}^{H}=-2\sigma^2[In(-\frac{1}{2}d\ (\hat{z}_{ij},\varphi(x_1))+1),\cdots,In(-\frac{1}{2}d\ (\hat{z}_{ij},\varphi(x_k))+1)] \tag{3-23}$$

式(3-22)的值与式(3-23)的值越接近，说明 x_{ij} 经过空间变换后，在高斯核函数对应的希尔伯特空间中的位置 $\varphi(x_{ij})$ 越接近 SMOTE 合成的样本点 z_{ij}。

借鉴文献[53]利用前 k 个与 SMOTE 产生的样本点距离最近原始少数类样本点作为约束来确定希尔伯特空间样本的原像的思路，为了能够很好的填充边界少数类，本文考虑利用 SVM 自动划分出的边界中的少数类作为 $\boldsymbol{D}_{x_{ij}}^{H}$ 中的距离约束，以此来取代原始约束，并采用网格法来寻找该近似原像。具体地：假设 SVM 训练后，在希尔伯特空间中划分出来的少数类边界样本的标号为 1,2,⋯,k，求出这 d 个特征在这 k 个少数类边界样本中的上边界和下边界，如式(3-24)和式(3-25)所示，其中(3-24)是所有少数类边界样本的下边界，(3-25)是所有少数类边界样本的上边界。

$$x_{low}=\begin{bmatrix}\min\{x_1^1,x_2^1,\cdots,x_k^1\}\\ \min\{x_1^2,x_2^2,\cdots,x_k^2\}\\ \vdots\\ \min\{x_1^d,x_2^d,\cdots,x_k^d\}\end{bmatrix} \tag{3-24}$$

$$x_{high}=\begin{bmatrix}\max\{x_1^1,x_2^1,\cdots,x_k^1\}\\ \max\{x_1^2,x_2^2,\cdots,x_k^2\}\\ \vdots\\ \max\{x_1^d,x_2^d,\cdots,x_k^d\}\end{bmatrix} \tag{3-25}$$

然后按式(3-26)划分每一个网格的粒度，将边界少数类空间划分成 k×d 个网格，每个网格代表一个欧几里得空间中的位置，要寻找到一个网格使得它映射到希尔伯特空间后与过采样产生的点最相近。具体地，每一个网格的大小为该特征维度上的最大值减去最小值再除以原始边界样本的总数 k，在后续搜索原像的过程中，将以每一个网格为单位，搜索整个网格空间。

$$x_{unit}^{i}=\frac{x_{high}^{i}-x_{low}^{i}}{k} \tag{3-26}$$

式(3-20)中的 z_{ij} 是在希尔伯特空间中进行 SMOTE 过采样所生成的少数

类样本点，是已知的；式(3-22)中的 x_{ij} 是要求的 z_{ij} 的原像，是未知的。式(3-26)表示第 i 个特征的网格粒度，在每一次 PSO 随机网格搜索中，每一维都加上 PSO 所优化的网格粒度的数目得到 x_{ij}，并将该次搜索的样本点作为求解变量 x_{ij} 的一次迭代。代入式(3-22)中，然后求得式(3-22)与式(3-23)的余弦距值的平方，如式(3-27)，直到迭代结束为止。最后，用余弦值的平方最大的点代替目标解 x_{ij} 作为 z_{ij} 的近似原像。

$$\cos^2 < D_{x_{ij}}, D_{x_{ij}}^H >= \frac{D_{x_{ij}} \cdot D_{x_{ij}}^H \cdot D_{x_{ij}} \cdot D_{x_{ij}}^H}{\left\| D_{x_{ij}} \right\|^2 \cdot \left\| D_{x_{ij}}^H \right\|^2} \tag{3-27}$$

引入 PSO 算法后，整个解决高维不平衡数据分类任务中的不平衡问题的流程如图 3-7 所示：

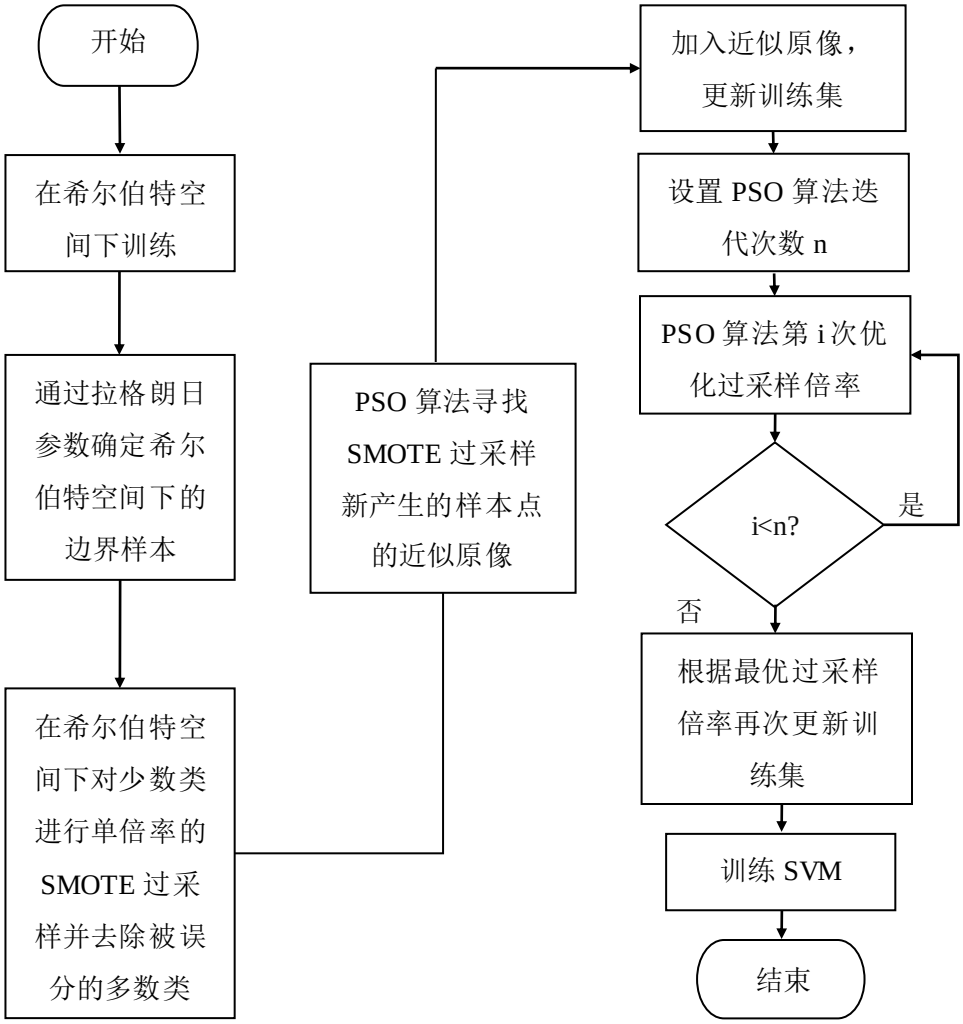

图 3-7 不平衡问题的解决流程图

利用 SVM 自动划分样本边界和在 SVM 中不平衡问题主要集中体现为边界样本不平衡问题的特点，PSO-Border-Kernel-SMOTE 算法在希尔伯特空间

下利用不同的两个少数类合成新的少数类，并寻找过采样产生的样本点在欧几里得空间中的近似原像，同时利用 PSO 算法自适应的对少数类边界样本点以及新产生的样本点的采样倍率进行优化，提升 SVM 的分类效果。从图 3-7 中可以看到，左侧部分的流程在希尔伯特空间下完成，右侧部分的流程主要欧几里得空间下完成，中间的部分是欧几里得空间下的操作和希尔伯特空间下的操作进行对接的关键。算法的伪代码如表 3-3 所示，其时间复杂度取决于 PSO 算法迭代次数的设置，因此为常数级时间复杂度 $O(c)$，c 为常数。

表 3-3 PSO-Border-Kernel-SMOTE 算法伪代码

输入：消除高维问题后的数据集 train_set

输出：SVM 模型

(1) 求解 SVM 的分类超平面，依据拉格朗日参数α定位希尔伯特空间下的边界样本，找到 k 个少数类边界样本点；

(2) for i in range(0,k):
　　for j in range(i+1,k):
　　在希尔伯特空间中利用 SMOTE 算法在 $\Phi(x_i)$ 和 $\Phi(x_j)$ 之间插值

(3) 假设希尔伯特空间下 SMOTE 算法生成的少数类样本点为 x_{ij}，则 x_{ij} 与 k 个原始少数类边界样本在欧几里得空间中的距离向量为：

$$D_{xij}=[\|x_{ij}-x_1\|^2,\|x_{ij}-x_2\|^2,\cdots,\|x_{ij}-x_k\|^2]$$

(4) 求 k 个原始边界少数类在欧几里得空间中与每个过采样产生的少数类样本点对应的近似原像 x_{ij} 的距离的平方组成的矩阵 $\boldsymbol{D}_{x_{ij}}^{H}$：

$$D_{x_{ij}}^{H}=-2\sigma^2[In(-\frac{1}{2}d^2(z_{ij},\varphi(x_1))+1),\cdots,In(-\frac{1}{2}d^2(z_{ij},\varphi(x_k))+1)]$$

