浙江工业大学

硕士学位论文

融合人的认知模型的模糊支持向量机算法及其应用

作者姓名　　赵丽丽

指导教师　　赵云波　教授

学科专业　　控制工程

学位类型　　工程硕士

培养类别　　全日制专业型硕士

所在学院　　信息工程学院

提交日期：2022 年 1 月

Fuzzy Support Vector Machine Algorithm Based on Human Cognitive Model and Application

Dissertation Submitted to

Zhejiang University of Technology

in partial fulfillment of the requirement

for the degree of

Master of Engineering

by

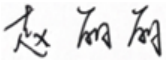
Lili-ZHAO

Dissertation Supervisor: Yunbo-ZHAO

Jan., 2022

浙江工业大学学位论文原创性声明

本人郑重声明：所提交的学位论文是本人在导师的指导下，独立进行研究工作所取得的研究成果。除文中已经加以标注引用的内容外，本论文不包含其他个人或集体已经发表或撰写过的研究成果，也不含为获得浙江工业大学或其它教育机构的学位证书而使用过的材料。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人承担本声明的法律责任。

作者签名：  日期：2021 年 12 月

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权浙江工业大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 1、保密□，在一年解密后适用本授权书。

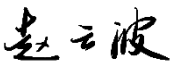
2、保密□，在二年解密后适用本授权书。

3、保密□，在三年解密后适用本授权书。

4、不保密□。

（请在以上相应方框内打“√”）

作者签名：  日期：2021 年 12 月

导师签名：  日期：2021 年 12 月

中图分类号　TP181　　学校代码　10337

UDC　681.5　　密级　公开

研究生类别　全日制专业型硕士研究生

浙江工业大学

工程硕士学位论文

融合人的认知模型的模糊支持向量机算法及其应用

Fuzzy Support Vector Machine Algorithm Based on Human Cognitive Model and Application

作者姓名　赵丽丽　　第一导师　赵云波教授

学位类型　工程硕士

学科专业　控制工程　　培养单位　信息工程学院

研究方向　人工智能

答辩日期：　2021　年　12　月　8　日

融合人的认知模型的模糊支持向量机算法及其应用

摘　要

不平衡分类是一种常见的分类问题，见于医疗诊断、风险预测、残次品挑选等广泛的领域。在将支持向量机（Support Vector Machine，SVM）方法应用于不平衡分类问题时，需要保证包含重要信息的正样本在训练过程中的受重视程度，克服数据集的不平衡对分类边界的影响，从而提升分类精度。

近年来，研究者提出通过机器学习再现人类概念学习能力的研究思路，以此试图缩小机器学习与人类学习之间的差距。已有研究表明，基于人类某些认知特性的神经网络方法在乳腺癌诊断预防中具有突出的表现。

本文基于人引导机器学习的思路，将人的认知模型与支持向量机算法相融合，构成一种新的机器学习方法。针对二分类任务的不平衡问题，设计了一种融合人的某种认知偏差模型的模糊支持向量机算法。通过实验证明，这种新方法在现有的公开的不平衡数据集中取得了更好的分类精度，其有效性在银行信用风险评估这一实际场景应用中进一步得到验证。本文的主要工作有以下方面：

（1）对比分析了支持向量机方法和人面对不平衡问题的不同表现能力。阐明不平衡分类问题影响支持向量机分类性能的因素，通过实验对比说明不平衡数据集的类间不平衡中数据的绝对不平衡和相对不平衡对支持向量机的影响程度；总结适用于解决不平衡分类问题的机器学习方法的评价指标。通过人体生理认知特征和心理认知特征两个维度，说明人类在处理不平衡分类问题方面的突出能力，并列举已有研究证明将这些特征转移到机器学习领域确实提高了原有算法的能力。

（2）设计了一种融合人的认知模型的模糊隶属度方法，将融合人的认知模型的模糊隶属度计算规则应用到支持向量机算法中，得到融合人的认知模型的模糊支持向量机算法(Fuzzy membership support vector machine based on cognitive bias，FSVM-BS)。利用人在特定场景下的某种认知特性及其模型，与 K 近邻算法结合对不平衡数据集中的样本进行分析、判断，针对其中的“负样本”进行模糊化处理，降低量大的负样本在 FSVM-BS 模型训练过程中的权重，使算法更重视对少量的正样本的特征学习，有效调整分类边界。利用现有的公开的二分类不平衡数据集对 FSVM-BS 算法进行验证，比较 FSVM-BS 算法与目前突出的

SVM 改进算法的分类性能，使用用于不平衡问题的评价指标对新算法和传统改进算法的测试效果进行分析评价。

（3）选取银行信用风险预测这一实际场景的数据集检验了提出的 FSVM-BS 算法在真实环境下的可行性和有效性。对银行信用风险评估的研究背景和现状进行分析，信用评估的相关数据集中信用“好”的样本明显多于“坏”的样本的特点，符合不平衡分类问题要求。其次，借助随机森林算法对数据集中的特征进行重要性排序，选取相关性更强的特征进行 FSVM-BS 算法训练，训练出来的模型与其他算法相比，有更高的 ROC 线下面积值，同时 FSVM-BS 算法模型在训练集和测试集间准确度的稳定性优于其他三种方法，证明了本文提出的算法 FSVM-BS 在实际应用中的可行性和有效性。

关键词：不平衡分类问题，人的认知特性，支持向量机，银行信用风险预测

FUZZY SUPPORT VECTOR MACHINE ALGORITHM BASED ON HUMAN COGNITIVE MODEL AND APPLICATION

ABSTRACT

Unbalanced classification is a common classification problem in medical diagnosis, risk prediction, defective product selection and other fields. When the Support Vector Machine (SVM) method is applied to the unbalanced classification problem, it is necessary to ensure that the positive samples containing important information are paid attention to in the training process, to overcome the impact of the unbalanced data set on the classification boundary, so as to improve the classification accuracy.

In recent years, researchers have proposed the idea of reproducing human concept learning ability through machine learning, in an attempt to narrow the gap between machine learning and human learning. Studies have shown that neural network based on some cognitive characteristics of human beings has an outstanding performance in breast cancer diagnosis and prevention.

Based on the idea of human-guided machine learning, this paper combines human cognitive model with support vector machine algorithm to form a new machine learning method. A fuzzy support vector machine algorithm based on human cognitive bias model was designed to solve the problem of disequilibrium of binary classification tasks. Experiments show that this new method achieves better classification accuracy in the existing public unbalanced data sets, and its effectiveness is further verified in the practical scenario of bank credit risk assessment. The main work of this paper has the following aspects:

(1) The different performance abilities of support vector machine and human in the face of unbalanced problems are compared and analyzed. This paper expounds the factors that influence the classification performance of support vector machines due to unbalanced classification problem, and illustrates the influence degree of absolute and relative imbalance of unbalanced data sets on support vector machines through experimental comparison. The evaluation indicators of machine learning methods for solving unbalanced classification problems are summarized. Through the two dimensions of human physiological and psychological cognitive features, this paper

illustrates the outstanding ability of human beings in dealing with unbalanced classification problems, and lists the existing studies to prove that transferring these features to the field of machine learning does improve the ability of the original algorithm.

(2) A fuzzy membership method of human cognitive model is designed, and the fuzzy membership calculation rules of human cognitive model are applied to support vector machine algorithm. Fuzzy membership support vector machine based on cognitive bias (FSVM-BS) is proposed. Use of people in a specific scenario of a certain cognitive features and its model, with K neighbor algorithm combining with the analysis of unbalanced data set of samples, judgment, and for the samples of "negative" fuzzy processing, reduce the large amount of negative samples in FSVM-BS model weights in the process of training, make the algorithm more emphasis on the characteristics of the small amount of samples are learning, effectively adjust classification boundaries. The FSVM-BS algorithm was verified by using the existing open binary unbalanced data set, and the classification performance of FSVM-BS algorithm was compared with that of the outstanding SVM improved algorithm. The test effect of the new algorithm and the traditional improved algorithm was analyzed and evaluated by using the evaluation index for the imbalance problem.

(3) The feasibility and effectiveness of the proposed FSVM-BS algorithm in the real environment are tested by selecting the data set of the actual scenario of bank credit risk prediction. By analyzing the research background and current situation of bank credit risk assessment, it can be found that there are more "good" samples than "bad" samples in the relevant data of credit assessment, which meets the requirements of unbalanced classification. Secondly, with the aid of random forest algorithm to importance of the characteristics of the data set, selecting the correlation stronger characteristics of FSVM-BS training algorithm, compared with other algorithms, the trained model has a higher ROC offline area value, at the same time FSVM-BS algorithm model between training set and test set the stability of the accuracy is better than the other three methods. The feasibility and effectiveness of the proposed algorithm FSVM-BS in practical application are proved.

KEY WORDS: Unbalanced classification problem, human cognitive characteristics, support vector machine, bank credit risk prediction

目 录

摘 要.....I
ABSTRACT.....III
插图清单.....VII
表格清单.....VIII
第一章 绪 论.....1
1.1 研究背景和意义.....1
1.2 国内外研究现状.....3
1.2.1 数据集层面.....3
1.2.2 算法层面.....5
1.2.3 集成分类方法.....6
1.3 本文工作与创新点.....7
1.4 本文组织架构.....8
1.5 本章小结.....9
第二章 支持向量机算法综述.....10
2.1 标准支持向量模型.....10
2.1.1 线性支持向量机与硬间隔最大化.....10
2.1.2 线性支持向量机与软间隔最大化.....15
2.1.3 常见核函数.....17
2.2 有偏支持向量机.....19
2.3 模糊支持向量机.....20
2.3.1 模糊集和模糊隶属度函数.....20
2.3.2 基于熵的模糊隶属度支持向量机模型.....21
2.4 本章小结.....22
第三章 支持向量机和人面对不平衡问题的不同表现能力.....23
3.1 不平衡问题下的支持向量机方法.....23
3.1.1 数据绝对不平衡对支持向量机算法的影响.....24
3.1.2 数据相对不平衡对支持向量机算法的影响.....25
3.1.3 不平衡问题影响支持向量机的其他因素.....26
3.1.4 不平衡问题分类方法的评价标准.....27
3.2 不平衡问题下人的行为特性.....30
3.2.1 人的生理认知特性及其应用案例.....31

3.2.2 人的心理认知特性及其应用案例.....32
3.3 本章小结.....33
第四章 融合人的心理认知特性的模糊支持向量机算法设计.....34
4.1 人的心理认知特性及其特征模型.....34
4.1.1 人的心理认知特性：对称偏差和互斥偏差.....34
4.1.2 松散对称模型.....35
4.2 融合人的心理认知模型的模糊支持向量机算法设计.....38
4.2.1 对不平衡数据中的负样本进行模糊化处理.....38
4.2.2 融合人的心理认知模型的模糊支持向量机.....41
4.3 数值实验.....43
4.3.1 实验设置.....44
4.3.2 实验结果.....45
4.4 本章小节.....47
第五章 基于 FSVM-BS 算法的具体场景应用及其分析.....49
5.1 信用风险评估的研究背景及研究现状.....49
5.1.1 背景介绍.....49
5.1.2 研究现状.....50
5.2 银行信用风险评估数据集介绍和处理.....52
5.2.1 数据说明.....52
5.2.2 数据描述和处理.....52
5.3 FSVM-BS 算法模型的实验设计及结果分析.....54
5.3.1 参数的设置.....54
5.3.2 模型实验结果与分析.....54
5. 4 本章小结.....61
第六章 总结与展望.....62
6.1 总结.....62
6.2 展望.....63
参考文献.....64
致　谢.....68
作者简介.....69
1 作者简历.....69
2 参与的科研项目.....69
3 学术论文.....69
4 发明专利.....69
学位论文数据集.....71

插图清单

图 2-1（a）（b）分别表示两种不同的支持向量的训练模型 11
图 2-2 图 2-1 训练模型对应的测试结果 11
图 2-3 支持向量机分类示意图 12
图 2-4 非线性分类问题与核技巧示例 18
图 3-1 绝对不平衡对支持向量机分类性能的影响； 24
图 3-2 相对不平衡对支持向量机分类性能的影响 25
图 3-3 类内子聚集 26
图 3-4 ROC 曲线线下曲线面积 AUC.....29
图 3-5 机器学习模型的建立流程 30
图 3-6 人的认知能力“进化”过程概括.....31
图 4-1 “对称偏差”和“互斥偏差”示意图 35
图 4-2 LS 模型的对称偏差特性.....37
图 4-3 LS 模型的互斥偏差特性.....38
图 4-4 不平衡数据集分类边界局部样本分布可能 39
图 4-5 利用认知偏差模型对负样本进行模糊化处理流程图 41
图 5-1 数据集中的违约样本、正常样本数量比例 53
图 5-2 *k* 取值为 5 对样本权重的影响.....55
图 5-3 *k* 取值为 7 对样本权重的影响.....55
图 5-4 *k* 取值为 9 对样本权重的影响.....56
图 5-5 *k* 取值为 11 对样本权重的影响.....56
图 5-6 *k* 取值为 13 对样本权重的影响.....56
图 5-7 *k* 取值为 15 对样本权重的影响.....57
图 5-8 不同 *k* 值下 20 次试验的错误统计次数及其概率.....58
图 5-9 SVM 算法在违约数据集上的 AUC 58
图 5-10 BSVM 算法在违约数据集上的 AUC.....59
图 5-11 FSVM 算法在违约数据集上的 AUC 59
图 5-12 FSVM-BS 算法在违约数据集上的 AUC.....60
图 5-13 四种模型的其他性能对比 61

表格清单

表 3-1 混淆矩阵 27
表 4-1 因果估计列联表 36
表 4-2 两种认知偏差模型和相关的重要性质 36
表 4-3 样本特征与样本类别间的因果关系 39
表 4-4 a、d 的大小关系和与之对应的样本类别判断.....39
表 4-5 FSVM-BS 算法实现步骤 43
表 4-6 不平衡数据集基本信息 44
表 4-7 四种分类器的 F-measure 值比较（加粗为最佳结果）.....45
表 4-8 四种分类器的 G-mean 值比较（加粗为最佳结果）.....46
表 4-9 四种分类器的 AUC 值比较（加粗为最佳结果）.....47
表 5-1 数据集 credit risk analysis 重要的特征信息.....53

第一章 绪 论

随着计算机技术的飞速发展，人们已经进入了大数据时代。将各行各业的信息归纳为数据，通过对数据的分析和学习，获得经验，进而指导这些行业的进一步发展。在对这些数据分析的过程中，我们不难发现很多不同标签的数据数量是不对等、不平衡的。机器学习方法在分析和学习不平衡的二分类数据时，由于数据集中不同标签的样本数不相等，机器学习模型得到的分类边界偏向于样本数较多的类别。当判断样本为低错误容忍度的疾病预测样本时，分类边界出现偏差会带来不必要的损失和伤害。为了减少数据集不平衡造成的误差，至今已有许多不同思想的优化方法。

本章的第 1.1 节介绍了研究不平衡分类问题的背景和意义；第 1.2 节从数据集、算法、集成方法三个角度总结解决不平衡分类问题的现有研究成果；第 1.3 节阐明本文的工作和创新点；第 1.4 节列举了本文的组织结构；第 1.5 节对本章的内容进行总结。

1.1 研究背景和意义

不平衡分类问题(Imbalanced Classification Problem),指机器学习解决分类问题时，应用的数据集包含某种标签的一方样本数量明显少于其他标签样本的数量，导致样本分类结果倾向于样本数量多的那一类的情况。一般地，在二分类相关的不平衡分类问题中，正样本和负样本的定义不是基于样本本身的“好”或“坏”的含义，而是把样本数量少，应获得更多关注的、属于目标导向的样本类别，称为正样本；样本数量较为丰富的一类样本标记为负样本。比如，癌症检测中的患有癌症的样本称为“正样本”，正常人的样本为“负样本”。二分类不平衡问题的不平衡率(Imbalanced Ratio，IR)，用样本中的正、负样本数量衡量，某样本的不平衡程度用公式这样表示[1]：

$$IR = \frac{\text{负样本数量}}{\text{正样本数量}} = \frac{n_{neg}}{n_{pos}}$$

在现实生活中，太过于频繁发生的事情由于其普通性往往不能引起人们的注意，相反，偶然发生的事件因为其特殊性、差异性往往引起我们的好奇。不平衡分类问题也是如此，我们最应该关注的是少数正样本的特征，因为少数正样本包含的信息的稀缺性、差异性、特殊性使其更具有价值。例如在乳腺癌诊断方面，我们重点关注的是患乳腺癌患者的样本信息，其中包含的病理、病症

对于诊断病情、指导治疗方案更具有意义。然而，由于医疗诊断场景的特殊性，患者的人数远少于正常人的人数，导致难以取得大数量的病理样本供医护工作者参考。我们只能从少量的患者身上获得少量的信息提供给医护工作者进行诊断和治疗。

以二分类问题举例，假设正样本、负样本的数量比例为1：99，一般的分类器为了保证整体分类性能，都会使分类边界倾向于负样本。在实际应用中，这种分类器即使分类精度也能达到 99%，但并不能满足我们真实的需求，尤其是我们重点关注的样本为少数类的正样本的时候。

不平衡分类问题遇到的困难及其在机器学习和数据挖掘的应用中起的重要作用已引起了学界的关注。在 2000 年举办的国际人工智能协会(The Association for the Advancement of Artificial Intelligence，AAAI)会议首次组织了关于不平衡问题讨论的研讨会，2003 年的机器学习国际会议(the International Conference on Machine Learning，ICML)上也就不平衡分类问题展开研究讨论。2004 年，美国计算机协会的知识发现和数据挖掘研讨会[2](ACM SIGKDD Explorations Newsletter)介绍了不平衡学习的进展和影响；在 2005 年的 ICDM(IEEE International Conference on Data Mining)会议上，不平衡类数据挖掘被列为数据挖掘领域的十大难题之一。

随着计算机技术和大数据的快速发展，我们寄希望于机器学习方法对数据经过分析和学习得到规律和经验，之后把这种经验反作用于我们日常的学习和工作。传统的机器学习算法主要是基于数据集相对平衡的样本进行设计的。在相对平衡的数据集中，不同类别的样本数量相对均衡，且不同类别的样本在误分类时对应的代价也基本相同。与正样本相比，负样本数量较多且包含的信息相对更加完善，但负样本包含的信息价值低于正样本。传统机器学习面对不平衡数据集，负样本因其数量较多，干扰了机器学习对数量明显少的正样本学习，从而正样本包含的信息容易被忽视甚至被遗漏。另一方面，现有的分类方法为了追求整体的分类精度不惜降低少数类的正样本的分类效果，这些都导致了不合适甚至不正确的分类结果。然而，在现实生活中这样理想的数据集很难有对应的场景。生活中常见场景的样本数据信息多是不平衡的，例如文本分类[3]中垃圾邮件和正常邮件、故障诊断[4]中残缺部件和完整部件、医疗诊断[5]中患者和非患者以及风险估计[6]中失信用户和正常用户等，这些日常场景中均存在一方数据与另一方不平衡分布的情况。重视少数类正样本的作用，弱化数据集不平衡的影响，提高少数类正样本的分类精度，对于解决个人、企业甚至整个社会层面的问题都具有很强的现实意义。

1.2 国内外研究现状

不平衡分类问题对多种传统分类算法的性能带来负面影响，为减小这种样本不平衡带来的影响，不少研究学者分析了不平衡分类问题中的存在的问题，提出了不同的应对方法，并通过实验证明其方法的有效性。这些改进的方法归纳为三个维度：数据集层面的方法、算法层面的方法、结合数据集和算法两个层面的方法。接下来，主要介绍关于以上三个层面的研究现状的具体方法。

1.2.1 数据集层面

不平衡分类问题是由于数据集中的正负样本数量存在明显偏差引起的，通过对数据集中的样本数量的修改来调整数据集的不平衡，减小样本的不平衡程度。用重新生成的相对平衡的数据集作为分类模型的训练数据，可以有效提高不平衡分类问题的精度。这个方法用于数据预处理阶段的准备工作。从数据集层面解决不平衡分类问题的方法主要分为三类：（1）过采样、（2）欠采样和前两者结合的（3）混合采样方法。

（1）过采样方法

过采样方法(over-sampling methods)，通过简单复制或生成合成数据在数据集中添加新的正样本，增加数据集中正样本的数量。最简单的过采样方法是随机过采样(ROS)，它简单地从原始数据集中复制正样本，直到数据集的不平衡率接近 1 为止[7]。此时正样本数量增加，但增加的样本只是重复了已有的数据，没有补充新鲜的、更全面的信息，这样会导致模型学到的内容过于重复出现过拟合的现象。为避免这种现象，不少学者为此做了不少尝试和创新。

其中，最常用的过采样方法之一是合成少数过采样技术（SMOTE），是一种在过采样算法的基础上进行改进的方法，它通过对正样本的分析和尝试，将人工模拟合成的新样本添加到原数据集中，从而降低样本的不平衡程度。该方法在尝试的过程中借用 KNN 思想，计算出每个正样本最近的 k 个样本，并从中随机选择 n 个样本连线，在所连的线段中随机选取数据作为新的正样本，最后将新的正样本与原本的不平衡数据混合，构成新的数据集[8][9]。这个方法的弊端在于利用 K 近邻规则时 k 值和随机选取样本 n 时，选取规则不明确，k 和 n 的取值需要用户自己决定。Borderline-SMOTE[10]一文在 SMOTE 方法的基础上，提出 borderline-SMOTE1 和 borderline-SMOTE2 两种新的过采样方法，这两种方法以分类边界处的正样本作为抽样样本进行抽样，来增加不平衡数据集中的正样本数量。类不平衡数据的 EM 聚类过采样算法[11]采用聚类技术，选取每个聚类簇的中心点作为采样点，既提高了数据中的正样本数量又避免了盲目采样引起的样本偏差。基于时间序列模型的非平衡数据的过采样算法[12]把少数类的正样本数据转化为时间序列并进行平稳性检测和处理，待序列平稳后建立合适的时间序列模型进行预报，从而达到数据集平衡。G-SOMO[13]作为一种基于自组织

映射和几何 SMOTE 的过采样方法，它使用自组织映射确定最佳区域以生成人工数据，并采用几何 SMOTE 算法生成仿真实例，通过 69 个数据集验证证明了方法的有效性。以上这些方法都通过采用过采样的方法增加不平衡数据集中的正样本数量，一定程度上改善了不平衡问题的分类精度。

（2）欠采样方法

欠采样方法(under-sampling methods)：通过从数量多的负样本中删除一些样本构造新的数据集，使负样本的数量减少，从而使样本中的正负样本数量趋于平衡[14]。最简单的欠采样方法与随机过采样方法思想相同，即从原始数据集中随机删除一些负样本，从而获得一个相对平衡的数据集。在这个过程中，被删去的负样本可能包含重要信息，降低数据集中样本特征的丰富程度，所以比起随机过采样方法，此方法在实际应用中使用较少[15]。

在欠采样方法的发展过程中，一方面要克服随机欠采样方法易丢失重要信息的缺点，另一方面又要保证算法的性能。比较受欢迎的一种欠采样方法为单边取样法（OSS），在这个方法中，所有的正样本均被保留保持不变，负样本根据最近邻分类法进行欠采样。在最新的研究成果中，基于特征边界信息的负采样方法[16]保留分布在最优非线性分类决策面附近的数据点，删除其他远离分类边界的负样本，既保留了重要的分类信息又能减少不平衡数据集中的负样本数量。基于密度峰值聚类的自适应欠采样方法[17]，首先删除最近邻搜索算法识别出的重复出现的负样本区域，之后采用改进的密度峰值聚类方法计算剩余子簇中的样本密度，以此得到采样权重进行欠采样。此外，一种基于聚类的加权边界点集成欠采样算法[18]以正样本为负样本的初始聚类中心进行聚类，把变异系数识别出的边界点加权后用到不平衡数据的处理中，将根据簇密度分出的低密度簇删除，最后获得欠采样后的负样本数据集和原始的正样本整合成新的用于 AdaBoost 训练的相对平衡的数据集，该方法对多数类的样本进行精简也提高了算法的执行效率，最终也提高了不平衡数据集分类的精确度。另外一种基于聚类的欠采样方法[19]采用 K-means 聚类算法对数据进行聚类，利用马氏距离分析各聚类中样本到质心的距离，保留各聚类中规律分布数据的选择方法，通过 44 个选自 KEEL 的数据集的实验验证，该方法优于文中提到的其余 7 种先进算法。

（3）混合采样方法

混合采样方法(the hybrid methods)：混合方法将前两种方法结合起来[14]，尽量避免过采样和欠采样方法中的缺点，通常从过采样方法开始，之后应用欠采样方法从负类样本或两个类中删除样本，以实现数据集平衡，这样的策略能获得更好的效果。

文章[20]针对重叠样本提出一种进化混合抽样技术，这个方法先除去无用的多数类样本，决策边界更清晰的同时也避免新合成的少类样本中夹杂噪声，在

KEEL 数据库的所有二值类非平衡数据集(100 个数据集)上进行的数值实验表明，该方法与其他常用的采样方法相比具有优越性。文献[21]提出一种区别于其他混合采样方法的基于聚类的不平衡数据分类混合采样方法，通过 15 个不同程度的不平衡数据集的实验，结果表明该方法的性能优于其他先进的方法。文章[22]提出一种基于熵的混合采样方法，利用信息熵考虑训练数据的分布情况，在欠采样过程中避开重要样本；在过采样时为了避免数据重叠，正类数量扩展直到负类每个子集的大小。文献[23]基于异类 K 距离识别不平衡数据集的样本分类边界，并对三种不同支持度的负样本采用不同的过采样方法和倍率，生成包含具有更重要信息的样本点；基于异类 K 距离，对远离分类边界的负样本进行欠采样，使达到数据集样本平衡。

1.2.2 算法层面

算法层面的方法主要分为两类：一类是通过直接修改分类器来提高模型的分类性能，另一类是代价敏感学习的方法，对不同的样本弱化分类数据不平衡带来的影响。算法层面的方法与分类器相结合，在模型训练的过程中发挥作用。

开发一个算法方案来解决现有的问题,特别是现有的问题可用数据的类分布不均匀,有必要了解相应的分类器学习算法和应用领域的同时,学习之前算法失败的原因。对于决策树而言，处理类不平衡问题的一种常用策略是选择适当的归纳偏差，一种方法是调整节点的概率估计，另一种方法是开发新的剪枝方法[24]。对于支持向量机，提出对不同的类使用不同的惩罚常数[25]，或基于核对齐思想调整类边界[26]等建议。基础的朴素贝叶斯算法的独立性假设在现实的应用环境中并不存在[27]。

（1）一类学习

一类学习(one-class learning)，也可称为单类学习,在实际应用中，由于场景特殊(如军事环境等)无法获得另外一种样本或者需要花费大量的成本才能得到样本时，一类学习是一种有效的解决方法。

使用一类学习方法最多的算法之一是支持向量机算法,这样的一类学习称为一类支持向量机[28](OSSVM)。基于一类支持向量机算法的类别增量方法通过已学的知识和模型进行类增量学习，一方面充分利用学到的数据信息不用存储减少了存储成本。一类板级支持向量机[29]（OCSSVM）的快速学习算法提出一种改进的序列最小化算法对 OCSSVM 进行快速训练的方法，在不明显降低精度的前提下提高了算法的可扩展性。在临床环境的医学图像数据中，提出一种基于深度学习的一类分类方法[30]，利用成像复杂性的概念使深度学习模型更优地捕获、学习与单类相关的固有成像特征。一种基于迁移学习的推荐数据流单类字典学习模型[31]，将迁移学习和判别字典学习引入到单类数据中学习，将概念从最近的历史模块转移到当前模块，构建新模块的预测分类器，从而减少不确定数据对分类器的影响。一种基于一类学习的多任务字典学习方法[32]将分析判别

字典融入到一类学习中，分析判别字典学习保证了对不同任务做出响应的字典具有独立性和尽可能多的判别性，以此提高分类判别的性能。

（2）代价敏感学习

代价敏感学习(cost-sensitive learning)，算法分类时并不是所有的样本都能被正确分类，错误分类的样本根据样本的类型不同导致不同的损失，用$C(i，j)$表示类别 i 预测为类 j 的代价，在这个过程中最小化高成本错误的数量和总的错误分类成本。代价敏感学习是算法层面适用于解决不平衡分类问题的常用方法之一。

Yanmin Sun[33]等人在 AdaBoost 的学习框架中引入代价项，提出了基于代价敏感的 boosting 算法。Veropoulos[34]提出 Biased 支持向量机(BSVM)，假设样本少的一类为负样本，并赋予比样本多的正类惩罚参数。J Xu 和 Y Cao[35]等一起提出秩向量机算法，在不同的秩对应的样本分类错误时设置不同的损失。Qi Fan[1]等人提出基于熵的模糊隶属度的模糊支持向量机算法(FSVM)，通过熵来确定样本类的模糊隶属度，并根据样本的类的模糊隶属度度量不同样本对分类器的重要性。Ruth C Fong[36]等人提出一种“神经加权”的机器学习范式，从观看图像的受试者身上获取人脑活动的 fMRI 测量值，并把这些数据注入到物体识别算法的训练过程中，在这个过程中使其与人脑进行对比，根据对比结果分配不同的惩罚力度，使算法能力最终趋近于人脑。