(5) for ij in range(0 , (k-1)*k/2)：
　　一轮 PSO 迭代寻找到 x_{ij} 的近似使下式最小：

$$\cos^2<D_{x_{ij}},D_{x_{ij}}^{H}>=\frac{D_{x_{ij}}\cdot D_{x_{ij}}^{H}\cdot D_{x_{ij}}\cdot D_{x_{ij}}^{H}}{\|D_{x_{ij}}\|^2\cdot\|D_{x_{ij}}^{H}\|^2}$$

　　更新 train_set：train_set←x_{ij}；

(6) for i in range(0,n):
　　用 PSO 优化 train_set 中的 k 个原始边界少数类和生成的 k*(k-1)/2 个少数类的过采样倍率；

(7) 按照 PSO 寻找到的最优过采样倍率对这 k+k*(k-1)/2 个少数类进行重采样，并更新训练集 train_set：
　　train_set←u 倍过采样倍率*$x_{ij;}$

(8) 用新的 train_set 训练 SVM；

(9) 返回 SVM 模型；

3.3 BRFE-PBKS-SVM 算法描述

BRFE-PBKS-SVM 算法。该算法的具体流程如图 3-8 所示：

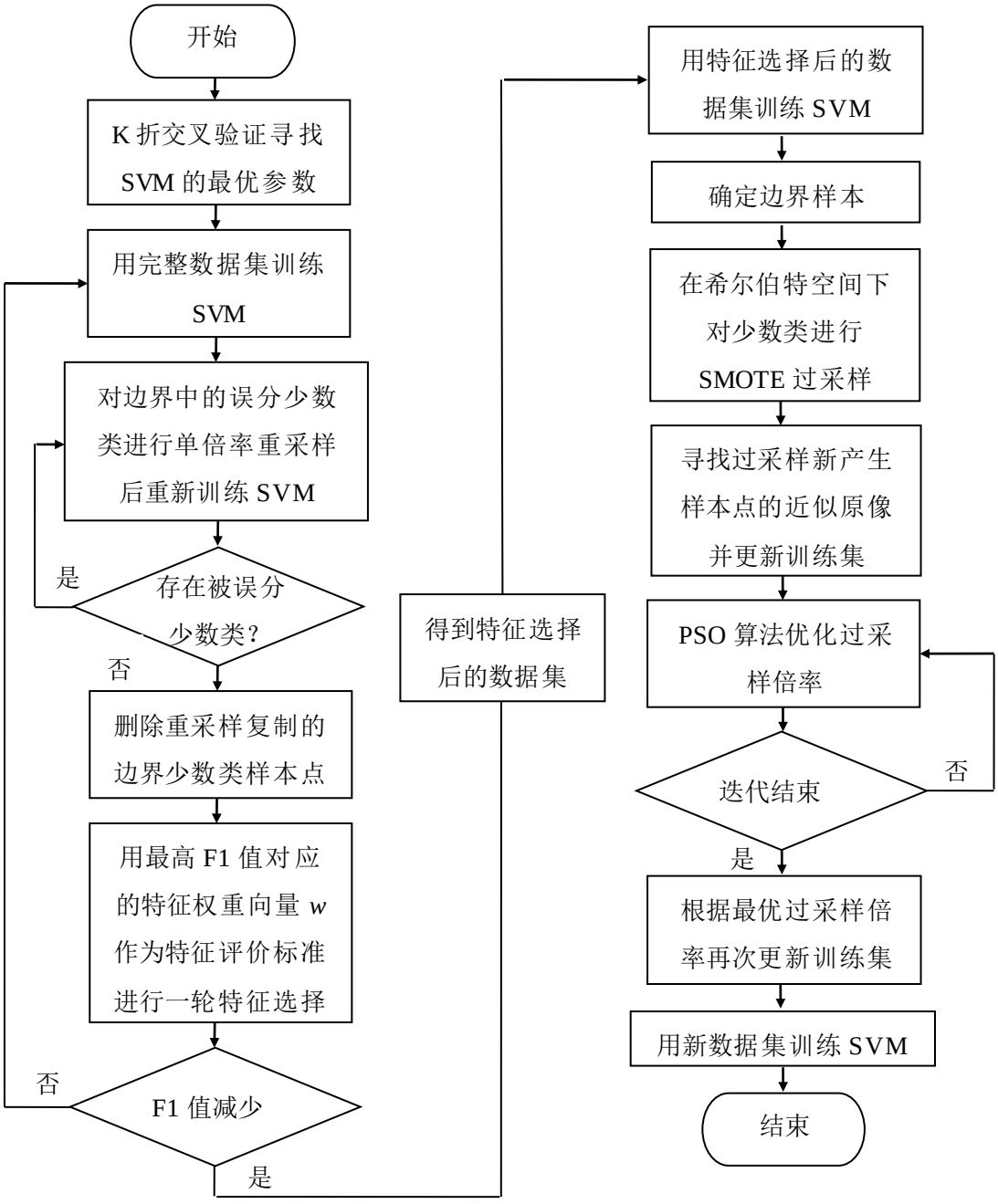

图 3-8 BRFE-PBKS-SVM 算法流程图

通过以上对高维不平衡数据二分类任务中高维问题和不平衡问题的解决思路，本课题提出了一种基于 SVM 的考虑不平衡特性的包裹式特征选择方法，并使用上述方法做特征选择后得到的数据集作为输入，提出了一种最有利于提升 SVM 分类效果的过采样方法，同时用 PSO 算法优化过采样倍率，

自适应的寻找最优参数，经过上述步骤的改进，提出了一种基于 SVM 的针对高维不平衡数据二分类的算法：BRFE-PBKS-SVM 算法。其总体时间复杂度为 $O(d^2+c)$，其中 d 为特征数目，c 为常数。算法的总体流程描述如下：

（1）用完整的数据集训练 SVM，并确定希尔伯特空间中的边界样本；

（2）对边界中每个被误分的少数类样本进行单倍率的重采样后重新训练 SVM 以重新确定边界并重复这一步骤直到不存在被误分的少数类为止；

（3）删除第 2 步中重采样复制的样本点；

（4）用第 2 步中最高 F1 值对应的 w 作为路径做一轮特征选择；

（5）回到第 2 步直到选择出最优的特征子集为止；

（6）用特征选择后得到数据集训练 SVM，并重新确定边界样本；

（7）在希尔伯特空间下对边界少数类进行 SMOTE 过采样；

（8）寻找在希尔伯特空间中过采样产生的新的少数类样本点对应的欧几里得空间下的近似原像，并把它们的近似原像加入特征选择后的数据集中；

（9）用 PSO 算法优化过采样倍率更新训练集；

（10）用新的数据集训练 SVM，并输出最终训练好的 SVM 模型。

3.4 本章小结

本章基于第 2 章的基础研究工作，首先着手解决高维不平衡数据分类任务中的高维问题：在考虑不平衡问题对特征选择带来的影响的情况下，不断对边界中被误分的少数类进行单倍率的重采样以获得新的特征评分，使有利于少数类识别的特征的评分得到提高，以此来修正特征选择路径，提出了一种基于 SVM 的 SVM-BRFE（SVM-Border-Resampling-Feature-Elimination）特征选择算法；该算法通过修正特征选择的方向，解决了高维不平衡数据分类任务中，不平衡特性影响特征选择的问题，得到了一个有利于提高少数类识别效果的特征子集。接着在希尔伯特空间中对边界少数类进行 SMOTE 过采样，并寻找过采样产生的少数类在对应欧几里得空间下的近似原像，然后用 PSO 算法自适应的寻找边界少数类的最优过采样倍率，提高 SVM 的分类效率，提出了一种 PBKS（PSO-Border-Kernel-SMOTE）过采样算法；PBKS 算法通过寻找希尔伯特空间中的样本点在欧几里得空间中的近似原像，解决了高维不平衡数据分类任务中，用基于 SVM 的包裹式特征选择方法得到较好的特征组合之后，产生的空间转换问题。最终参考国内外解决高维不平衡问题的主流方法，采用先解决高维问题再解决不平衡问题的思路，结合两部分算法，提出了一种基于 SVM 的高维不平衡数据二分类算法：BRFE-PBKS-SVM 算法。

第 4 章　实验结果与分析

本章首先通过实验证明在高维不平衡数据分类任务中，存在着高维问题和不平衡问题相互干扰所形成的新问题。然后利用 UCI 的数据集进行实验，和未改进之前的基于 SVM 的特征选择算法作对比，验证 SVM-BRFE 特征选择算法在解决高维不平衡数据分类任务中的不平衡问题的有效性；并且结合后续的采样过程，与普通采样过程作对比，验证 BRFE-PBKS-SVM 算法的有效性。最后，本章将 BRFE-PBKS-SVM 算法在 UCI 上的数据集取得的效果与一些现有的解决高维不平衡问题的方法作对比，证明 BRFE-PBKS-SVM 算法的优越性。

4.1 实验数据与参数设置

UCI 是一个著名的、公开的机器学习数据库，为使实验结果更具说服力，本章所有实验的数据集，均来源于 UCI。实验数据如表 4-1 所示。

表 4-1 实验数据集描述

No.	Data-sets	Instances	Features	%Min.
1	Detect Malacious Executable	373	513	28.57
2	SECOM	1567	591	6.64
3	Musk	6598	168	15.55
4	Heart Disease	267	44	20.60
5	Urban Land Cover	675	147	32.20
6	Multiple Features	1608	649	12.06