1.2.3 集成分类方法

集成方法(ensemble learning)，是一种多个弱学习机混合学习的学习模式，通常以神经网络、支持向量机等弱分类器为基学习机，采用某种方式将这些学习机结合为一个集成学习机。与其他传统机器学习方法相比，集成方法在解决不平衡分类问题时具有更好的泛化能力和分类精度。一般用于生成基学习机的方法，根据是否同质，大致可分为两种类型：异质类型，结合不同的学习算法后针对同一个数据集学习；同质类型[37]，将同一学习算法在不同训练集中学习。

RC Bhagat 和 SS Patil[38]在 2015 年提出用随机森林强化 SMOTE 算法解决不平衡问题的能力，在这个新的学习模式中，通过二值化技术将原始数据集拆分，得到不同的二进制类的子集，之后对其采用 SMOTE 方法获得平衡的二分类的子集，最后构建随机森林分类器对样本进行分类。陈圣灵[39]等人于 2018 年提出 Rotation SMOTE 算法，该算法在 boosting 过程中，根据基分类器的预测结果对正样本进行过采样，增加正样本的权重；基于 FocalLoss 基本思想，依照基分类器的预测结果直接对 AdaBoost 权重进行优化。Fan X[40]等人于 2019 年提出一种双向长短时记忆卷积神经网络加权极限学习机(BC-WELM)，提高了模型在生物医学相关的不平衡问题中的分类性能。杨昊天等[41]在软件缺陷预测中应用了一种基于 SMOTE 混合采样和集成学习算法 XGBoost 结合的方法，在相关的软件

缺陷数据集上方法的优越性得到有效验证。Anbazhagan Mahadevan 等[42]2 人在 2020 年提出一种模型，将 bagging 集成和 AdaBoost boosting 集成两种方法合并成一个更强大的集成结构，该模型通过 Amazon 产品数据集实验，在 G-均值、F-measure 和 ROC-AUC-Score 方面性能优于其他模型。Choudhary Roshani 和 Shukla Sanyam 在 2021 年发表的文章[43]提出一种基于聚类的加权核化极限集成学习，将一个复杂的不平衡问题分解为更简单的子问题，利用代价敏感分类器解决这些子问题，然后利用投票方法将各个分类器的结果进行组合。

1.3 本文工作与创新点

不平衡分类问题真实存在于日常生活的方方面面，提高不平衡数据的分类研究在机器学习中有着极其重要的地位。本文以不平衡数据为前提，结合支持向量机的基本思路和优化方法，把人的认知偏差模型添加到支持向量机算法的改进模型里面，通过用人的心理认知特性模型重新定义一种融合人的认知模型的模糊隶属度计算规则，给不同的训练样本赋予不同的学习权重，把这个新的模糊隶属度定义方法和支持向量机方法结合，构建更有效的处理不平衡数据的支持向量机的分类模型，并和已有的改进方法做对比，分析各种方法的优缺点，进而提高支持向量机处理不平衡分类问题的能力。本文的创新点主要包括两个方面：

（1）提出一种融合人的认知模型的模糊隶属度计算方法。依据 K 近邻算法得到样本点归属于正类或负类的概率，以此作为人的认知模型计算样本模糊隶属度的参数。对于归属于正类的样本点不进行模糊处理，权重仍保持为1；对于归属于负类的样本基于人的心理认知特性模型进行模糊化处理，从而降低原本数量多的负样本的权重，减小数据集不平衡对分类器分类性能的影响。将样本所属类别和此样本的学习重要程度两者间的因果关系用基于人的心理认知特性模型来定义，保证重要的样本在学习过程中得到更多的关注。

（2）利用融合人的认知模型的模糊隶属度方法和传统的支持向量机算法结合，得到一种融合人的认知模型的模糊支持向量机方法。利用 K 近邻算法统计出某未知样本点附近的正负样本分布情况，使用人的心理认知特性模型计算样本的模糊隶属度，被赋予模糊隶属度的样本用于支持向量机算法的训练过程，模糊隶属度高的样本给予的学习权重高于模糊隶属度小的样本数据，最终形成融合人的认知特性的模糊支持向量机模型。把人处理不平衡数据时的某种优势引入到机器学习领域，提高支持向量机方法在处理不平衡数据时的性能。

1.4 本文组织架构

本文针对不平衡问题提出了一种基于人类认知偏差模型的非平衡认知模糊支持向量机算法，重点研究了基于人类认知偏差模型的样本模糊处理方法。该方法给位于分类边界的不同的样本赋予不同的权重，使样本在支持向量机模型的训练过程中得到相应的关注。通过 15 个不同不平衡程度的数据集和真实的银行信贷风险预测不平衡数据集，对本方法的有效性和实用性进行验证。文章所阐述的内容由以下六个章节组成：

第一章：绪论，主要介绍了本文的研究背景及意义，介绍了关于不平衡分类问题研究相关的国内外研究现状，从不同层面介绍了解决不平衡分类问题的改进方法，最后阐述了本文的主要工作、创新点和论文的具体内容结构安排。

第二章：相关的理论知识，主要介绍了关于支持向量机的理论基础和其发展、衍生的模型，及其优化过程。其中，着重介绍了解决不平衡分类问题的基于熵的模糊隶属度的支持向量机模型算法。

第三章：比较和分析支持向量机模型和人面对不平衡分类问题的不同表现能力。其中通过实验探究不平衡程度对标准支持向量机算法性能的影响程度，分析影响分类结果的因素，并列出评价不平衡分类问题的模型性能的评价指标。同时，结合具体案例介绍人在面对小数量的、存在偏差的任务时的能力及其特点。

第四章：FSVM-BS 算法设计，主要介绍了从逻辑学角度出发的人的认知特性的理论基础及其数学模型，进而设计 K 近邻规则和人的心理认知模型结合，形成一种新的模糊隶属度函数计算方法，并把这种融合人的心理认知模型的模糊隶属度方法与现有的支持向量机算法相结合，得到融合人的认知模型的模糊支持向量机模型。FSVM-BS 算法的实验验证，其中包括所用实验环境和所用数据集的介绍，利用多个不平衡的二分类数据集对 FSVM-BS 算法和其他改进方法做对比实验。并通过 F-measure、G-mean 和 AUC 标准评价算法的性能，证明其有效性。

第五章：通过真实的、具体的银行信用风险预测数据集，进行数据处理和实验验证，结果表明本文提出的方法在实际生活场景的具体应用中有效。分析银行信用风险预测的必要性和国内外研究现状；针对银行信用风险预测数据集中“违约”样本过少导致的数据集不平衡问题和小微贷款机构难以获得大量真实数据的困难，以及对符合合格信贷特征的潜在违约者的精准预测，采用本文提出的算法对此预测，检验本文提出方法在实际场景应用的有效性。

第六章：对全文的研究内容进行总结，说明本文提出方法的创新性、有效性，同时也指出本文研究内容存在的缺陷和不足，并对下一步的研究工作进行展望和规划。

1.5 本章小结

本章主要阐述了不平衡分类问题的研究背景和意义，并针对国内外就不平衡分类问题的研究现状进行了总结。阐明本文的研究内容和结构安排，为后续章节的研究工作做了铺垫。

第二章 支持向量机算法综述

支持向量机(Support Vector Machine，SVM)，作为一个良好的数学统计学习方法，已经成功地应用于许多实际的分类问题。它通过扎实的数学基础、较高的泛化能力，在寻找全局最优决策面方面表现突出[44]。因此，在过去的时间里，支持向量机理论的理解、算法的优化和具体案例的应用都有了很大的发展[45]。

本章的2.1节介绍标准支持向量机模型的基本理论和公式推导过程，包括硬间隔、软间隔支持向量机和几种常用核函数。第2.2节和第2.3节分别介绍了用于提升不平衡分类问题精度的有偏支持向量机和模糊支持向量机方法。第2.4节对本章节内容进行总结。

2.1 标准支持向量模型

2.1.1 线性支持向量机与硬间隔最大化

支持向量机是一种二分类模型。它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大化使它区别于感知机。支持向量机的学习策略为间隔最大化，可形式化为凸二次规划问题，等价于正则化的铰链损失函数的最小化问题。支持向量机学习算法是求解凸二次规划的最优化算法。

假设给定一个特征空间上的训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_N, y_N)\}$$

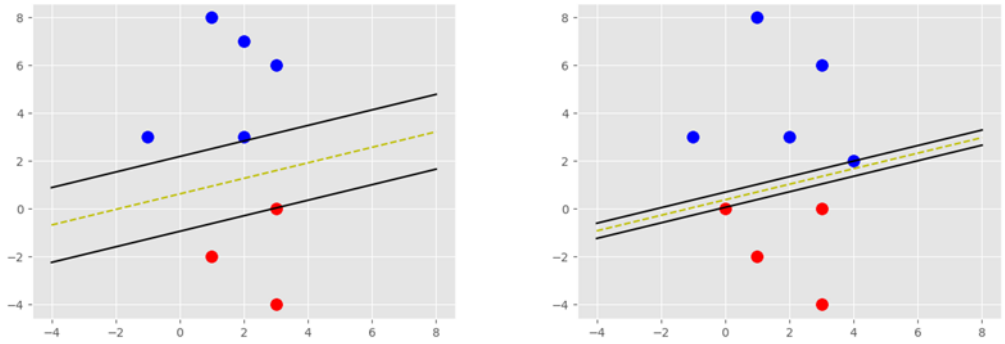
其中，$x_i \in X = R^n$，$y_i \in Y = \{-1, +1\}$，$i = 1,2, \dots, N$，x_i为第i个特征向量，也称为实例。y_i为x_i的类标签，当y_i为+1 时，称x_i为正样本，反之为负样本，称(x_i, y_i)为样本点。假设样本点线性可分，学习的目标是在特征空间中找到一个将实例分开到不同类的面，这个面称为分离超平面。分离超平面可以通过间隔最大化，或等价地求解相应的凸二次规划问题学习得到：

$$w^T x + b = 0 \tag{2-1}$$

它由法向量w、截距b决定。一般地，当训练数据集线性可分时，存在无穷个分离超平面正确地分离正、负两类数据。利用间隔最大化找出最优超平面时，这个平面是唯一的，这也是支持向量机算法与感知机的不同之处。

从几何的角度出发，支持向量是决定最优分离超平面的样本数量的最小个数，是使约束条件等号成立的点。在决定分离超平面时只有支持向量起作用，如果支持向量有所变动，所求的解随之也会发生变化；但如果间隔边界以外的

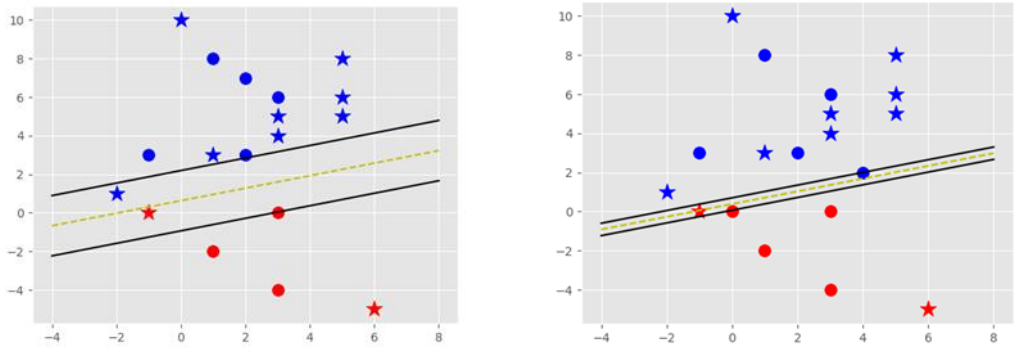
点被移动，甚至被移除，解不变。可观察图 2-1、图 2-2 上的红、蓝色样本点。不同的支持向量训练得到的间隔超平面不同，预测分类结果也会受影响。



(a) (b)

图 2-1 （a）（b）分别表示两种支持向量的训练模型

Figure 2-1. (a) and (b) training models of two support vectors.



(a) (b)

图 2-2 图 2-1 训练模型对应的预测结果；

（a）间隔内的蓝色星星可能为噪声，或错误分类；对应的（b）中的蓝色星星被正确分类

Figure 2-2. Fig 2-1 Prediction results corresponding to the training model.

(a) The blue stars within the interval may be noise or misclassified; The corresponding blue star in (b) is correctly classified

假设支持向量正样本x_+和负样本x_-，两者间的距离表示为

$$Margin = \sqrt{(x_+ - x_-)^T(x_+ - x_-)} \tag{2-2}$$

正样本x_+在分类边界

$$w^T \cdot x_+ + b = +1 \tag{2-3}$$

负样本x_-在分类边界

$$w^T \cdot x_- + b = -1 \tag{2-4}$$

上，正样本$\boldsymbol{x}_+$和负样本$\boldsymbol{x}_-$间的距离认为是正负类样本所在的超平面间的间隔(如图 2-3 所示)，其满足如下关系：

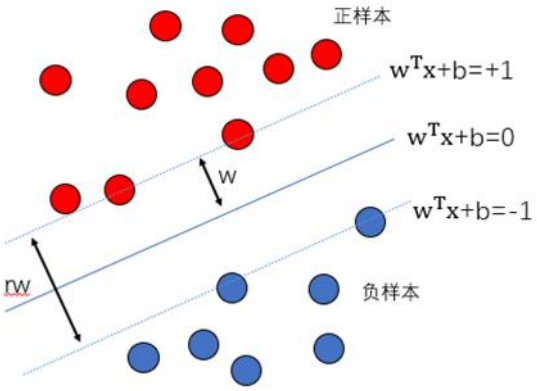


图 2-3 支持向量机分类示意图

Figure 2-3. Support vector machine classification diagram

$$x_+ = x_- + rw \tag{2-5}$$

$$w^T(x_- + rw) + b = 1 \tag{2-6}$$

代入调整后，

$$(w^T x_- + b) + rw^T w = 1 \tag{2-7}$$

可得，

$$r = \frac{2}{w^T w} \tag{2-8}$$

$$x_+ - x_- = rw = \frac{2w}{w^T w} \tag{2-9}$$

那么，正样本x_+和负样本x_-之间的距离可表示为$\frac{2w}{w^T w}$，即求分类间隔最大化得问题可以转化为求$\frac{2w}{w^T w}$最大化：

$$\begin{aligned} \text{Margin} &= \max_{w \in R^d} \sqrt{\left(\frac{2w}{w^T w}\right)^T \left(\frac{2w}{w^T w}\right)} \\ &= \max_{w \in R^d} \sqrt{\frac{4w^T w}{w^T w \times w^T w}} \\ &= \max_{w \in R^d} \frac{2}{\sqrt{w^T w}} \end{aligned} \tag{2-10}$$