表 4-1 描述了所有实验所用数据集的具体属性，其中 No.列为数据集编号，Data-Set 为数据集名称，Instances 表示样本总数，Features 为数据集包含的属性数量，%Min 表示少数类样本所占比例。少数类所占的比例是训练集与验证集中所有少数类所占的比例，实际每一个数据集的训练集里，不平衡比例更高。

4.1.1 数据预处理

UCI 中获得的数据集，有不少都存在着属性值缺失、格式不一致、类别标签意义不明确等问题，因此不能直接交给算法模型训练，需要经过处理之后使

之变成“干净”的、可供训练的数据集。数据的清洗和处理往往对最终的模型效果的好坏至关重要。在本课题中，主要做的预处理步骤如下：

（1）类标转换

有时候数据集的类别标签是多于两个的，对于这种数据集，将其中一种类别当作少数类，而其他所有类别的数据都当作多数类。所以在进行实验室之前，根据少数类和多数类样本特性，数据集的类别标签需要进行转换。

（2）格式转换

由于本文所提出的算法，主要是针对 SVM 进行的一系列改进所得到的，而在本章的实验中，所采用的算法包是 libsvm-3.2.1 版本，训练集的格式需要转换为标准 SVM 格式：对于类别标签，分别用+1 表示少数类，-1 表示多数类；对于每个样本点的特征，用稀疏表示，具体地，假设第 i 个特征的值为 value，则对应该列属性应改写成 i:value，特征值为 0 的属性设置为空缺。

（3）缺失值填充

原始数据集中会有一些样本的属性值存在缺失的情况，主要存在以下两种情况的数据缺失：首先，对于连续型特征值的缺失，用平均值填补；其次，对于缺失的离散型特征，用该样本点对应的类别中该特征出现次数最多的值来填补。如果数据集存在类别标签缺失的情况，对于这些样本，采用直接删除的方法。

（4）归一化

在所用数据集中，每个属性的取值范围会千差万别，为了消除这种大数连续值对最终效果的影响和便于属性之间的数值计算方便，本实验中对所有数据集中的属性值做了归一化处理，如式(4-1)所示：

$$a^{new} = \frac{a_i - a_{\min}}{a_{\max} - a_{\min}} \tag{4-1}$$

这里，a^{new}是通过归一化得到的值，a_i是归一化之前的值，a_{max}为该属性中最大的取值，a_{min}为该属性中最小的取值。

（5）验证集划分

从 UCI 得到的这几个数据集中，有的数据直接划分好了训练数据集和验证数据集，对于这种类型的数据，可以通过上述 4 个预处理步骤之后，直接调用本文所提出的算法来训练模型，然后再用验证数据集来验证算法的效果；而有的数据没有给出验证数据集，对于这样的数据，则需要预先划分出验证集。在本章中，对于没有给出验证集的数据，将随机取出其中 20%的数据作为验证集，以便测试算法的效果。

4.1.2 算法参数设置

参数对于考究一个算法的有效性和优越性，是十分重要的因素；参数的调节，不仅关系到对训练数据过拟合、欠拟合的问题，还会影响在验证集上的实验结果，对是否能公平的衡量一个算法的分类结果起着决定性的作用。在libsvm-3.2.1 中，二分类支持向量机可调的参数主要有核函数类型、惩罚参数大小、核函数中的常数参数这 3 种类型，本章对所有的训练集，均采取 5 折交叉验证和网格法来寻找平均分类效率最好的参数。对于每个数据集，一旦最优参数确定，在后续的实验中将固定使用该最优参数。

在 5 折交叉验证下，6 个数据集的最优参数和最优参数下 SVM 对训练数据集的分类效果，如表 4-2 所示。

表 4-2 K 折交叉验证下最优参数及分类效果

No.	Data-sets	-c	-t	-d	-g	-r	ACC	F1
1	Detect Malacious Executable	3.272	0	/	0.0125	/	83.1%	0.866
2	SECOM	3.21	2	/	0.0625	/	67.6%	0.414
3	Musk	0.02	2	/	0.0059	/	84.3%	0.795
4	Heart Disease	0.018	2	/	0.0227	/	71.6%	0.504
5	Urban Land Cover	0.00003	0	/	0.0068	/	90.8%	0.860
6	Multiple Features	9.67	0	/	0.0015	/	88.4%	0.811

图中，-c 代表惩罚参数，即式(2-10)中的 C。-t 代表核函数选项：0 表示线性核，其核函数如式(2-34)所示；1 表示多项式核，其核函数见式(2-32)；2 代表高斯核，其核函数如式(2-33)所示。-d 代表多项式核函数的幂次，即式(2-32)中的 p。-g 代表式(2-33)中的高斯核函数中的 $1/-2\sigma^2$。-r 只有在多项式核中才有该参数，默认设置为 1。ACC 和 F1 分别代表 K 折交叉验证取得上述最优参数后，用最优参数训练完整训练集，并用完整训练集来测试该模型所得到的 ACC 值和 F1 值。这样做的好处是，我们在训练数据集的时候设置的参数，不至于过拟合训练集，最终能得到的模型是一个较稳定的 SVM 模型。

4.2 高维不平衡数据中新型问题的存在性验证及结果分析

本节将通过实验，来证明不平衡特性的存在会对特征选择过程造成影响，通过给出在同种特征选择算法下，考虑不平衡问题的情况下的特征选择路径、新特征选择算法的特征选择路径，对比这两种特征选择路径及其 F1 值，来说

明不平衡问题对特征选择过程的干扰；然后采用同种特征选择算法，分别比较它们在特征选择之前运用 SMOTE 过采样方法所取得的分类效果的提高程度和在特征选择之后运用 SMOTE 过采样方法所取得的分类效果的提高程度，以此来证明高维问题对解决不平衡问题的解决会造成影响。

4.2.1 验证不平衡问题干扰特征选择

本小节为了证明不平衡问题干扰特征选择过程的存在性，主要对 SVM-RFE 特征选择算法和经过改进的 SVM-BRFE 特征选择算法进行比较，观察其特征选择过程与最终的 ACC 值和 F1 值，以及分析造成其中的差异的原因。具体实验结果如表 4-3 所示。

表 4-3 SVM-RFE 与 SVM-BRFE 特征选择算法对比

No.	Data-sets	SVM-RFE			SVM-BRFE		
		Fea	ACC	F1	Fea	ACC	F1
1	Detect Malacious Executable	206	86.3%	0.882	288	**89.5%**	**0.893**
2	SECOM	134	83.4%	0.477	97	**84.0%**	**0.512**
3	Musk	6	**91.8%**	0.833	23	90.1%	**0.845**
4	Heart Disease	2	82.3%	0.536	13	**87.0%**	**0.696**
5	Urban Land Cover	123	90.1%	0.891	135	**93.2%**	**0.912**
6	Multiple Features	108	**90.0%**	0.780	125	88.3%	**0.835**

表中 Fea 表示选择剔除掉的特征个数，ACC 和 F1 分别表示全局准确率和 F1 值。从图中可以看到，SVM-RFE 算法在改进之后，所选择出的特征不同，由于特征子集太过庞大，所以不具体列出选择剔除掉的特征，经过使用改进之后的特征评价标准来对每一个特征评分之后，特征选择的路径也随之改变，这说明，在经过对边界少数类进行重采样之后，再进行特征选择，使得特征选择的过程在对少数类投入更多的关注下进行，最终的特征选择结果也会随之改变；经过考虑边界少数类的情况之后，可以看到表中的 F1 值均有提高，而全局准确率 ACC，基本上都有所提升，只有最后两个数据集，再改进了特征选择算法之后，它们的 F1 值有所提升，但 ACC 值却下降了，这说明了两个问题：

第一，由于在本文的特征选择过程中，选择子集的好坏是用剔除掉某一特征之后的F1 值来衡量，F1 值的提升表明算法模型对少数类的识别率有所增加，全局准确率的降低说明后两个数据以牺牲相对较少的多数类的识别率来增加了少数类的识别率。这也说明了评价方式对于衡量不平衡数据分类效率的重要性。

第二，F1 值的提升，说明了在改进了特征选择算法之后，不仅选择出的特征子集较改进前的特征选择不一样，而且最终在 SVM 上的分类效果，也有所提升，这说明了在高维不平衡数据中，确实存在着不平衡问题干扰特征选择过程的问题，当在特征选择的过程中消除或者削弱了该问题之后，所选择出的特征子集在同一参数下的 SVM 中，所取得的分类效果也获得了提升。

4.2.2 验证高维问题影响采样效果

在第二章中已经提过，由于高维不平衡数据中数据也同时展现出高维特性，导致样本的特征之间很有可能会有一些冗余的特征组合或者是对分类效果起到负面作用的特征组合，在先解决不平衡问题再解决高维问题的思路中，我们希望过采样所产生的样本点能够忠实的还原原始样本空间中的特征分布，即对分类效果有负面影响的特征组合，在新的样本点中，这些特征组合也应该对最终分类结果起到降低的效果，以便在后续特征选择的过程中剔除掉这些特征组合，但由于 SMOTE 过采样方法的随机性，所以无法做出保证。

本小节为了证明上述问题的存在，将 SVM-RFE 算法与 SMOTE 算法组合，设置了一组对比实验：对没有进行 SVM-RFE 特征选择的数据，观察其进行了 SMOTE 过采样之后与没有进行 SMOTE 过采样之前的分类效果的提升水平；对进行了 SVM-RFE 特征选择之后的数据，观察其进行了 SMOTE 过采样之后与没有进行 SMOTE 过采样之前的分类效果的提升水平；通过比较这两组实验的提升效果，来分析高维问题对数据采样的效果的影响。实验效果如图 4-1 和图 4-2 所示，其中横坐标代表 6 个数据集的标号：

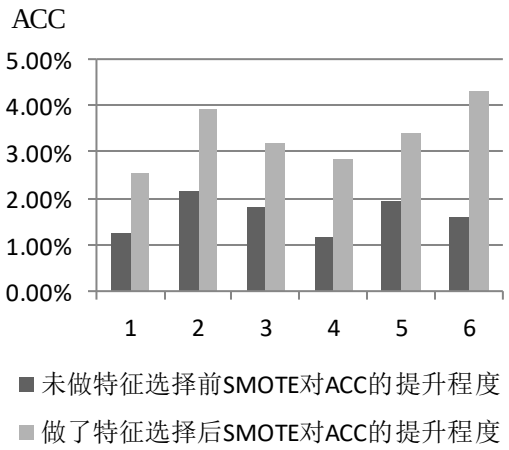

图 4-1 特征选择前后对 ACC 值的提升对比

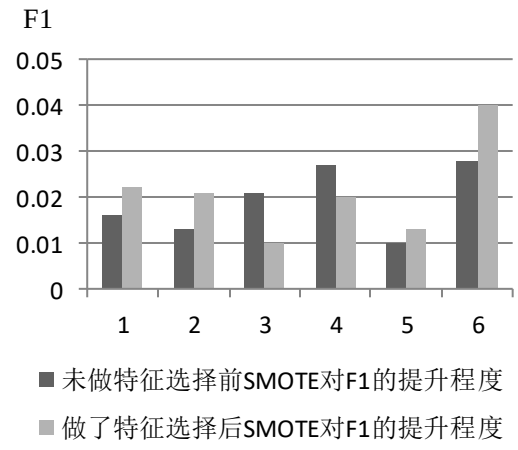

图 4-2 特征选择前后对 F1 值的提升对比

图 4-1 和图 4-2 在采用同一参数的情况下，分别从 F1 值和 ACC 值进行考

察，比较了 SMOTE 算法在特征选择前、后对分类效果的提升程度。从图中可以看到，在做了特征选择之后再对数据进行 SMOTE 过采样，其 F1 值和 ACC 值的提升，均比做特征选择之前的提升要大，但是准确率的最高提升幅度只有 4.2 个百分点，F1 值的最高提升幅度也只有 0.04，提升幅度都很小，因此可以得到以下两点结论：

第一，在高维不平衡数据集中，高维问题的存在确实会对 SMOTE 的采样效果有所影响，SMOTE 过采样由于引入了样本点之间的相关性而不是真实的还原少数类特征之间的关系，导致在大多数情况下未进行特征选择前进行 SMOTE 过采样所获得的提升效果不如做了特征选择之后进行 SMOTE 过采样所获得的提升效果好。

第二，在利用 SVM 解决高维不平衡数据分类任务时，SMOTE 过采样方法确实存在着弊端，导致提升效果不佳；SVM-RFE 特征选择算法虽然已被证明在单独存在高维问题的数据中能取得不错的效果，然而在同时存在高维问题和不平衡问题的数据中，也同样存在着缺陷，因此用 SVM-RFE 算法结合 SMOTE 过采样算法来解决高维不平衡问题所取得的效果不佳。

4.3 BRFE-PBKS-SVM 算法的有效性验证

BRFE-PBKS-SVM 算法分成两部分，第一部分是特征选择部分，第二部分是数据采样部分，通过将两部分结合，形成了一种专门针对解决高维不平衡数据分类问题的算法。在该算法中，后半部分所需要解决的，是运用基于 SVM 来解决高维不平衡数据分类任务中的不平衡问题之后，所产生的新问题。本节将利用第二章所提到的评价标准，分别比较 BRFE-PBKS-SVM 算法对少数类识别率的提高、总体效率的提高以及算法稳定性的对比。

4.3.1 原子标准验证

本小节将用第二章所提到的原子标准来判别 BRFE-PBKS-SVM 算法是否能提高少数类的识别率，主要用到的是精确率 Precision 和召回率 Recall，其计算公式见式(2-36)和式(2-37)。

精确率主要是指所有被算法模型判断为某一类型的样本中，被正确判别的样本所占据的比例，它反映了分类器对数据空间中某一类样本的分布的拟合程度。召回率主要反映的是某一类样本中，被正确识别出来的总数占该类所有样本总数的百分比，可以通过召回率来衡量算法改进之后，少数类的识别率是否

有所提高；在某一固定参数下，改进的算法使少数类样本的识别率的提升，往往是以牺牲一定量的多数类的识别率为代价的，只要保持整体 F1 值不断提升，就说明所换取到的少数类识别率，比牺牲掉的多数类的识别率更具有价值。整体 F1 值的公式如式(2-41)所示，它反映的是精确率与召回率的综合情况。

本小节用 4 种算法组合对本文所引用的 6 个数据集进行训练，并用验证集对算法模型进行测试，得到的多数类和少数类的精确率与召回率具体数值如表 4-4 和表 4-5 所示，其中表 4-4 是各种算法组合对少数类的精确率与召回率，表 4-5 是各种算法组合对多数类的精确率与召回率。图 4-3 到图 4-8 对应 4 种算法在 6 个数据集中对多数类的召回率、对少数类的召回率以及总体 F1 值的三值折线图对比，可以通过它们来直观的分析各个数据集中，不同算法的召回率和 F1 值变化情况。在图中，横坐标 1 到 4 分别按顺序表示表中的 4 种算法组合。

表 4-4 各算法少数类的精确率与召回率对比

No.	RFE-SMOTE-SVM		RFE-PBKS-SVM		BRFE-SMOTE-SVM		BRFE-PBKS-SVM	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
1	96.3%	82.7%	97.2%	86.8%	96.9%	89.6%	**98.0%**	**92.3%**
2	18.0%	89.4%	45.6%	**91.5%**	61.2%	84.3%	**90.0%**	69.2%
3	81.0%	83.6%	82.9%	93.7%	81.5%	85.1%	**86.9%**	**94.0%**
4	41.7%	**83.3%**	**91.7%**	55.0%	58.3%	77.8%	83.3%	71.4%
5	87.8%	92.6%	91.3%	90.2%	93.6%	92.0%	**97.1%**	**93.8%**
6	86.8%	87.7%	89.0%	86.0%	87.6%	**91.3%**	**91.5%**	90.4%

表 4-5 各算法多数类的精确率与召回率对比

No.	RFE-SMOTE-SVM		RFE-PBKS-SVM		BRFE-SMOTE-SVM		BRFE-PBKS-SVM	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
1	98.6%	94.6%	93.1%	93.1%	94.7%	94.7%	**95.4%**	**95.4%**
2	92.8%	91.6%	81.9%	80.0%	**95.4%**	**95.7%**	88.1%	86.7%
3	97.4%	97.4%	93.8%	94.5%	**97.6%**	**97.6%**	95.3%	94.5%
4	**97.6%**	**97.6%**	78.6%	78.6%	95.2%	95.2%	90.5%	90.5%
5	96.4%	94.9%	94.9%	94.9%	95.8%	95.8%	**96.7%**	**96.7%**
6	**96.2%**	**96.2%**	95.6%	95.6%	95.3%	95.3%	94.4%	94.4%

从表中 4-4 中可以看到，BRFE-PBKS-SVM 算法在 4 个算法中，对少数类都取得了最高的召回率，相比于未改进的 SMOTE 算法，PBKS 过采样算法对

少数类召回率的提升程度显著，并且随着少数类召回率的提升，其精确率有所下降；而从表 4-5 中则可以看到，BRFE-PBKS-SVM 算法对多数类的召回率几乎没有达到最高。这说明 BRFE-PBKS-SVM 算法对少数类的关注程度增加了，但对多数类的关注程度却降低了。