为了计算方便，上式可写为：

$$\begin{aligned} &\min_{w \in R^d} w^T w \\ &= \min_{w \in R^d} \frac{1}{2} w^T w \\ &= \min_{w \in R^d} \frac{1}{2} \|w\|^2 \end{aligned} \tag{2-11}$$

此时，$\max_{w \in R^d} \frac{2}{\sqrt{w^T w}}$和$\min_{w \in R^d} \frac{1}{2}\|w\|^2$是等价存在的。于是得到线性可分支持向量机学习的最优化问题

$$\min_{w\in R^d}\frac{1}{2}\|w\|^2 \qquad (2\text{-}12)$$
$$s.t. \qquad y_i(w\cdot x_i+b)-1\geq 0, i=1,2,\dots,N$$

这是一个凸二次规划问题，$s.t.$是约束条件，通过对此凸二次规划问题优化得到最优解 w 和 b，由此得到分离超平面

$$w\cdot x+b=0 \qquad (2\text{-}13)$$

和分类决策函数

$$f(x)=\text{sign}(w\cdot x+b) \qquad (2\text{-}14)$$

应用拉格朗日对偶性，不仅容易得到线性可分支持向量机原始问题的最优解，还方便引入核函数，推广运用到非线性分类问题。首先对每个不等式约束条件引入拉格朗日乘子，构建拉格朗日函数：

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i y_i(w\cdot x_i+b)+\sum_{i=1}^{N}\alpha_i \qquad (2\text{-}15)$$

其中，$\alpha=(\alpha_1,\alpha_2,\dots,\alpha_N)^T$为拉格朗日乘子向量。根据拉格朗日对偶性，原始问题的对偶问题是极大极小问题，过程上先求$L(w,b,\alpha)$对 w 和 b 的极小，再求对α的极大：

$$\max_{\alpha}\min_{w,b}L(w,b,\alpha) \qquad (2\text{-}16)$$

求$\min_{w,b}L(w,b,\alpha)$，将拉格朗日函数$L(w,b,\alpha)$分别对w、b求偏导,并令其偏导等式右边等于 0：

$$\nabla_w L(w,b,\alpha)=w-\sum_{i=1}^{N}\alpha_i y_i x_i=0 \qquad (2\text{-}17)$$

$$\nabla_w L(w,b,\alpha)=\sum_{i=1}^{N}\alpha_i\, y_i=0 \qquad (2\text{-}18)$$

由上可得

$$w=\sum_{i=1}^{N}\alpha_i y_i x_i \qquad (2\text{-}19)$$

$$\sum_{i=1}^{N}\alpha_i\, y_i=0 \qquad (2\text{-}20)$$

将其代入拉格朗日函数，得到

$$\begin{aligned} L(w,b,\alpha) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) \\ &\quad -\sum_{i=1}^{N}\alpha_i\, y_i\left(\left(\sum_{j=1}^{N}\alpha_i y_j x_j\right)\cdot x_i + b\right) + \sum_{i=1}^{N}\alpha_i \\ &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i \end{aligned} \tag{2-21}$$

即

$$\min_{w,b} L(w,b,\alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i \tag{2-22}$$

在这个数学表达式中，$y_i y_j$作用在于看两个数据点是否属于同一类别，若属于同一类别使此项表达式的值增加，否则减小；$(x_i \cdot x_j)$来衡量两个数据之间的相似性。

将式（2-22）转化为对偶问题，求$\min_{w,b} L(w,b,\alpha)$对α的极大

$$\begin{aligned} &\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i \\ &s.t.\ \sum_{i=1}^{N}\alpha_i y_i = 0 \\ &\alpha_i \geq 0，\ i = 1,2,\dots,N \end{aligned} \tag{2-23}$$

将上述目标函数求极大转化为求极小，得到式（2-24）与之等价的对偶最优化问题：

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \\ &s.t.\ \sum_{i=1}^{N}\alpha_i y_i = 0 \\ &\alpha_i \geq 0，\ i = 1,2,\dots,N \end{aligned} \tag{2-24}$$

得到最优解

$$\alpha^* = (\alpha_1^*, \dots, \alpha_j^*)^T \tag{2-25}$$

根据 KKT 条件，可得

$$w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i \tag{2-26}$$

$$b^* = y_i - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j) \tag{2-27}$$

分离超平面可写成

$$\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_j) + b^* = 0 \tag{2-28}$$

分类决策函数为

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_j) + b^*) \tag{2-29}$$

2.1.2 线性支持向量机与软间隔最大化

线性可分问题的支持向量机学习方法，面对训练收集时产生噪声或者度量过程中丢失正确数据又或是其他原因导致样本点统计错误时是不适用的，因为这时上述方法中的不等式约束并不能都成立。需要将硬间隔最大化，使其修改为软间隔最大化。

假设给定一个特征空间上的训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_N, y_N)\}$$

其中，$x_i \in X = R^n$，$y_i \in Y = \{-1, +1\}$，$i = 1, \dots, N$，x_i为第 i 个特征向量，也称为实例，y_i为x_i的类标签。现假设训练数据集线性不可分。一般情况下，训练数据中会有一些异常值。去除这些异常值后，剩下的大多数样本点组成的集合线性可分。线性不可分指某些样本点不满足函数间隔大于等于 1 的约束，为解决这个问题，每个样本点对应地引进一个松弛变量$\xi_i \geq 0$，使函数间隔加上松弛变量大于等于 1。于是，约束条件变为

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \tag{2-30}$$

同时，对每个松弛变量ξ_i都对应一个代价 C，这里的$C > 0$，C 值越大对误分类的惩罚越大。目标函数由原来的$\frac{1}{2}\|w\|^2$变为

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \tag{2-31}$$

此时的目标函数的含义不仅要求分类间隔最大，同时约束误分类点的个数，分错个数越少越好，这样的目标函数称作软间隔最大化。

线性不可分的线性支持向量机的学习问题变成凸二次规划问题：

$$\begin{aligned} &\min_{w \in R^d} \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{N} \xi_i \\ s.t. \quad &y_i(w \cdot x_i + b) \geq 1 - \xi_i, i = 1,2, \dots, N \\ &\xi_i \geq 0, i = 1,2, \dots, N \end{aligned} \tag{2-32}$$

软间隔最大化问题，得到的分离超平面为

$$w \cdot x + b = 0 \tag{2-33}$$

与之对应的分类决策函数

$$f(x) = \text{sign}(w \cdot x + b) \tag{2-34}$$

把式（2-34）的凸二次规划问题转化为对偶问题：

$$\begin{aligned} L(w,b,\xi,\alpha,\mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ -\sum_{i=1}^{N}\alpha_i(y_i(w \cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i \\ \alpha_i \geq 0, \mu_i \geq 0 \end{aligned} \tag{2-35}$$

接下来求拉格朗日函数的极大极小问题，先求$L(w,b,\xi,\alpha,\mu)$对 w，b，ξ的极小，同样地，将拉格朗日函数$L(w,b,\xi,\alpha,\mu)$分别对 w，b，ξ求偏导，并令其偏导等式右边等于 0：

$$\nabla_w L(w,b,\xi,\alpha,\mu) = w - \sum_{i=1}^{N}\alpha_i y_i x_i = 0 \tag{2-36}$$

$$\nabla_b L(w,b,\xi,\alpha,\mu) = -\sum_{i=1}^{N}\alpha_i y_i = 0 \tag{2-37}$$

$$\nabla_{\xi_i} L(w,b,\xi,\alpha,\mu) = C - \alpha_i - \mu_i = 0 \tag{2-38}$$

可得，

$$w = \sum_{i=1}^{N}\alpha_i y_i x_i \tag{2-39}$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0 \tag{2-40}$$

$$C - \alpha_i - \mu_i = 0 \tag{2-41}$$

将式（2-39）至（2-41）代入拉格朗日函数，得

$$\min_{w,b,\xi} L(w,b,\xi,\alpha,\mu) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i \tag{2-42}$$

再对$\min_{w,b,\xi} L(w,b,\xi,\alpha,\mu)$求$\alpha$的极大，即得对偶问题：

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i \tag{2-43}$$

$$s.t.\ \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{2-44}$$

$$C - \alpha_i - \mu_i = 0 \tag{2-45}$$

$$\alpha_i \geq 0 \tag{2-46}$$

$$\mu_i \geq 0, i = 1,2,\dots,N \tag{2-47}$$

上述约束条件代入变换后发现μ_i被消掉，只留下变量α_i，上述约束条件可以简写为

$$0 \leq \alpha_i \leq C \tag{2-48}$$

假设得到对偶问题的一个解

$$\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_N^*)^T \tag{2-49}$$

根据 KKT 条件，可得

$$w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i \tag{2-50}$$

$$b^* = y_i - \sum_{i=1}^{N}\alpha_i^* y_i (x_i \cdot x_j) \tag{2-51}$$

分离超平面可写成

$$\sum_{i=1}^{N}\alpha_i^* y_i (x \cdot x_j) + b^* = 0 \tag{2-52}$$

分类决策函数为

$$f(x) = \text{sign}(\sum_{i=1}^{N}\alpha_i^* y_i \left(x \cdot x_j\right) + b^*) \tag{2-53}$$

2.1.3 常见核函数

对于线性分类问题，本章前文描述的线性分类支持向量机能有效解决。但对于非线性分类问题，需要引入核函数，利用核技巧将原空间的数据映射到新空间，把非线性数据映射到高维空间、甚至是无穷空间，从而把非线性分类问题转化为线性分类问题，如图 2-4，表示核函数将线性不可分的原始数据映射到特征空间中，实现特征空间内数据线性可分。

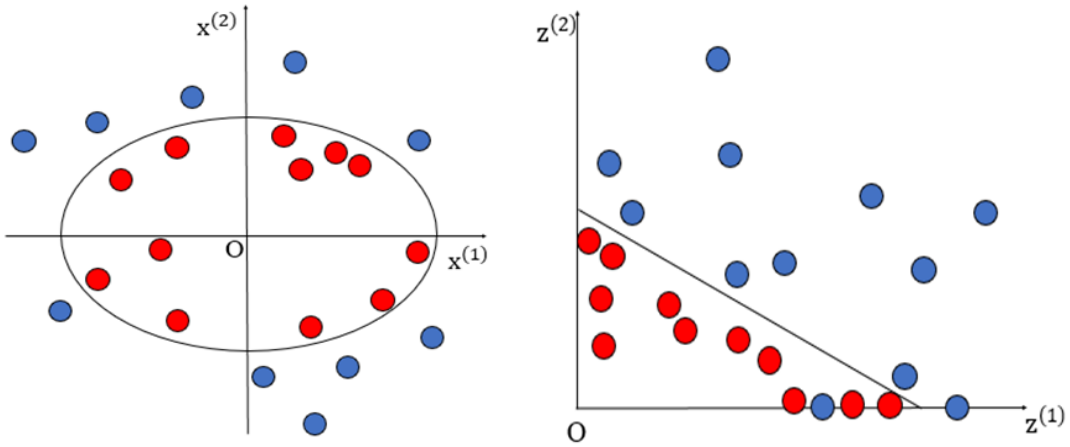
对核函数做出定义，设χ是输入空间(欧氏空间R^n的子集或离散集合)、$\mathcal{H}$为 Hillbert 空间，如果存在一个从χ到$\mathcal{H}$的映射

$$\phi(x): \chi \to \mathcal{H} \tag{2-54}$$

使得对所有$x, z \in \chi$，函数$K(x,z)$满足条件

$$K(x,z) = \phi(x) \cdot \phi(z) \tag{2-55}$$

此时称$K(x,z)$为核函数，$\phi(x)$为映射函数，$\phi(x) \cdot \phi(z)$为$\phi(x)$和$\phi(z)$的内积。



(a) (b)

图 2-4 非线性分类问题与核技巧示例；（a）非线性分类（b）映射到特征空间

Figure 2-4. Nonlinear classification problems and examples of kernel techniques; (a) Nonlinear classification (b) mapping to the feature space

在支持向量机的应用中，对偶问题中目标函数的内积$x \cdot x_j$可以用核函数$K(x,z) = \phi(x) \cdot \phi(z)$替换，此时的对偶问题变成

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \tag{2-56}$$

同样的，分类决策函数的内积也用核函数代替，于是分类决策函数成为

$$\begin{aligned} f(x) &= \text{sign}\left(\sum_{i=1}^{N}\alpha_i^* y_i \phi(x) \cdot \phi(z) + b^*\right) \\ &= \text{sign}(\sum_{i=1}^{N}\alpha_i^* y_i K\left(x \cdot x_j\right) + b^*) \end{aligned} \tag{2-57}$$

在这个过程中，在给定的核函数条件下，从新变换的特征空间中的训练样本学习线性支持向量机，实现利用解线性分类问题的方法解决非线性分类问题。在具体实际的应用中，根据不同的应用场景选择不同的核函数，核函数是否适合需要实验进行验证。

目前，最常用的核函数主要包括以下几种形式：

（1）线性核函数

$$K(x,z) = (x,z) \tag{2-58}$$

与之关联的支持向量机是

$$f(x) = \text{sign}(\sum_{i=1}^{N}\alpha_i^* y_i (x \cdot z) + b^*) \tag{2-59}$$

（2）多项式核函数

$$K(x,z) = (x \cdot z + 1)^p \tag{2-60}$$

与之关联的支持向量机是一个 p 次多项式分类器，此时的分类决策函数为

$$f(x)=\text{sign}(\sum_{i=1}^{N_s}a_i^*\,y_i(x_i\cdot x+1)^p+b^*) \tag{2-61}$$

（3）高斯核函数

$$K(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) \tag{2-62}$$

其中，σ为核参数，对应的支持向量机是高斯径向基函数分类器，此时的分类决策函数为

$$f(x)=\text{sign}(\sum_{i=1}^{N_s}a_i^*\,y_i\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)+b^*) \tag{2-63}$$

（4）Sigmoid 核函数

$$K(x,z)=\tan h(u(x\cdot z)+r) \tag{2-64}$$

与之关联的支持向量机是“激活函数”分类器，此时的分类决策函数为

$$f(x)=\text{sign}(\sum_{i=1}^{N_s}a_i^*\,y_i\tan h(u(x\cdot z)+r)+b^*) \tag{2-65}$$