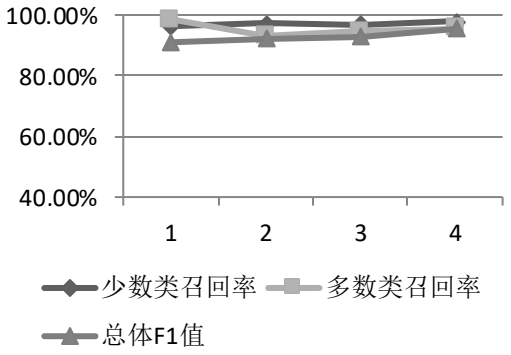

图 4-3 Detect Malacious Executable 三值折线图

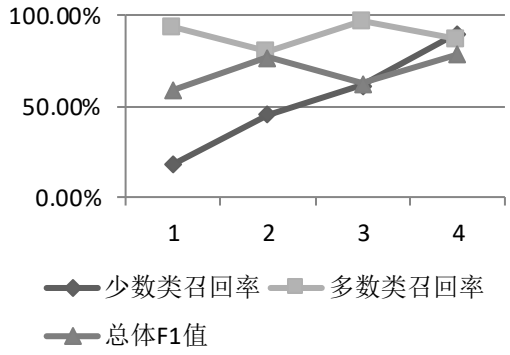

图 4-4 SECOM 三值折线图

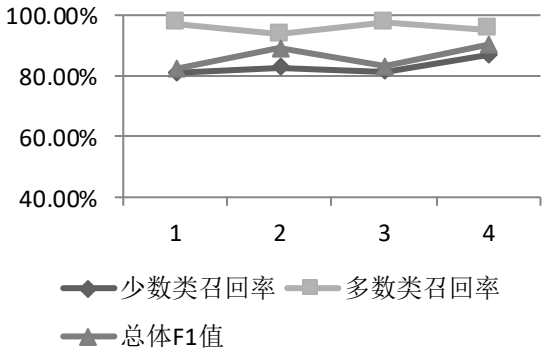

图 4-5 Musk 三值折线图

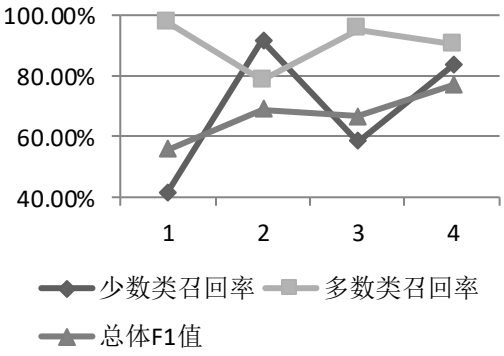

图 4-6 Heart Disease 三值折线图

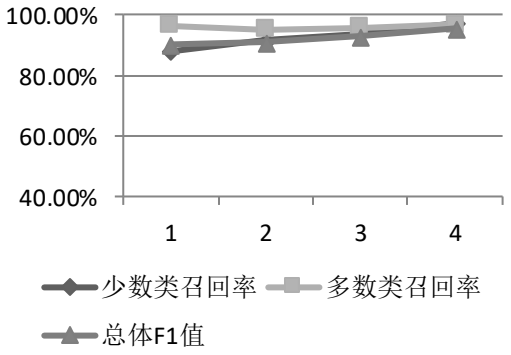

图 4-7 Urban Land Cover 三值折线图

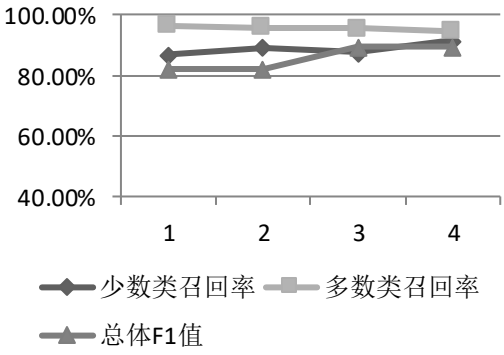

图 4-8 Multiple Features 三值折线图

从图 4-3 到图 4-8 中可以看到在每一个数据集中，4 种算法组合所对应的少数类召回率、多数类召回率以及总体 F1 值的变化趋势：

第一，在 6 个三值折线图中，每一个都对应 4 个算法组合在一个数据集上的效果图；所有 6 个图中的第 4 个算法组合的 F1 值、少数类召回率均比第 1 个算法组合的 F1 值、少数类召回率有所提高；而 6 个图中，有 3 个图的第 4 个算法的多数类的识别率和第 1 个算法的多数类识别率相比有所下降。上述趋势说明 BRFE-PBKS-SVM 在追求 F1 值的提升的同时，对少数类的识别水平也提高了，有效的提高了 SVM 对少数类的关注度，被误分的多数类的数量也随之增加，但通过 F1 值的变化情况来看，算法总体的分类效率是不断提升的。

第二，图 4-4 到图 4-6 中可以看到，对于第 2、3、4 个数据集来说，第 2 个算法 RFE-PBKS-SVM 的总体 F1 值比第 3 个算法 BRFE-SMOTE-SVM 要高，说明对于这三个数据集来说，不平衡问题对数据分类造成的影响比高维问题造成的影响大；从图 4-8 中，第 2 个算法和第 3 个算法对于少数类的召回率变化，也可以得到这个结论。

第三，在这 6 个图中，除了图 4-3 之外，BRFE-PBKS-SVM 算法对所有数据集的多数类召回率都不是最高的，且 6 个图中，就多数类的召回率来说，符合如下规律：同种过采样算法下，改进了特征选择算法之后，多数类的召回率会有所下降；同种特征选择算法下，改进了过采样算法，多数类的召回率会有所下降。这说明了 BRFE-PBKS-SVM 算法无论是在特征选择部分的改进，还是在过采样算法部分的改进，都是在追求 F1 值不断提高的前提下，牺牲尽量少的多数类识别率来换取尽量多的少数类识别率。

通过以上原子评价标准的三值折线图分析，说明经过改进后的 BRFE-PBKS-SVM 算法对少数类识别率的提高确实有效，少数类的召回率随着算法的改进不断提升。

4.3.2 复合标准验证

本小节通过使用第三章中所提到的全局准确率 ACC 值以及复合标准 F1 值来全面验证 BRFE-PBKS-SVM 算法对于高维不平衡数据分类效果的改善，并与基于 SVM 的已有特征选择算法和 SMOTE 过采样算法的组合算法进行算法组合，对实验效果进行对比、分析，以验证 BRFE-PBKS-SVM 算法的总体有效性。本节将要采用的算法组合是 SVM-RFE 与 SMOTE 组合、SVM-RFE 与 PSO-Border-Kernel-SMOTE 组合、SVM-BRFE 与 SMOTE 组合、以及 BRFE-PBKS-SVM。

表 4-6 各算法 F1 值与 ACC 值对比

No.	RFE-SMOTE-SVM		RFE-PBKS-SVM		BRFE-SMOTE-SVM		BRFE-PBKS-SVM	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
1	91.4%	0.922	94.8%	0.924	**97.8%**	0.926	96.9%	**0.943**
2	81.8%	0.593	83.9%	0.786	84.2%	0.619	**84.6%**	**0.790**
3	95.0%	0.823	96.1%	0.889	95.3%	0.832	**96.9%**	**0.903**
4	85.2%	0.556	81.5%	0.688	87.0%	0.667	**88.9%**	**0.769**
5	93.5%	0.901	93.7%	0.908	95.1%	0.928	**96.8%**	**0.954**
6	94.3%	0.820	**94.3%**	0.820	93.8%	0.894	93.8%	**0.894**

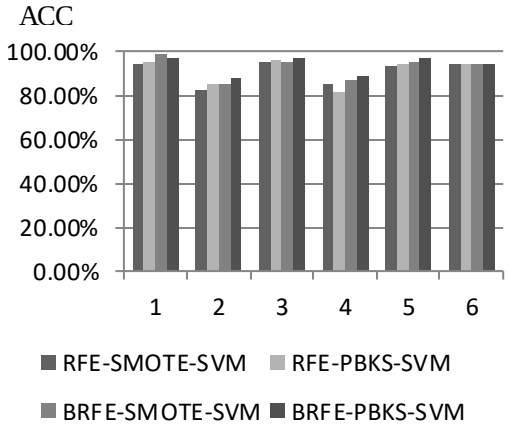

图 4-9 各算法 ACC 直方图

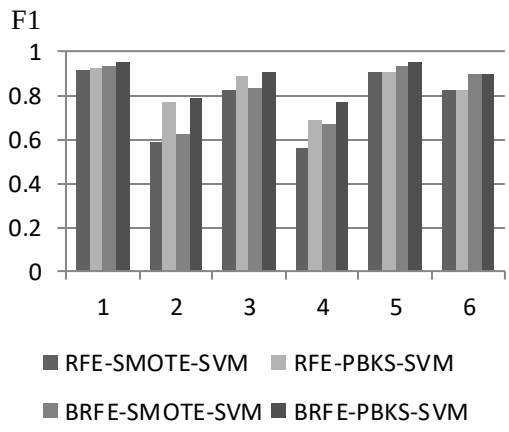

图 4-10 各算法 F1 值直方图

表 4-6 中，通过第二列和第四列的对比、第六列和第八列的对比，可以看出普通的 SMOTE 过采样方法与 PBKS 过采样方法在 SVM 中的 ACC 值效果对比；通过第二列和第六列的比较、第四列和第八列的比较，可以看到 SVM-RFE 特征选择算法与 SVM-BRFE 特征选择算法的效果对比。

图 4-9 和 4-10 的横坐标代表 6 个数据集。从图 4-9 中可以直观的看到在用 ACC 值衡量分类性能的情况下，各算法的效果展示。从图 4-9 中可以看出，就全局准确率 ACC 来说，在第 2 到第 5 个数据集中，BRFE-PBKS-SVM 算法在所有算法组合里，是最优的；而在采用同样的过采样算法的情况下，经过改进的 BRFE 特征选择算法组合所取得的效果最好，这是因为 BRFE 特征选择算法在特征消除的过程中考虑了不平衡问题;在采用同样的特征选择算法的情况下，改进的 PBKS 过采样算法组合所取得的效果最好，这是因为它们都是在多项式核函数或者高斯核函数对应的希尔伯特空间下训练的数据，由于 PBKS 算法过采样产生的样本点能更好的填充希尔伯特空间下的边界，空间上分布更合理，

因此能使得分类效果提升较多。

但是在第一个和第五个数据中，就 ACC 值来说，BRFE-PBKS-SVM 算法却不是最优的算法，其原因如下：

第一，对于不平衡数据来说，F1 值对衡量算法效率的公平性，比 ACC 值更好，所以本文所提出的算法，均是在追求 F1 值最优化的前提下驱动进行的，而 F1 值的最优，不一定能使 ACC 值提升，就不平衡数据分类来说，F1 值的提升有可能是牺牲了部分多数类的正确识别来换取一些更有价值的少数类的识别，因此 BRFE-PBKS-SVM 算法的全局准确率 ACC 值在第一和第五个数据中不是最高的。

第二，表 4-2 给出了 6 个数据集在五折交叉验证情况下的最优参数，其中第一和第五个数据集选择线性核函数能取得最好效果，线性核函数的公式如式(2-34)所示，其内积与欧几里得空间的内积定义一致，训练空间与输入空间都是欧几里得空间，因此未改进的 SMOTE 算法与 PBKS 过采样算法在空间上等价，但由于 PBKS 过采样算法比 SMOTE 过采样算法引入了更强的约束，所以产生的样本点更稳定。

通过第三列和第五列、第七列和第九列的对比，可以看到在特征选择算法改进前后，运用同种过采样方法的 F1 值变化情况，而第三列和第七列、第五列和第九列的对比，则反映了在运用同种特征选择算法的情况下，过采样方法改进前后的效果情况。

从表 4-6 中可以看出，就 F1 值来说，在所有六组数据中，在运用同种特征选择算法的情形下，采用改进前的 SMOTE 过采样算法和改进后的过采样算法，其效果有所不同，改进后的 PBKS 过采样算法所提高的效果均比改进前所提升的效果有所增加。

观察分析图 4-10 的各算法 F1 值的直方图对比，发现所有算法中，经过改进后的 BRFE-PBKS-SVM 算法所取得的 F1 值是最佳的，说明再以最优化 F1 值为目标的追求下，改进自 SVM-RFE 算法和 SMOTE 过采样算法的 BRFE-PBKS-SVM 算法，能对分类效果有一定的提升。观察最后一个数据，可以看到在同种特征选择算法的前提下，采用改进前的 SMOTE 过采样算法和改进后的 PBKS 过采样算法，其效果并没有提升，正如上文所分析的，这也是因为第六个数据选择的核函数是线性核函数，它对数据的输入空间和训练空间都是欧几里得空间，在欧几里得空间下，PBKS 过采样算法在空间上等价；而在第五个数据中，PBKS 过采样算法在用经过改进的特征选择算法进行了特征选择之后，尽管训练空间都是欧几里得空间，但其效果却比运用未经改进的

SMOTE 特征选择算法要好，通过观察试验中 PBKS 过采样产生的样本点并结合这个分类结果，发现 PBKS 过采样算法产生的样本点集合方差较小，分布较集中，说明 PBKS 过采样算法比 SMOTE 过采样算法更为稳定。

4.3.3 ROC 曲线及 AUC 值验证

第二章中所提到的 ROC 曲线是目前常用于评价算法模型好坏的一种曲线，ROC 曲线所围成的面积值 AUC 的最大值为 1；AUC 值越大，表示相应的分类器越好，稳定性越高。在本小节中，分别画出了在六个数据集中，四种算法组合的 ROC 曲线，并计算出了其对应的 AUC 值以及各个算法组合在不同数据集上的 AUC 值的直方图，其 AUC 值如表 4-7 所示，其中横坐标代表 6 个数据集：

表 4-7 各种算法组合在各个数据集上的 AUC 值

No.	AUC			
	RFE-SMOTE-SVM	RFE-PBKS-SVM	BRFE-SMOTE-SVM	BRFE-PBKS-SVM
1	0.981	0.977	0.983	**0.985**
2	0.834	0.893	**0.901**	0.897
3	0.964	0.955	0.964	**0.964**
4	0.853	**0.897**	0.879	0.891
5	0.980	0.983	0.987	**0.993**
6	0.986	0.988	0.988	**0.989**

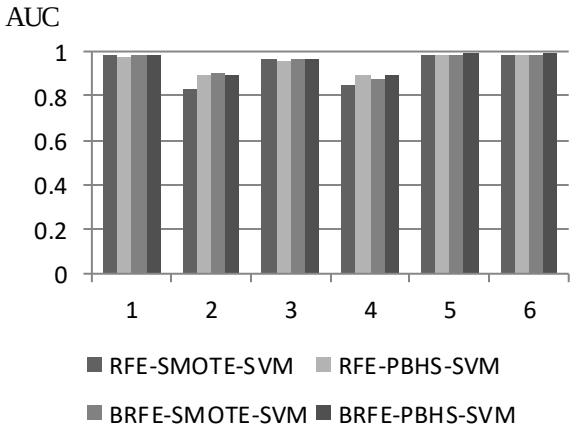

图 4-11 各算法 AUC 值直方图

结合表 4-7 和图 4-11 分析四种算法的 ROC 曲线的面积，发现在六组数据中，除了第二个和第四个数据外，BRFE-PBKS-SVM 算法都能取得最大的 AUC 值，而在第四个数据集中，即使改进后的算法没能取得最优的 AUC 值，其差

值也只有 0.006，总体上说明算法 BRFE-PBKS-SVM 有着良好的稳定性。图 4-11 显示了 4 种基于 SVM 的算法组合在各个数据集中的 AUC 值均相差不大，这也从侧面证明了 SVM 对完成高维不平衡数据的分类任务有着较好的稳定性以及优越性。

从图 4-12 到图 4-17 中，线条围起的面积即表 4-7 中的 AUC 值。对角线表示的是一个最差的分类效果水平，它对应的 AUC 值是 0.5，当一个分类器在某个数据集上的 ROC 曲线位于这条对角线之下时，它的 AUC 值将小于 0.5，这将意味着分类器在该数据集上的分类效率不如一个随机猜测的分类器效果好。ROC 曲线越趋向于左上方，代表相应的算法的效果越显著，AUC 值越接近于 1；例如图 4-16 中，算法 BRFE-PBKS-SVM 在第五个数据集上的 ROC 曲线，从表 4-7 可知，该曲线对应的 AUC 值为 0.993。