2.2 有偏支持向量机

面对不平衡分类问题，支持向量机具有较好的鲁棒性，对待所有样本一视同仁，忽视正、负样本存在的不同，导致决策边界偏向于数量多的负类一侧。因此，正确识别少数类的正样本需要更强大的方法，这个方法为不同标签的每个样本选择合适的学习权重，是支持向量机解决不平衡分类问题的一个着手点。

Veropoulos 等人[46]提出了一种不同错误代价的方法，这个方法对正、负样本赋予两个不同的惩罚参数，称为有偏支持向量机(Biased support vector machines，BSVM)。设少数类的正样本数量为n_1，多数类的样本个数为n_2，这个方法的优化问题模型表示如下：

$$\begin{aligned}&\min_{w,b,\xi_i}\frac{1}{2}\|w\|^2+C_+\sum_{i=1|y_i=+1}^{n_1}\xi_i+C_-\sum_{i=1|y_i=-1}^{n_2}\xi_i\\&s.t.\quad y_i(w\cdot x_i+b)+\xi_i\geq 1\\&\xi_i\geq 0,i=1,2,\dots,n\end{aligned} \tag{2-66}$$

其中，C_+和C_-分别是正、负样本分类错误的惩罚系数。为保证足够重视少数类的正样本，通常少数类的正样本比多类的负样本有更高的错误分类成本，即一般情况下$C_+>C_-$。

通过构建拉格朗日函数求解上述优化问题：

$$L = \frac{1}{2}\|w\|^2 + C_+ \sum_{i=1|y_i=+1}^{n_1} \xi_i + C_- \sum_{i=1|y_i=-1}^{n_2} \xi_i - \sum_{i=1}^{n} \alpha_i (y_i(w \cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} \mu_i \xi_i \tag{2-67}$$

其中，$\alpha = (\alpha_1, \dots, \alpha_n)^T$和$\mu = (\mu_1, \dots, \mu_n)^T$是拉格朗日乘子。通过 KKT 条件可得

$$\nabla_w L(w, b, \xi, \alpha, \mu) = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \tag{2-68}$$

$$\nabla_b L(w, b, \xi, \alpha, \mu) = \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{2-69}$$

$$\nabla_{\xi_i} L(w, b, \xi, \alpha, \mu) = C_+ - \alpha_i - \mu_i = 0, i = 1,2, \dots, n_1, y_i = +1 \tag{2-70}$$

$$\nabla_{\xi_i} L(w, b, \xi, \alpha, \mu) = C_- - \alpha_i - \mu_i = 0, i = 1,2, \dots, n_2, y_i = -1 \tag{2-71}$$

把以上四个式子的关系代入，可得优化问题模型的对偶问题：

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j (x_i \cdot x_j)\, \alpha_i \alpha_j - \sum_{j=1}^{n} \alpha_j$$
$$s.t. \sum_{i=1}^{n} y_i \alpha_i = 0 \tag{2-72}$$
$$0 \le \alpha_i \le C_+, i = 1,2, \dots, n_1, y_i = +1$$
$$0 \le \alpha_i \le C_-, i = 1,2, \dots, n_1, y_i = -1$$

2.3 模糊支持向量机

为了使支持向量机能更重视少数类的正样本，让少量的正样本在模型训练过程中发挥作用，提出了模糊支持向量机（Fuzzy Support Vector Machine, FSVM），该方法基于样本的确定性来确定模糊隶属度，给具有更高确定性的样本分类更大的模糊隶属度，使得不同的样本对分类器做出不同的贡献，一定程度上保证了少类的正样本在模型训练中起的作用。

2.3.1 模糊集和模糊隶属度函数

在模糊逻辑中，模糊集是指具有模糊属性的对象的集合，用来确定模糊集特征的隶属度函数，依据不同领域表达形式具有背景差异化的特点。因此，隶属度的确定方式具有不同的形态：

（1）在基于神经网络的脊柱侧弯矫形器设计模型中，神经元间传递权隶属度函数通过神经网络的学习能力自行修正，辅助医师设计脊柱侧弯矫形器[47]。模型的第四层表示输出向量的隶属函数，其表达式为：

$$R_i: \text{If } x_1 \text{ is } A_1^i \text{ and } x_2 \text{ is } A_2^i \ldots \text{and } x_k \text{ is } A_n^i \\ \text{then } y^i = a_0^i + a_1^i x_1 + \cdots + a_k^i x_k (i = 1,2,\ldots,l) \tag{2-73}$$

其中，A_n^i、a_n^i为模糊规则的参数。

（2）在解决数字图像效果不佳等问题时，文章[48]利用图像灰度级作为论域确定模糊隶属度，提出卷积神经网络下的基于图像灰度级的模糊熵对图像边缘灰度值进行计算。在此方法中提到的图像的模糊集合的隶属度$u_m\big(x(i,j)\big)$和表示集合中的模糊属性的模糊熵H_m分别为：

$$u_m\big(x(i,j)\big) = \frac{1}{1 + x(i,j) - m/C} \tag{2-74}$$

$$H_m\left(u_m\big(x(i,j)\big)\right) = -u_m\big(x(i,j)\big)\log_2 u_m\big(x(i,j)\big) \tag{2-75}$$

其中，m 代表伸缩因子参数，C 代表常数, 模糊熵H_m随$\big(x(i,j)\big)$的大小变化而变化。测量图像灰度模糊集的隶属度函数为：

$$E_m(A) = \frac{1}{M \times N}\sum_i^M \sum_j^M H_m u_m\big(x(i,j)\big) \tag{2-76}$$

2.3.2 基于熵的模糊隶属度支持向量机模型

在模糊支持向量机方法的具体应用中，关键在于如何确定模糊隶属度。香农定义的熵的理解表示一个系统的“内在混乱程度”。随着信息论的发展，文章[49][50]等采用基于信息熵的模糊隶属度度量方法，提出利用熵来评价每个样本的类确定性，该方法根据样本的类别确定性来确定样本的模糊隶属度，正样本分配较大的模糊隶属度，负样本的模糊隶属度基于它的类的熵来决定，如若该样本的类确定性较低，在训练过程中容易误导分类决策面，则弱化其在学习过程中的作用。这样，会达到强化少类的正样本在训练过程中被重视的效果。

假设训练样本$\{x_i, y_i\}_{i=1}^N$，$y_i \in \{+1,-1\}$，$y_i = +1$表示样本x_i属于正样本，否则表示样本x_i属于负样本。用p_+和p_-表示样本x_i在约定集合内属于正类或者负类的概率，样本x_i的熵为

$$H_i = -p_{+i} In(p_{+i}) - p_{-i} In(p_{-i}) \tag{2-77}$$

同一模糊集下的p_+和p_-数值越接近，说明样本点所在的集合内正、负样本存在数量相当。已知，距离样本边缘越近的点对噪声越敏感，其所在类别的确定性越低。通过具体的数值代入上式比较可知，p_{+i}、p_{-i}越接近得到对应的熵也越大。因此可得到结论，熵越大的样本越靠近样本分类边界。

在基于熵的模糊支持向量机方法中，将负样本分为 m 个样本子集合，每个子集合Sub_l中的样本模糊隶属度相同，表示为：

$$FM_l = 1 - \beta * (l-1), l = 1,2,\ldots,m \tag{2-78}$$

其中，模糊隶属度参数$\beta \epsilon (0, \frac{1}{m-1}]$小于等于正样本的权重。总的训练样本$x_i$的模糊隶属度如下定义：

$$s_i = \begin{cases} 1.0 & \text{if } y_i = +1 \\ FM_l & \text{if } y_i = -1 \,\&\, x_i \in Sub_l \end{cases} \tag{2-79}$$

把上述的基于熵的模糊隶属度参数代入到支持向量机算法中，得到基于熵的模糊隶属度支持向量机方法（Entropy-based Fuzzy Support Vector Machine，EFSVM）：

$$\begin{aligned} &\min_{w \in R^d} \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{N} s_i \xi_i \\ s.t. \quad & y_i(w^T \varphi(x_i) + b) \geq 1 - \xi_i, i = 1,2,\dots,N \\ & \xi_i \geq 0, i = 1,\dots,N \end{aligned} \tag{2-80}$$

C 为正则化参数，ξ为松弛变量x_i的约束。约束条件中，w 是决策函数的权值向量，$\varphi(x_i)$表示x_i映射到更高的维度特征空间的非线性函数。在不平衡数据集上实验验证，EFSVM 算法提出的基于熵的模糊算法提高了确定性样本重要性的隶属度，保证了正样本在确定决策面时的作用。当然，这个方法也存在不足：平等对待的子集合训练点中如果包含噪声或者孤立点，这些点也会被给予一样的权重，会使算法过分敏感，出现过拟合的情况。

2.4 本章小结

本章内容主要描述了支持向量机算法的相关理论基础，其中包括算法的基本原理、推导、优化过程和几种核函数内容的介绍。另外，针对基础支持向量机算法面对不平衡数据集的分类问题存在的不足，介绍了有偏支持向量机和基于熵的模糊支持向量机两种方法，两种方法在一定程度上了弥补了标准支持向量机算法的缺陷。在介绍基于熵的模糊支持向量机方法之前，简单说明了熵的两种定义方法。

第三章 支持向量机和人面对不平衡问题的不同表现能力

传统分类器算法的设计基于分类样本的数量、分布特点整体上均处于一个相对平衡的状态而来，用来进行训练和测试模型的训练集、测试集也是基于此。在面对不平衡问题时，直接用传统的机器学习方法进行训练，测试得出的结果往往不如所愿。对比之下，在小样本和不平衡样本环境中，人可以简单直接地捕捉样本的特征，通过少量的训练样本获得样本概念，甚至具备快速迁移学习的能力。

本章第 3.1 节通过实验对比分析了不平衡数据集对支持向量机方法的影响程度和影响因素，总结了适用于评价不平衡分类算法性能的评价指标；第 3.2 节结合已有研究案例说明人具有的某种生理、心理认知特点在面对小样本、有偏差的样本中快速准确学习的优势；第 3.3 节对本章内容进行总结。

3.1 不平衡问题下的支持向量机方法

数据的不平衡一般分为类内不平衡和类间不平衡两种状态。类内不平衡，指某子类别与其它子类别的样本数量分布密度呈现不平衡或者是少数类的样本存在形式具有碎片化特点的情况。比如，某一类别的样本在不平衡数据集中的分布不均匀，在数据的分布形式中呈现多个“零散”的同样类别的子集合。

类间不平衡，指数据样本量的不平衡，在二分类问题中表现为数据中的一类种样本的数量远远少于另一种样本的数量，使得不同类别样本的数量不平衡。这种不平衡的表现形式有以下两种，一种是数据的绝对不平衡，另一种是数据的相对不平衡。绝对不平衡，指的是由于数据的来源特殊（军事、医疗等）导致难以获得大量数据，以致数据量真正的稀少，作为样本不能够全面地表达该类信息，难以形成代表自己类别的分类规则，从而影响支持向量机算法对少类样本的分类精度；相对不平衡，数据中的正样本数量本身不是想象中的那么少，且数据量丰富，但与多数类的负样本数量相比，其在样本中占的比例过小，这种稀缺导致传统分类器为了提高整体的分类精确度，而误将正样本归类为负样本，最终降低了分类器的分类效果。

接下来针对类间不平衡设计对比试验，从数据绝对不平衡、数据相对不平衡以及样本容量的角度出发，通过仿真显示这些不平衡对支持向量机算法的分类效果的影响，其中数据来源于 numpy 中的 random.RandomState 伪随机发生器

随机生成，其他无关参数保持默认值不变。图 3-1、3-2 中棕色点表示正样本，蓝色点表示负样本。

3.1.1 数据绝对不平衡对支持向量机算法的影响

绝对不平衡情况下，设置正样本数量为 50 个，改变负样本的数量：50、100、200、500、800、1500，得到六组不同程度的绝对不平衡数据，这六组数据的不平衡程度分别为 1、2、4、10、16 和 30。探究不同程度的绝对不平衡下，支持向量机算法对小数量不平衡数据集中正、负样本的分类情况，如下图所示：

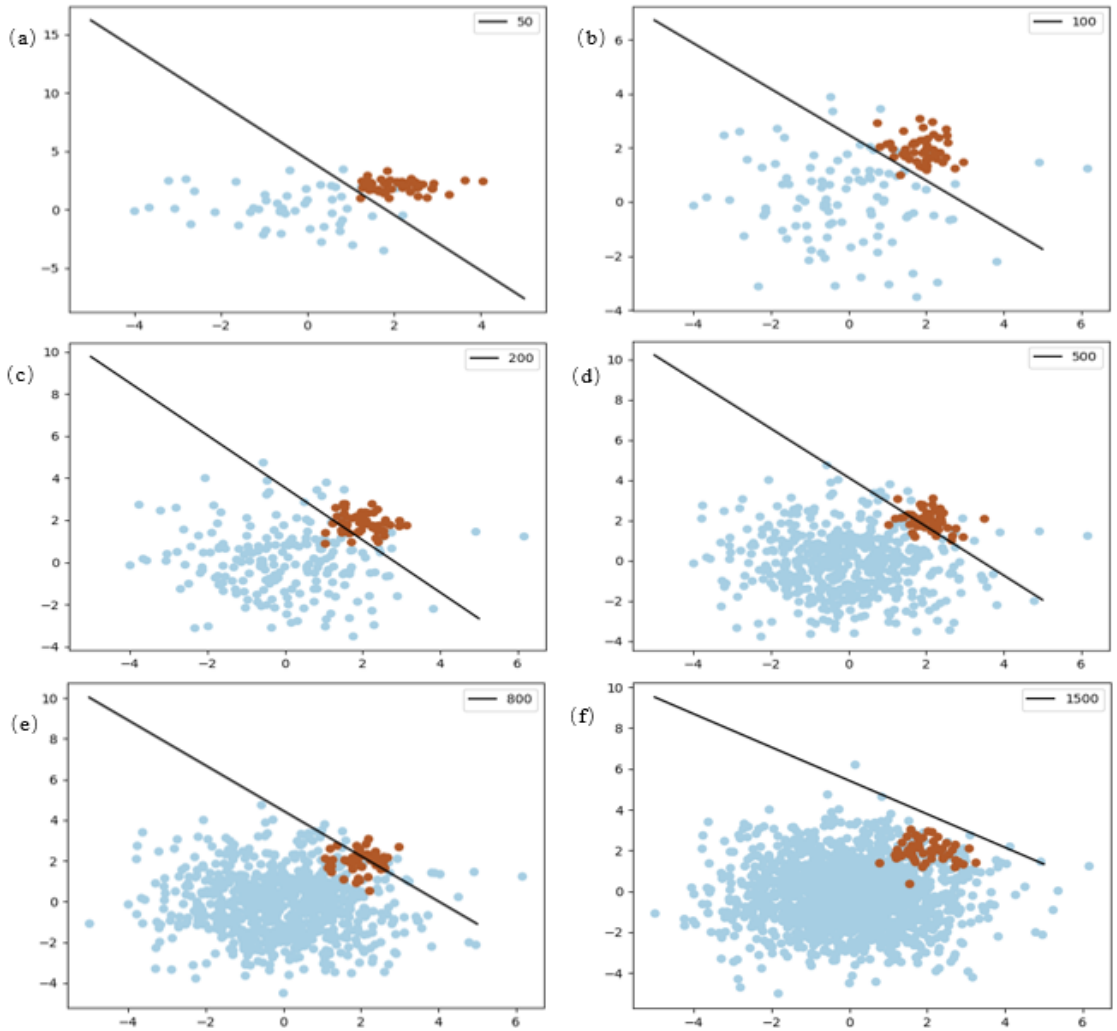


图 3-1 绝对不平衡对支持向量机分类性能的影响；（a）（b）（c）（d）（e）（f）分别为负样本数量为 50、100、200、500、800、1500 的分类结果