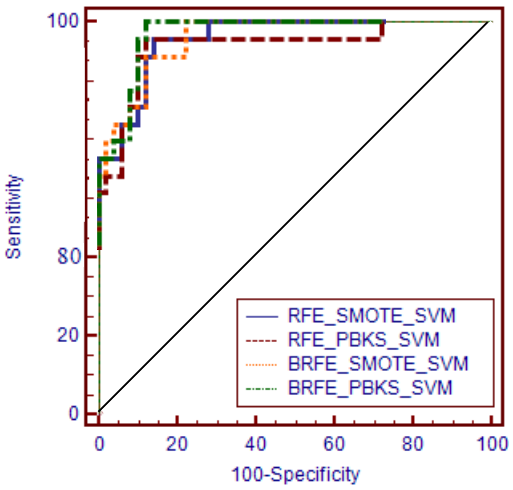

图 4-12 Detect Malacious Executable 的 ROC 曲线

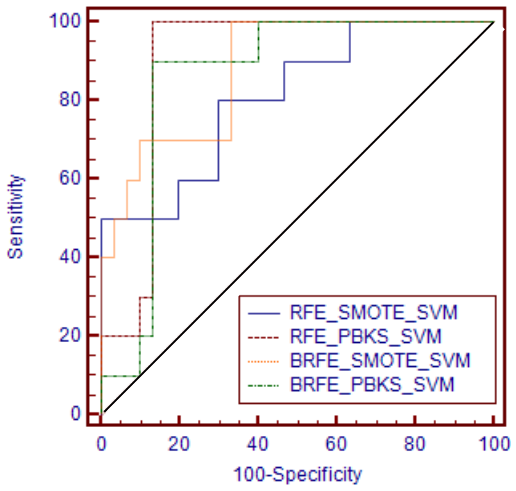

图 4-13 SECOM 的 ROC 曲线

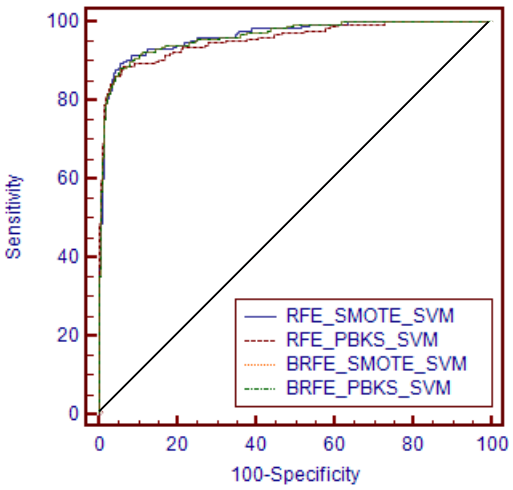

图 4-14 Musk 的 ROC 曲线

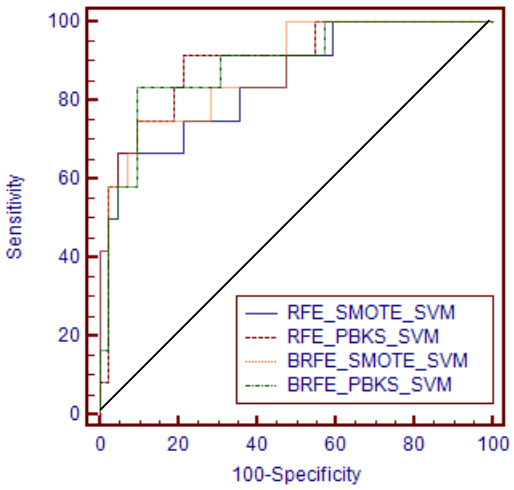

图 4-15 Heart Disease 的 ROC 曲线

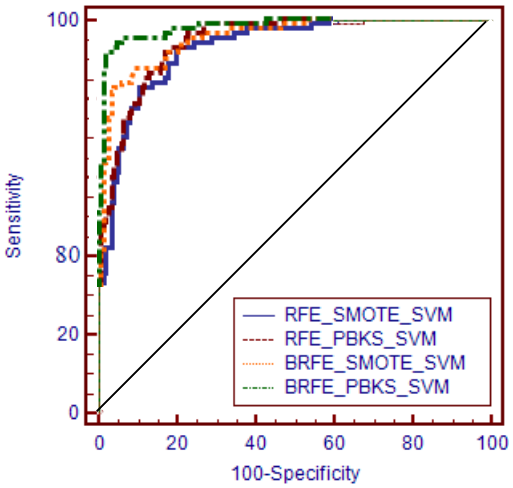

图 4-16 Urban Land Cover 的 ROC 曲线

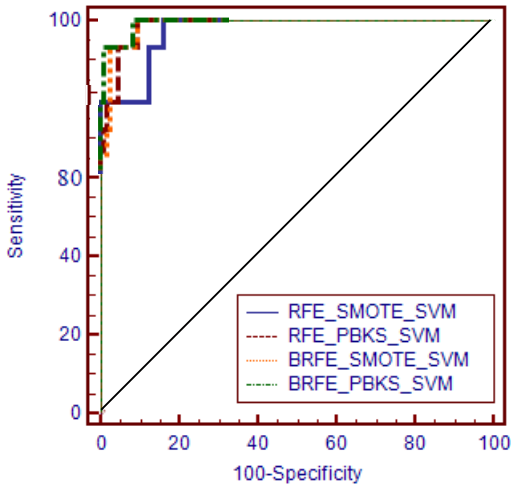

图 4-17 Multiple Features 的 ROC 曲线

从实验得到的六个 ROC 曲线图中发现，除了第二个和第四个数据集之外，在剩下的数据集里，四种算法所围成的面积相差均不大，都能取得较好的效果，并且 BRFE-PBKS-SVM 算法都能在这四个数据集中取得最大的 AUC 值；而在第二和第四个数据集中，四种算法效果差异性较大，并且 ROC 曲线极度不平滑，BRFE-PBKS-SVM 算法也没能取得最佳的分类效果，但与分类效果最好的算法的 AUC 值相差并不大，且都能取得较随机分类器好的 ROC 面积。这说明，基于 SVM 的针对高维不平衡数据分类任务的 BRFE-PBKS-SVM 算法，能稳定有效的完成高维不平衡数据的分类任务，并能取得可观的效果。

4.4 本章小结

本章基于第 3 章的基础研究工作，为防止过拟合问题首先进行 K 折交叉验证寻找最佳的参数设置，为了实验结果的可比性和公平性，在后续的实验中的每一个数据集均采用各自的最优参数作为其固定参数。

然后，本章就不平衡问题与高维问题相互干扰形成的新问题的存在性做了验证，通过实验证明了在高维不平衡数据中，存在着第 2 章所论述的不平衡问题影响特征选择过程的问题和高维问题影响过采样效果的问题。

最后，本章就第 3 章提出的解决高维不平衡问题的算法 BRFE-PBKS-SVM 算法的有效性进行了验证，将它与其改进的原算法做了实验对比，通过考察对比各个方面的试验指标，证明了 BRFE-PBKS-SVM 算法在处理高维不平数据分类问题中的优良性能。

结 论

高维不平衡数据作为一种新兴数据类型，在近几年吸引了不少学者的关注和研究。在高维不平衡数据分类任务中，现在绝大多数的研究都是将其分成互不相连的两部分单独解决，并没有考虑到彼此之间的关联和影响。近几年的文献中，高维不平衡数据分类问题，一般都是采用先做特征选择再进行采样的思路来解决，但是在特征选择的过程中，由于特征选择的过程是单独考虑解决高维问题，并没有考虑数据的不平衡特性对高维问题所产生的影响，而在后续的采样过程中，也没有考虑数据的高维特性对不平衡问题的解决所带来的影响。针对高维不平衡数据的特点，本课题通过在特征选择部分和数据采样部分进行了相应的改进，提出了针对高维不平衡数据分类任务的解决方案：

（1）在特征选择部分，提出了一种能充分考虑数据不平衡特性的改进的特征选择算法：SVM-BRFE。通过利用 SVM 在希尔伯特空间中自动划分边界样本的特性，在每一轮特征选择中，多次对边界中的误分少数类样本进行单倍率重采样，以调整每一个特征的权值评分，使得对有利于提升少数类识别率的特征组合的权值有所提高。同时在每一轮的特征选择的过程中剔除一个最有利于提高分类效果并且特征评分尽量最低的特征，使得分类效果逐步提升。本课题通过实验，对比了原始 SVM-RFE 算法和改进后的算法 SVM-BRFE 在特征选择上的性能，通过对比其分类效果证明了高维不平衡数据中存在着数据不平衡特性干扰特征选择的情况和 SVM-BRFE 算法的有效性。

（2）在数据采样部分，针对 SVM 的训练空间和输入空间不一致的问题，进行了如下改进：在 SVM 的训练空间，即希尔伯特空间中，对 SVM 自动划分的边界中的少数类进行 SMOTE 过采样，然后寻找过采样产生的样本点在原始欧几里得空间中的近似原像，并利用 PSO 算法自适应的寻找少数类样本的最优过采样倍率。通过这些改进提出了一种基于 PSO 的核边界 SMOTE 过采样算法 PBKS 算法，并通过实验验证了 PBKS 算法在数据采样部分的有效性。最后将两部分算法组合，通过和其他算法对比，验证了 BRFE-PBKS-SVM 算法在高维不平衡数据分类任务中的有效性。

通过分析不平衡数据特点以及实验的结果，本课题所提出的方法还存以下待改善之处：

（1）在特征选择的过程中，由于每一轮迭代搜索只选择一个特征，因此耗时比较庞大。在后续的研究中会重点考虑提升算法时间开销问题。

（2）此方法是只针对 SVM 的包裹式算法，后续的工作应考虑是否能将该算法的思路扩展，使其更具有普适性。

参考文献

[1] Vandenberghe R, Nelissen N, Salmon E, et al. Binary Classification of F-flutemetamol PET Using Machine Learning: Comparison with Visual Reads and Structural MRI[J]. NeuroImage, 2013, 64:517-525.

[2] Provost F. Machine Learning from Imbalanced Data Sets 101 (Extended Abstract)[C]// Soft Computing and Pattern Recognition (SoCPaR), 2011 International Conference of, IEEE, 2015:435-439.

[3] Huang Y, Hung C, Jiau H. Evaluation of Neural Networks and Data Mining Methods on A Credit Assessment Task for Class Imbalance Problem[J]. Nonlinear Analysis: Real World Applications, 2006, 18:720-747.

[4] Bühlmann P, Geer SVD. Statistics for High-Dimensional Data[J]. Springer, 2011, 39(10):1-1.

[5] Guo B, Damper R I, Gunn S R, et al. A fast separability-based feature-selection method for high-dimensional remotely sensed image classification[J]. Pattern Recognition, 2008, 41(41):1653-1662.

[6] Yu L, Liu H. Efficiently handling feature redundancy in high-dimensional data.[C]// ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2003:685-690.

[7] Yu H, Ni J. An Improved Ensemble Learning Method for Classifying High-Dimensional and Imbalanced Biomedicine Data[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, IEEE, 2014, 11(4):657-666.

[8] Dittman D J, Khoshgoftaar T M, Napolitano A. Selecting the Appropriate Data Sampling Approach for Imbalanced and High-Dimensional Bioinformatics Datasets[C]// IEEE International Conference on Bioinformatics and Bioengineering. IEEE, 2014:304-310.

[9] García V, Sánchez J S, Mollineda R A. Classification of High Dimensional and Imbalanced Hyperspectral Imagery Data[C]// Iberian Conference on Pattern Recognition and Image Analysis. Verlag Berlin: Springer, 2011:644-651.

[10] Yin L, Leong T. A Model Driven Approach to Imbalanced Data Sampling in Medical Decision Making[J]. Study Health Technology Information, 2010, 89:856-860.