Figure 3-1. Effects of absolute imbalance on classification performance of support vector machines; (a), (b), (c), (d), (e) and (f) are respectively negative results of 50, 100, 200, 500, 800 and 1500 samples

由图 3-1 的（a）（b）可知，当两个类别的学习样本数量不等时 SVM 分类器的性能不一定会降低。由图 3-1 的这组全部的图可知，在数据集正样本量较小的情况下，随着负样本量的不断增加，数据集的不平衡程度也随之不断增加，分类边界越来越倾向于类别多的负类一侧，使正类的分类结果的准确性不断下

降。特别地，当不平衡程度逐渐增大到一定程度，支持向量机算法对正样本的正确率降低到0。因此，在设计分类器时，给负样本添加一个错分代价参数在一定程度上能减少大量的负样本对决定分类边界的干扰程度，但为了避免分类边界向正类过度偏移，这个错分代价参数的选取需要推敲。

3.1.2 数据相对不平衡对支持向量机算法的影响

相对不平衡情况下，设置正样本数量为 10000 个，改变负样本的数量：10000、20000、50000、80000、100000、150000，得到六组不同程度的绝对不平衡数据，这六组数据得到的数据集不平衡率分别为 1、2、5、8、10 和 15（因算力限制，此处的不平衡率最高设置为 15）。探究不同程度的绝对不平衡下，支持向量机算法对较大数量数据集下的不平衡数据集中正、负样本的分类情况，如图 3-2 所示：

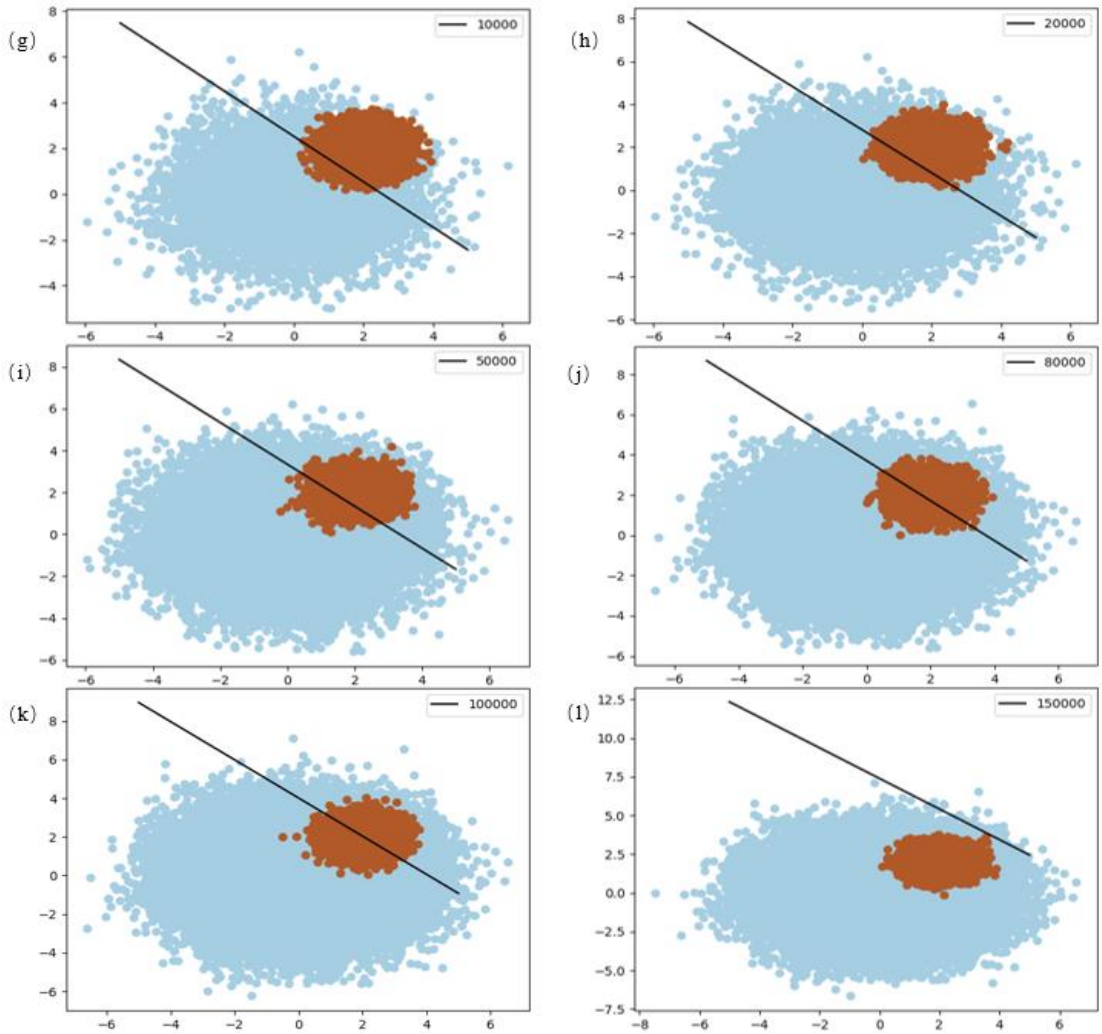


图 3-2 相对不平衡对支持向量机分类性能的影响；（g）（h）（i）（j）（k）（l）分别为负样本数量为 10000、20000、50000、80000、100000、150000 的分类结果

Figure 3-2. Influence of relative unbalance on classification performance of support vector machines; (g), (h), (i), (j), (k) and (l) are negative classification results with sample numbers of 10,000, 20,000, 50,000, 80,000, 100,000 and 150,000, respectively

由图 3-2 可知，在正负样本数量均较大的数据集下，随着不平衡程度的增加，数量占优势的负样本分类精确度也随之增加。而正样本的分类精度随着负样本数量的增加而降低，当不平衡率达到 15 时，所有的正类都被错误分类为负类。不过我们也能发现，随着正样本容量的不断增加，数据集中关于正类样本的边界信息也更加完善，一定程度上减弱了绝对不平衡的类间不平衡对支持向量机算法分类性能的影响。因此，在学习样本容量越大的情况下，数据的相对不平衡对支持向量机算法的性能影响较小。

考虑到数据不平衡的两种状态的实验验证，可以得出结论：数据不平衡对支持向量机算法的性能影响，绝对不平衡样本分类的影响要大于相对不平衡。因此，我们研究的数据不平衡问题主要是针对小样本情况下的不平衡分类问题。

3.1.3 不平衡问题影响支持向量机的其他因素

类内不平衡和类间不平衡这两种影响支持向量机算法的情况，从本质上仔细划分的话归纳为四类，分别是样本稀少、样本边界重合、数据碎片化和类内子聚集。另外，传统的评价机器学习性能的指标在面对不平衡数据学习时显得不适用。这些因素，也是分类器针对不平衡问题学习的难点所在。

（1）少数类样本稀少。因样本环境等因素的影响，增加少数类样本的数量存在成本较高或根本不存在可能性的问题。真正数量上的绝对稀缺导致含有的样本特征不够丰富，分类器难以在极度有限少的训练样本中形成、总结出清晰的分类规则；数量上的相对稀少，导致分类器为了获得全局分类准确率“牺牲”了少数类样本的分类精度，将少数类的样本错分到多数类样本一侧。

（2）类内子聚集导致的不平衡。类内不平衡的子聚集现象主要表现为少数类的样本集合在整体数据集中的分布呈现“碎片式”、“不均衡”等特点。有些碎片化分布的少数类样本，子集合样本容量足够小或者距离样本空间太远，例如下图所示的蓝色样本，导致分类器在学习过程中被忽视当作噪声样本进行处理，使原本数量少的少数类样本信息变得更为稀少，更深层次的影响分类器对少数类样本的分类效果。

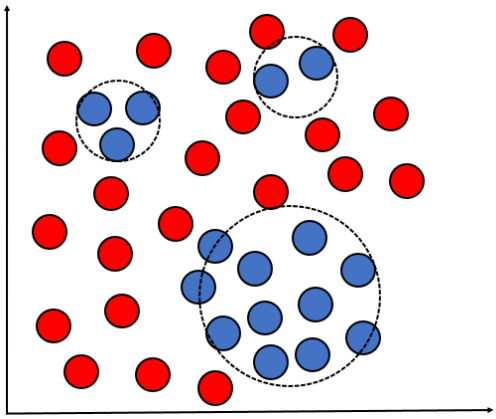


图 3-3 类内子聚集

Figure 3-3. In-class subaggregation

（3）数据重叠和噪声。在二分类不平衡分类问题中，少数类的正样本样本空间与多数类的负样本样本空间存在重叠，使得分类器难以识别和分类边界样本，这给分类器在确定分类边界时带来一定程度的干扰，影响分类结果。

（4）评价指标不恰当。在不平衡数据集中，少数类的样本数量在整个数据集集合中所占的比例较小，以传统的评价指标作为评价处理不平衡问题的分类器，将会把少数类的样本错分为多数类一侧的结果并不会影响评价指标。而在不平衡问题中，当我们需要重点关注少数类的样本时，以“牺牲”少数类的分类效果提高整体分类效果的评价指标显然是不可取的。

3.1.4 不平衡问题分类方法的评价标准

针对构建好的分类算法，需要选择一个合适的评价指标对算法的性能进行评估。通过性能评估，清楚模型的泛化能力，得知模型的好坏，进而对于相对差的模型进一步调参优化。对于普通的相对均衡的数据集分类问题来说，通常使用预测准确率(accuracy，ACC)作为其分类标准，表示样本中被正确分类的比率，其具体数学表达式如下所示：

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{3-1}$$

式中各部分代表的含义如下表混淆矩阵所示，TP 表示真正类（True positive）、FN 表示假负类（False negative）、FP 表示假正类（False positive）、TN 表示真负类（True negative）：

表 3-1　混淆矩阵

Table 3-1. Confusion matrix

	预测为正类	预测为负类
事实为正类	TP	FN
事实为负类	FP	TN

对于不平衡数据集，考虑到不平衡分类问题的特殊性，仍然使用 ACC 作为分类器性能评判标准是不适用的：当数据集中的数据不平衡程度比较高时，如果分类器把所有的样本全部预测为负样本，TN 将逼近甚至等于所有样本总和，此时的 ACC 值非常高；然而，数据集中的少量正样本也被预测为负样本，这样的分类器在解决现实问题时可能会造成不可估量的损失。因此，为了能更好的衡量解决不平衡分类问题的分类器的性能，我们需要一些复杂的全面的特殊的指标指标，研究者制定了一套新的适用于不平衡分类问题的评标指标。

G-mean 度量具有独立于类之间示例分布的独特特性，分别对少类的正样本和多类的负样本的分类情况采用不同的评价标准，来评估一个学习算法的综合性能。在使用 G-mean 评价指标对不平衡分类评判时，应用了灵敏度(sensitivity，

SE)和特异度(specificity，SP)两个概念，灵敏度和特异度对少量的正样本的预测效果做出评判：

$$SE = \frac{TP}{TP + FN} \tag{3-2}$$

$$SP = \frac{TN}{TN + FP} \tag{3-3}$$

$$G\_mean = \sqrt{SE \times SP} \tag{3-4}$$

SE 表示在所有被预测为正样本的样本中，为真正样本的情况，SP 表示所有预测为负样本的样本中，真负样本占的比率。G-mean 的大小受 SE 和 SP 的共同影响，只有正负样本分类准确率较高时，灵敏度和特异度的值才会特别大，相对应 G-mean 的值才会高。由公式可知 SE 和 SP 这两个值不会因为样本不平衡受到干扰，自然 G-mean 也比较稳定不会受到样本不平衡程度的干扰。如果分类器只是对少数类的正样本或多数类的负样本分类正确，另一类样本的分类准确率较低，G-mean 值也较低。因此，G-mean 用来作为评价不平衡分类的性能标准是合适的[51]。

除了 G-mean 作为不平衡分类的评判标准，F-measure(亦称 F-score)也是常用的评判标准之一。F-measure 结合准确率(precision)和查全率(recall)作为分类有效性的度量，根据用户设置的系数对查全率或查全率的加权重要性的比率。因此，F 度量比准确度度量更能深入了解分类器的性能[2]。

$$\text{precision} = \frac{TP}{TP + FP} \tag{3-5}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{3-6}$$

$$F_\beta - \text{measure} = \frac{(1 + \beta^2)\text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}} \tag{3-7}$$

precision 表示在所有被预测的正样本中为真正样本的概率，recall 表示在所有为正样本的样本中被预测为正样本的程度。准确率和查全率这两个指标有时是互为矛盾的：样本不是预测为正类则为负类。要想获得较高的准确率，那就要降低一点查全率。于是，引入β对两者进行加权平均，β取不同的值，相应的两个指标对 F-measure 的影响程度不同。当β大于 0 时，F-measure 的结果更倾向于准确率；当β小于 1 时，结果更看重于准确率，β通常取值为 1，此时的 F-measure 是标准度量，准确率和查全率同样重要。当准确率和查全率其中一个比较低时或者两者都比较低时，F-measure 值也比较低。只有当分类器的准确率和查全率都很高时，才能得到较高的 F-measure。因此，F-measure 能正确地评判不平衡样本的分类器的性能。

ROC[51]，全称为 Receiver Operating Characteristic Curve，中文名称为“受试者工作特征曲线”，最早用于雷达信号检测，区分信号与噪声。这个曲线的横坐