[11] Kubat M, Holte R C, Matwin S. Machine Learning for The Detection of Oil Spills in Satellite Radar Images[J]. Machine Learning, 1998, 78:195-215.

[12] Jung S, Marron J S. PCA consistency in High Dimension, Low Sample Size

context[J]. Annals of Statistics, 2009, 37(6):4104-4130.

[13] Zhuang X S, Dai D Q. Improved discriminate analysis for high-dimensional data and its application to face recognition[J]. Pattern Recognition, 2007, 40(5):1570-1578.

[14] Muhammad Arif. Similarity-Dissimilarity Plot for Visualization of High Dimensional Data in Biomedical Pattern Classification[J]. Journal of Medical Systems,2012,36(3):1173-1181.

[15] Maryam Imani,Hassan Ghassemian. Binary coding based feature extraction in remote sensing high dimensional data[J]. Information Sciences,2016,342(C):191-208.

[16] Bharat Singh,Nidhi Kushwaha,Om Prakash Vyas. A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty[J]. Journal of Data Analysis and Information Processing,2014,02(4):95-105.

[17] Narissara Eiamkanitchat,Nipon Theera-Umpon,Sansanee Auephanwiriyakul,Chunlin Chen. On Feature Selection and Rule Extraction for High Dimensional Data: A Case of Diffuse Large B-Cell Lymphomas Microarrays Classification[J]. Mathematical Problems in Engineering,2015(9):1-12.

[18] Lin W J, Chen J J. Class-imbalanced classifiers for high-dimensional data.[J]. Briefings in Bioinformatics, 2013, 14(1):13-26.

[19] Deegalla S, Bostrom H. Reducing High-Dimensional Data by Principal Component Analysis vs. Random Projection for Nearest Neighbor Classification[C]// International Conference on Machine Learning and Applications, IEEE Computer Society, 2014:245-250.

[20] Hulse J V, Khoshgoftaar T M, Napolitano A, et al. Feature Selection with High-Dimensional Imbalanced Data[C]// IEEE International Conference on Data Mining Workshops, IEEE Computer Society, 2009:507-514.

[21] Blagus R, Lusa L. Evaluation of SMOTE for High-Dimensional Class-Imbalanced Microarray Data[C]//International Conference on Machine Learning and Applications, IEEE, 2012:89 - 94.

[22] 尹华. 面向高维和不平衡数据分类的集成学习研究[D].武汉大学,2012.

[23] Zhang Chunkai, Jia Pengfei. DBBoost: Enhancing Imbalanced Classification by a Novel Ensemble Based Technique[C]//International Conference on Medical Biometrics, IEEE, 2014:210-215.

[24] Kubat M, Matwin S. Addressing the Curse of Imbalanced Training Sets: One-sided Selection Sampling[C]//Proceedings of 14th International Conference on Machine Learning, ICML, 1997:179-186.

[25] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002,

43: 321-357.

[26] Han H, Wang W, Mao B. Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Sets Learning[C]//Proceedings of International Conference on Intelligent Computing, Berlin Heidelberg: Springer, 2005:878-887.

[27] Yen S, Lee Y. Cluster-based Under-sampling Approaches for Imbalanced Data Distributions[J]. Experts Systems with Applications, 2009, 36:5718-5727.

[28] Lópeza V, Fernándezb A, Moreno-Torresa J, Herreraa F. Analysis of Preprocessing vs Cost-sensitive Learning for Imbalanced Classification. Open Problems on Intrinsic Data Characteristics[J]. Expert Systems with Applications, 2012, 53:6585-6608.

[29] Polikar R. Ensemble Based Systems in Decision Making[J]. IEEE Circuits and Systems Magazine, 2006, 31:21-44.

[30] 谢纪刚,裘正定. 非平衡数据集 Fisher 线性判别模型[J]. 北京交通大学学报,2006,05:15-18.

[31] Yin H, Gai K. An Empirical Study on Preprocessing High-Dimensional Class-Imbalanced Data for Classification[C]// IEEE, International Conference on High PERFORMANCE Computing and Communications. IEEE, 2015:24-26.

[32] Gashler M, Martinez T. Temporal nonlinear dimensionality reduction[J]. 2011:1959-1966.

[33] Kira K, Rendell L A. The feature selection problem: traditional methods and a new algorithm[C]// Tenth National Conference on Artificial Intelligence, San Jose, California:AAAI Press, 1992:129-134.

[34] Liu H, Setiono R. Feature Selection And Classification - A Probabilistic Wrapper Approach[C]// International Conference on Industrial & Engineering Applications of Ai & Es. IEEE, 1997:419-424.

[35] Tibshirani R. Regression Shrinkage and Selection via the Lasso[J]. J of the Royal Statistical Society, 1996, 58(1):267-288.

[36] Cortes C, Vapnik V. Support-Vector Networks.[J]. Machine Learning, 1995, 20(3):273-297.

[37] Guyon I, Weston J, Barnhill S, et al. Gene Selection for Cancer Classification using Support Vector Machines[J]. Machine Learning, 2002, 46(1-3):389-422.

[38] Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data[J]. Bmc Bioinformatics, 2013, 14(1):1-16.

[39] Maldonado S, Weber R, Famili F. Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines[J]. Information Sciences, 2014, 286:228-246.

[40] Yin L, Ge Y, Xiao K, et al. Feature selection for high-dimensional imbalanced data[J]. Neurocomputing, 2013, 105(3):3-11.

[41] Liu Y, Chen Y. Face Recognition Using Total Margin-based Adaptive Fuzzy Support Vector Machines[J]. Neural Networks, 2007, 18:178-192.
[42] Qi L, Jiang H. Semismooth Karush-Kuhn-Tucker Equations and Convergence Analysis of Newton and Quasi-Newton Methods for Solving these Equations[J]. Mathematics of Operations Research, 1997, 22(2):301-325.
[43] Kuo T F, Yajima Y. Ranking and selecting terms for text categorization via SVM discriminate boundary[J]. International Journal of Intelligent Systems, 2010, 25(2):137–154.
[44] Boser B E, Guyon I M, Vapnik V N. A Training Algorithm for Optimal Margin Classifiers[J]. Proceedings of Annual Acm Workshop on Computational Learning Theory, 1996, 5:144-152.
[45] Lu J, Zhang C, Shi F. A Classification Method of Imbalanced Data Base on PSO Algorithm[C]//ICYCSEE, Singapore: Springer, Harbin, 2016:121-134.
[46] Wu C, Wang X, Bai D, et al. Fast Incremental Learning Algorithm of SVM on KKT Conditions[C]//International Conference on Fuzzy Systems & Knowledge Discovery. IEEE Computer Society, 2009:551-554.
[47] Vapnik V, Levin E, Cun Y L. Measuring the VC-dimension of a learning machine[J]. Neural Computation, 1994, 6(5):851-876.
[48] 薛毅. 支持向量机与数学规划[D]. 北京工业大学, 2003.
[49] Fawcett T. An Introduction to ROC Analysis[J]. Pattern Recognition Letters, 2006, 27(8):861-874.
[50] He H, Garcia E. Learning From Imbalanced Data[J]. Knowledge and Data Engineering, 2009, 21(9):1263-1284.
[51] Delong E R, Delong D M, Clarke-Pearson D L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.[J]. Biometrics, 1988, 44(3):837-45.
[52] 张新建. 再生核的理论与应用[M]. 科学出版社, 2010.
[53] Kwok T Y, Tsang W H. The pre-image problem in Kernel methods[J]. IEEE Transactions on Neural Networks, 2004, 15(6):1517-25.
[54] Kennedy J, Eberhart R. Particle swarm optimization[C]// IEEE International Conference on Neural Networks, IEEE. 1995:1942-1948.

攻读学位期间发表的学术论文

[1] Lu J, Zhang C, Shi F. A Classification Method of Imbalanced Data Base on PSO Algorithm[C]//ICYCSEE, Singapore: Springer, Harbin, 2016:121-134.

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《基于支持向量机的高维不平衡数据二分类方法的研究》，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

两年半的时间转瞬即逝，在硕士阶段的时间里，我收获颇多，在此论文完成之际，我要向我的母校、导师以及一直给予我帮助的同学朋友们表示真心的感谢。

首先我要感谢我的导师张春慨教授，从本课题的选题、开题、中期检查以及最后的定稿与答辩，张老师都给予了我细心的指导。张老师治学严谨、为人谦逊，在论文完成期间，给予了不少意见；生活上，更是以一位长辈的身份对我悉心照料，使我收获颇丰。

其次，我要感谢我的父母。感谢父母一直在生活上和精神上给予我的支持和鼓励，父母的一路陪伴，给了我克服困难的决心和勇气，使我在前行的路上不断进步。

接着，我要感谢实验室的同窗好友及师兄弟们，他们在我这两年半的学习生活中给予了不少帮助。在课题的研究以及难点的突破中，他们给予了我中肯的建议，开拓了我的思路。在日常的生活中，他们带来的快乐和轻松的气氛，使我能乐观的面对一切挑战与困难。

最后我要感谢我的母校——哈尔滨工业大学，在这两年半的生涯里，母校浓浓的科研氛围无时不刻不在熏陶着我。“规格严格，功夫到家”的教诲，将永远激励着我前进。