标表示正样本的错误预测率(False Positive Rate，FPR)、纵坐标表示正样本的正确预测率(True Positive Rate，TPR)：

$$FPR = \frac{FP}{FP + FN} \tag{3-8}$$

$$TPR = \frac{TP}{TP + FN} \tag{3-9}$$

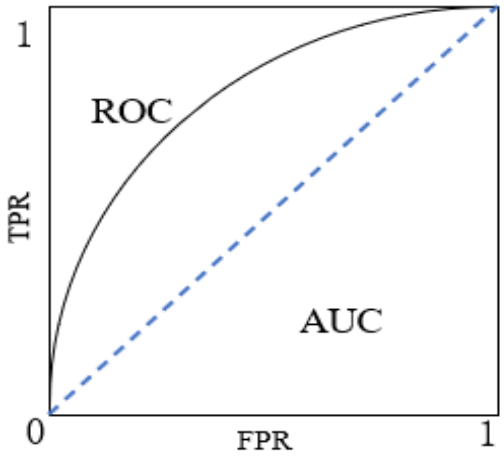


图 3-4 ROC 曲线线下曲线面积 AUC

Figure 3-4. Area under ROC curve AUC

理想状态下，TPR 趋近于 1，FPR 趋近于 0。ROC 曲线上的每一个点都有一个对应的阈值，此阈值下会有与之对应的 TPR 和 FPR。当阈值很大时，TP 和 FP 均为 0，对应于 ROC 曲线的坐标原点；当阈值最小时，TN 和 FN 为 0，对应于右上角的点(1,1)。所以，ROC 曲线的取值一般位于(0,0)到(1,1)之间。当 TPR 值越大同时 FPR 值越小，即 ROC 曲线上凸越明显，对应的模型的性能越好[52]。通过以上分析，FPR 和 TPR 分别基于正样本的预测结果来判别，避免了样本不平衡分布的影响。选取 FPR 和 TPR 作为 ROC 的性能指标，ROC 的评判标准不会受样本不平衡干扰。

但是 ROC 曲线只能形象化地描述不平衡样本分类的性能，不能对其进行量化的评价。为了能够量化的评价分类器的性能，提出 AUC[52]这一概念，它全称为 Area Under Curve，中文名称为曲线线下面积。顾名思义，线下面积表示 ROC 曲线下方与坐标轴围成的面积。曲线线下面积 AUC 越大，代表 ROC 曲线向左上凸出的程度越明显，说明正样本错误预测率低时对应的正样本正确预测率高，与我们追求的尽可能的提升正样本的预测精度的目标一致，即 AUC 的值越大，模型的性能越好。

3.2 不平衡问题下人的行为特性

传统的机器学习算法需要数量巨大且相对平衡的样本来进行训练才能获得较为理想的性能，这对算法本身的建立、算法的算力都带来不小的了挑战。况且在实际情况中样本的规模和平衡程度受到不同客观条件的约束，也相应地限制了机器学习算法在很多日常场景下的应用。相比之下，人作为一种精密的运转机器，本身携带的很多特质引起了大家的重视，越来越多的科学家试图从人身上寻找或总结出人类特有思维逻辑或心理变化的特征，进而对现有机器学习的算法进行改进，获得更多有效的方法提高机器学习算法的性能。

人的认知能力，即由于认知限制、动机因素或与自然事物和环境适应的过程中进化而成得人类共有的判断和决策中的系统性误差。虽说是一种系统性误差，但人的认知能力在人的思维、语言、交流等认知活动中发挥了明显的作用，是作为人类和动物或者机器区分开来的关键因素之一。这种认知能力不会因为应用环境中样本的规模大小或数据平衡程度而受到干扰，在很多应用场景中具有某种独特的优势。

人的认知能力的训练过程是发生在人所存在的真实情境中，机器学习算法的实际应用场景也是围绕人的活动场景发生，比如代替人标记图片、开车、玩游戏等。因此，从某种角度思考，人和机器学习算法的“训练场景”是相似甚至是相同的，训练后得到的“下意识”和“算法模型”所服务的对象也几近一致，也就是说，人进化训练的过程恰好与机器学习算法的应用场景之间具有某种一致性关系的联系（如图 3-5、图 3-6）。在机器学习算法中借鉴人的某种认知优势，把“人”的优势和“机器”的长处相结合，形成一种“1＋1＞2”的机器学习模型。

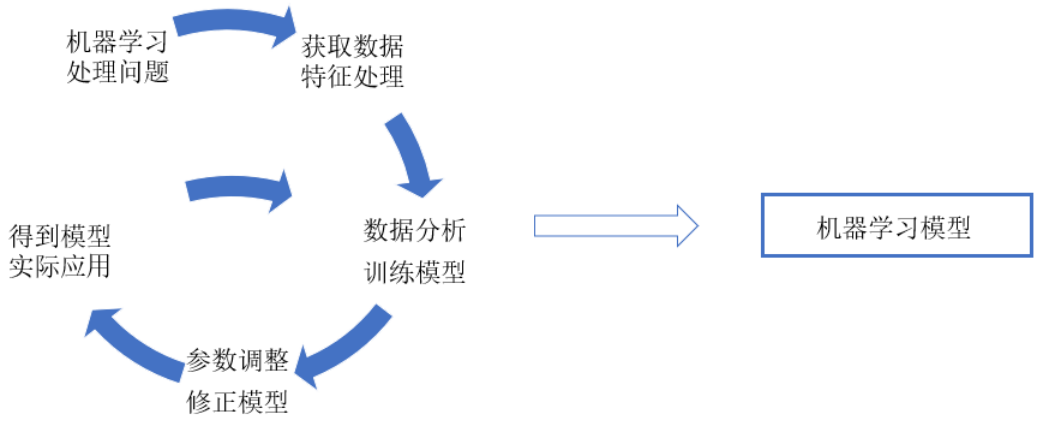


图 3-5 机器学习模型的建立流程

Figure 3-5. Machine learning model building process

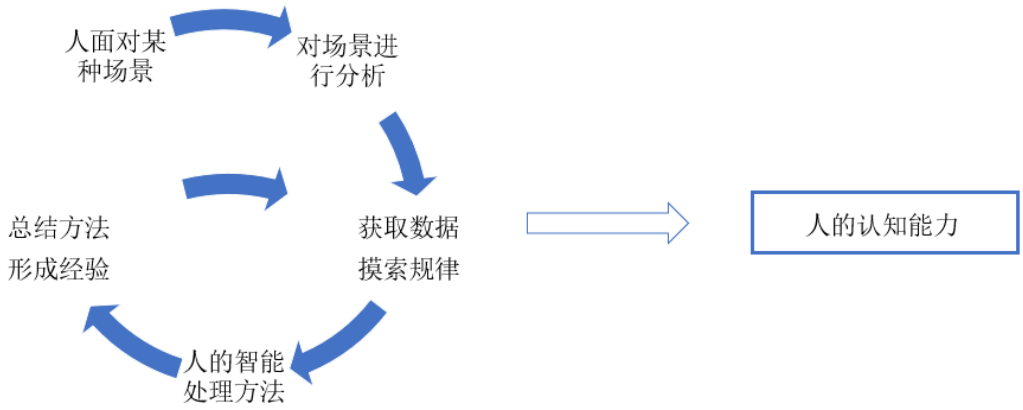


图 3-6 人的认知能力“进化”过程概括

Figure 3-6. Summary of the "evolution" process of human cognitive ability

根据认知特性的具体表现，暂把此文表述的人的认知特性简单归纳为人的生理认知特性和人的心理认知特性两类。人的生理认知特性，借助某种现代技术得到人脑思考、学习和决策时的生理结构变化，或是根据心电信号、肌电信号和脑血氧信号等的变化，通过解构这种生理结构变化了解人身体内部活动构建的一种认知模型。人的心理认知特性，从逻辑学出发的基于已有的知识储备作为先验概率对之后发生的事情进行预测的快速学习进而决策的过程。人的这两方面的认知特性都帮助人在面对少量的、不平衡的样本情境下，较为高效地分析情境、总结规律、解决问题。

3.2.1 人的生理认知特性及其应用案例

作为机器学习的灵感来源，人大脑的收集数据进行处理（比如拍照、开车、打乒乓球等）的过程，可以引导机器学习算法的提升。机器学习与神经科学和心理学领域也有越来越多的重叠，为了使机器学习更好地再现人的能力，在一些工作中能代替人类，现已有研究者提出了一种想法：是否可以通过测量大脑活动来明确地指导机器学习算法的训练，以使算法更像人类，从而改进机器学习算法，使机器学习更接近类脑特征。

人的脑部为人体最重要的脏器之一，生物学家们注意到人的大脑功能，映射到大脑皮层是按空间分区的，相同的，脑内次级结构也是按空间划分的，在不同的空间区域的脑组织结构负责处理不同的工作。功能性磁共振成像（functional magnetic resonance imaging，fMRI）可以利用磁振造影原理来测量人脑识别、阅读图片、文字时的神经元活动所引发的血液动力，通过血液动力的变化从而能够识别人此时所见到的图像或者阅读的词语时的脑成像。简单来说，指活体脑组织兴奋时，它实时地观测人脑特定的大脑活动的皮层区域的生理变化，描绘出脑组织兴奋的区域。

文章[37]利用人类受试者观看图像的功能磁共振成像记录指导图像机器学习面对图像分类任务的决策过程。通过来自于人类视觉皮层观看相同图像时的功能磁共振成像记录的值对个体训练图像进行加权，训练了四种视觉对象类别的监督分类模型(即人类、动物、地点、食物)，这些模型训练完成后，可以在没有神经数据帮助的情况下对图像进行分类。文章中的实验部分用到的视觉皮层功能磁共振成像的数据来于 Berkeley 实验室进行的一项研究：一名受试者观看 1386 张像素为 500*500 的自然场景下的图片的同时，一台视觉核磁共振仪器通过血液动力变化图像中的体素反应实时记录受试者的脑部活动信息，获得接受图片信息的大脑活动。通过对受试者观看图像时的脑部活动信息的建模和分析发现，某个区域的激活模式代表了神经元对图像上譬如形状、方向、颜色等信息引起的视觉刺激的反应。强烈的反应信号表明刺激更多地与特定的视觉区域有关，而较弱的反应表明刺激与它的关联较小。这放在机器学习的背景上表现为“训练权重”，在分类边界内的样本权重高，在分类边界外的样本权重低。

在机器学习中，损失函数用于对错误分类的数据进行惩罚，使算法最终损失最小化。铰链损失函数给所有分类错误的数据分配一个惩罚，这个惩罚与预测的错误程度成正比。关联性强的神经元和对应视觉刺激之间的函数关系纳入机器学习模型的训练过程中，重新定义与分类边界相关性强的数据分类错误时的惩罚力度，在传统的铰链损失函数基础上建立一种新的惩罚关系：

$$\phi_{\varphi}(x,z)=\max\big(0,(1-z)\times M(x,z)\big) \tag{3-10}$$

其中，

$$M(x,z)=\begin{cases}1+c_x, & \text{if } z<1\\ 1, & \text{otherwise}\end{cases} \tag{3-11}$$

式子中的 z 表示预测的正确性，$M(x,z)$为根据 fMRI 测量数据中发现的人类决策策略分配的错误分类的惩罚比例。c_x表示 x 对应的 fMRI 数据中得到的活动权重，基于 fMRI 的活动权重由特定类别中正类、负类样本图片相关的体素活动中获得，活动权重c_x值越大，表明获取 fMRI 图像的样本的置信度越高。与它的一般形式相比，新的惩罚关系对权重较大的样本分类错误时给予了更严厉的惩罚。

最后的实验也表明，直接从人脑测量的信息可以更好地帮助机器学习算法做出更“类人”的决定。脑神经科学与机器学习的跨学科融合，确实给我们打开了新的提升机器学习性能的思路。

3.2.2 人的心理认知特性及其应用案例

人类学习者不需要大量的负类样本来学习正类实例，可以从少量的、有限的正样本获得规则归纳出正样本的概念，并把这个概念很快地应用到下一个任务或情境中，这种快速学习新概念的能力被称为认知偏差。例如，通过第一次在动物园看到河马，婴儿可以获得关于河马的许多信息：它看起来像什么，它

有多大，以及河马与其他动物有什么不同的特征，从而较快速地获得河马的概念，下次见到形似这种概念的动物，可以呼之欲出“河马”。

相比之下，传统的机器学习方法需要大量的数据来解决相同类型的问题。在机器学习中，可能需要数百或数千个训练样本进行训练和学习，才能识别出“河马”。人类可以从单个类的样本中归纳出一个新概念，而机器学习需要通过许多数据和许多标签的训练才能达到类似或者相同程度的效果。试图把人的某种心理认知特性添加到机器学习算法模型中，为提高算法性能、缩小机器智能和人的智能的差距提供了一种新的思路。

经过神经科学研究的启发，文章[53]把上述人的心理认知特性添加到神经网络中，尝试从认知科学的角度再现神经元之间类似“一个神经元被激活的同时另一个神经元也会被激活”的特征。依据赫布理论提到的神经元持续、反复的被刺激使得神经元稳定性能够持续提升这一现象，添加了心理认知偏差模型的神经网络可以根据网络节点的状态进行“恢复”，调整神经网络传输节点的信号强度，（k-1）层的节点与 k 层的节点间的传递关系重新定义如下：

$$\Delta_{LS} w_{i,j}^{k-1,k} = -\alpha \delta_j^k y_j^k \left(1 - y_j^k\right) LS\left(y_j^{k-1}\right) \tag{3-12}$$

其中，$LS\left(y_j^{k-1}\right)$为基于人的心理认知特性的节点权重。在乳腺癌分类任务中，通过某种心里认知特性调整神经元传播过程中节点的值的新神经网络结构，与生理、神经科学和人类推理结果具有高度的相关性，提高神经网络在训练数据不足情况下的性能，弥补神经网络方法在训练数据数量不足或分布不均衡时性能下降的不足。它的表现优于其他具有代表性的机器学习方法的性能。

文章[54]在朴素贝叶斯方法中添加人的心理认知特性，一定程度避免了“朴素”条件在真实环境不适用带来的负面影响，添加后的模型在处理垃圾邮件分类问题时的准确度更高。把人的心理认知特性与神经网络、朴素贝叶斯算法的结合效果表明，与其他有代表性的机器学习方法相比，把人的某种认知偏差与现有的机器学习算法结合这种模型在小样本和不平衡样本的情况下取得了优异的性能。

3.3 本章小结

本章主要描写了支持向量机算法和人针对不平衡问题受到的影响和表现能力。分析了不平衡数据集对支持向量机算法的影响因素和原因，主要探讨了数据的类间不平衡和类内不平衡对算法的影响程度比较，也指出了传统的机器学习算法的评价指标不适用于评价不平衡问题分类器的性能并补充介绍了适合不平衡问题分类器的评价方法。另外，总结了人在少量的、不平衡的学习情境下的能力及其这种能力在机器学习方面的应用。

第四章 融合人的心理认知特性的模糊支持向量机算法设计

机器学习与神经科学和心理学领域也有越来越多的重叠，把人的智能引入到机器学习算法的训练过程中，是提高机器智能的一种新的研究思路。受人的认知能力不依赖大量训练样本的启发，本文将人的认知能力的某种模型添加到支持向量机的训练过程中，对位于不平衡数据集分类边界的样本赋予不同的学习权重，使分类器重视对少数量样本的学习，尽可能降低数据集不平衡对分类精度的影响。上章的 3.2 节举例说明了人的心理认知特性在具体场景中的有效应用。这章在理解这种心理认知特性的基础上，重新定义一种融合人的认知模型的模糊隶属度计算规则，设计一种融合人的心理认知特性的模糊支持向量机方法。

本章的 4.1 节介绍了人的心理认知特性的理论基础、特点以及由此推导出来的特征模型；4.2 节提出了融合人的心理认知特性的模糊支持向量机算法的设计思路、过程和得到的算法模型；4.3 节针对 4.2 节提出的方法进行数值实验验证，结果表明方法在提升不平衡分类问题精度的有效性；4.4 节对本章内容进行总结。

4.1 人的心理认知特性及其特征模型

4.1.1 人的心理认知特性：对称偏差和互斥偏差

偏差模型之所以能称为“模型”，需要满足以下条件：（1）具有一定的通用性，可以在各种情况下进行评估使用。（2）偏差程度可以根据不同的情境自动调整。针对人的逻辑学的研究进展中发现：人类的认知偏差能力有效地支持了概念的取得，这使得许多研究人员试图通过机器学习来再现人类层面的概念学习，对称偏差和互斥偏差这两种认知偏差可以有效地用于机器学习任务。

在生活中，如果外边的草坪湿了，人们会有“刚才是不是下雨了”的猜测；或者先看到下雨了，会想着外边的“草坪会湿”。之所以会有这样的联系，是因为下雨和草坪湿之间存在相互的因果关系。这种推论现象在我们的日常生活中随处可见，把因果关系中的“因”和“果”反向推导，由因得果、由果推因，把符合$B(q|p)\sim B(p|q)$特征的推测过程称为“对称偏差”，是指“如果p发生，那么q发生”联想得到“如果q发生，那么p发生”的结论。之所以称为“偏差”，在某种特定情况下，“下雨了”和“草坪湿了”之间的关系是不确定的，导致草坪湿的原因也可能是花洒刚给草坪浇过水。对称偏差以图 4-1 具体举例做进一步说明：当人通过把“苹果”“西瓜”这个标签分别与目标对应联系起来

后，通常也会根据目标来寻找对应标签。这个例子看起来行得通，但事实上有时并不成立，即使这样，对称偏差反而促进了人更快地学习和决策。

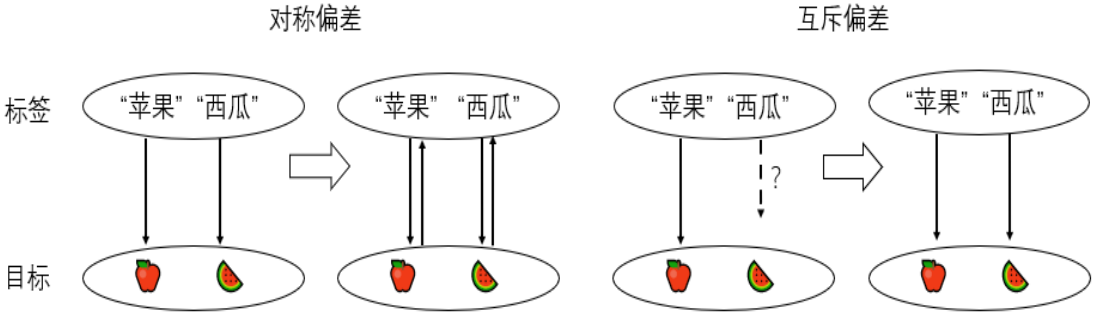


图 4-1 “对称偏差”和“互斥偏差”示意图

Figure 4-1. Symmetry bias and mutual exclusion bias

互斥偏差是另一种现象：若有“如果p发生，q就会发生”的前提，则人倾向于得到结论“如果p不发生，q也不会发生”。互斥偏差最早源于词汇学习研究，在小朋友学习新的词汇时，他通常会把新教的内容和他不熟悉的东西联系在一起，比如说，桌子上摆放了b、b'两种物种且他知道b的名字为a，那你问他“a'是哪一个”的时候，他通常会指向b'。在逻辑学中，互斥偏差有助于儿童的词汇学习。在我们的日常交流中，互斥偏差现象也很常见：一个妈妈告诉她的儿子，“如果你不打扫你的房间，你就不能出去玩”。在这句话中，p代表“不打扫房间”，q代表“不能出去玩”。她儿子听到这句话的理解是“如果我打扫房间了，我就能出去玩了”，于是她的儿子打扫了房间。用p和q表示他儿子的理解，$\bar{p}$表示“打扫房间”，$\bar{q}$代表“能出去玩”。在这个例子中，她的儿子好像误解了p和q这两句话的关系。但从效果上来看，母子两个人相互之间的沟通是成功的。我们把符合$B(q|p)$~$B(\bar{q}|\bar{p})$特征的认知逻辑称为互斥偏差。同样用图 4-1 中标签和目标的关系来解释：当为“苹果”的标签与目标联系好之后，标签为“西瓜”对应的目标自然会联系到另一个没确定的目标。互斥偏差有助于我们有效避免把不同的事物混淆起来。

如上所述，对称偏差和互斥偏差有时会导致不正确的逻辑，但人类在无意识中以适当的方式控制这种偏差的强度,这不仅没有影响到人正常沟通和学习，反而在我们的日常生活中得到了有效应用。如果我们能有效获得“适当的方式”，这两种认知模型将有助于机器学习的学习效果。

4.1.2 松散对称模型

为了便于后边的分析，定义概率模型的两种认知偏差和另外两个重要性质。对称偏差和互斥偏差两种对称认知偏差模型分别简称为 S（Symmetry bias）和 MX（Mutual exclusivity bias），相关的性质简称为 XM（excluded middle）和 ER（Estimation relativity）。用$B(q|p)$表示列联表 4-1 中p和q的因果关系进行概率

建模，表示p发生条件下人们主观上认为q发生的可能。当 B 满足表 4-2 右列的关系中的其中一个，认为 B 具有表 4-2 左列的四个性质：

表 4-1 因果估计列联表

Table 4-1. Contingency table of causal estimates

	q_j	$\bar{q}_j$
p_i	a	b
$\bar{p}_i$	c	d

表格 4-1 表示了事件p是否发生与其影响因素q是否出现之间的关系：表中有四个参数a、b、c和d，分别表示频率或者pq、$p\bar{q}$、$\bar{p}q$和$\bar{p}\bar{q}$同时出现的联合概率$P(p_i, q_j)$、$P(p_i, \bar{q}_j)$、$P(\bar{p}_i, q_j)$、$P(\bar{p}_i, \bar{q}_j)$。

表 4-2 两种认知偏差模型和相关的重要性质

Table 4-2. Two cognitive bias models and their important properties

名称	性质
Symmetry bias (S)	$B(q\|p)\sim B(p\|q)$
Mutual exclusivity bias (MX)	$B(q\|p)\sim B(\bar{q}\|\bar{p})$
The law of excluded middle (XM)	$B(q\|p)\sim 1-B(\bar{q}\|p)$
Estimation relativity (ER)	$B(q\|p)\sim 1-B(q\|\bar{p})$

其中，对称偏差性质和互斥偏差性质如上所述。所谓排中律，表示在同一逻辑思维过程中“非真即假”的规律，两种思想不能同时为假，即“要么 A 要么非 A”，表示在p的前提下，q要么发生要么不发生。估计相关性表示在q的发生与否与p是否发生无关，在决策过程中起到很好的作用[1]。对称偏差、互斥偏差、排中律和估计相关性之间的条件保持一致，如果B满足对称偏差，那就不仅满足$B(q|p) = B(p|q)$，同时也要满足$B(\bar{q}|p) = B(p|\bar{q})$、$B(q|\bar{p}) = B(\bar{p}|q)$和$B(\bar{q}|\bar{p}) = B(\bar{p}|\bar{q})$。估计相关性可以从互斥偏差和排中律这两个关系中推断出来，可以认为是两者的组合。

从对称偏差、互斥偏差、排中律和估计相关性的角度对称偏差和互斥偏差两个模型进行建模分析。借助条件概率建立对称偏差模型，用先前事件q的发生来预测后边的事件p，给定p条件下q发生的一般条件概率：

$$CP(q|p) = P(q|p) = \frac{P(p,q)}{P(p)} = \frac{P(p,q)}{P(p,q) + P(p,\bar{q})} = \frac{a}{a+b} \tag{4-1}$$

当a、b、c和d表示pq、$p\bar{q}$、$\bar{p}q$和$\bar{p}\bar{q}$同时出现的联合概率时，由以上推导过程同理能推断出：

$$CP(\bar{q}|\bar{p}) = \frac{\bar{p}\bar{q}}{\bar{p}\bar{q} + \bar{p}q} = \frac{d}{c+d} \tag{4-2}$$

此时发现，对称偏差 CP 模型满足排中律，但是并不满足对称偏差、互斥偏差和估计相关性。只有满足$a=d$和$b=c$条件时，通过a来识别d，用b来识别c，此时列联表和 CP 模型都对称，CP 偏差模型能同时满足对称偏差和互斥偏差。在 CP 模型中，把$a+b$作为整体(能出现的所有情况)只考虑了a当作分子的情况，忽视了事件不发生对结果的影响，即忽略了c和d对模型 CP 的影响。

互斥偏差 RS 模型把关注点放在$a+d$上，用$N=a+b+c+d$所有元素总和作为整体, 这种情况把其他可能造成相关影响的因素也考虑在内，$N=a+b+c+d$作为事件的总数，当它表示联合概率时，$N=1$。此时可以用对称偏差 CP 的绝对估计形式来表示 RS：

$$RS(q|p)=\frac{a+d}{a+d+b+c} \tag{4-3}$$

$RS(q|p)$不仅完全满足对称偏差、互斥偏差、排中律和相对估计，也完全满足贝叶斯规则：$a+d$大到接近$a+b+c+d$时，那么表示事件发生的概率接近于整体事件发生的概率。

对称偏差模型 CP 在表示人类的认知时，对称性不强，而用互斥偏差模型 RS 表示又对称性过大。在这里，通过具有参数和控制对称性的中间对称模型来参数化对称强度，使模型的偏差程度得到合理的调整，在决策或者其他任务中更人性化，公式表示为：

$$LS_{\alpha,\beta}(q|p)=\frac{a+\beta d}{a+\beta d+b+\alpha c} \tag{4-4}$$

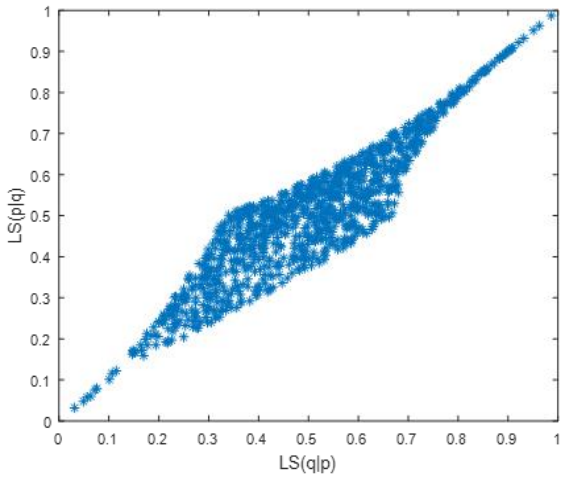


图 4-2 LS 模型的对称偏差特性

Figure 4-2. The symmetric deviation characteristics of LS model

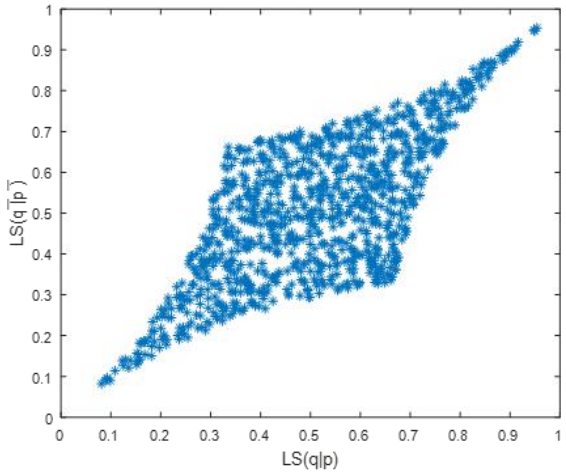


图 4-3 LS 模型的互斥偏差特性

Figure 4-3. The mutual exclusion bias characteristic of LS model

把这个基于概率推理的公式模型称作松散对称模型(Loose symmetry)，简称 LS 模型。松散对称模型通过α和β来调节对称偏差的绝对性和互斥偏差的相关性程度，这也有效避免了分母为 0 的情况出现。通过 Matlab 对 LS 模型建模得到图 4-2 和 4-3，不难发现$LS(q|p)$和$LS(p|q)$、$LS(q|p)$和$LS(\bar{q}|\bar{p})$均具有一定的相关性。当$LS(q|p)$处于中间值 0.5 附近时，$LS(p|q)$与$LS(\bar{q}|\bar{p})$相等，这一模型同时具有对称偏差和互斥偏差的性质。

4.2 融合人的心理认知模型的模糊支持向量机算法设计

在对非平衡数据进行分类时，数据集中的正样本数与负样本数之间存在明显的不对称性，导致分类结果倾向于偏向大量负样本的一侧。为了降低两个样本间的不平衡带来的影响，使分类算法在训练过程中弱化负样本的权重、更关注正样本，同时，又不过分弱化负样本的作用，尽可能保证负样本的分类精度。为解决以上问题，设计一种新的模糊隶属度定义方法，把人的松散对称模型引入到模型的训练过程中，保证正样本在训练中起的作用的同时，使算法能够根据不同的样本，尤其是负样本赋予不同的样本权重进行训练。

4.2.1 对不平衡数据中的负样本进行模糊化处理

在算法训练过程中，接近分类边界的样本会影响分类边界的确定，而这些样本点属于某一类别的置信度不高，考虑设计引入松散对称模型来确定样本的模糊隶属度。由于数据集中的正样本个数太少，使用 K 近邻法判断一个样本点的类别，这种判断的结果可能不完全正确。即使一个样本点被认为是负样本，实际上它也可能只是一个符合负样本特征的正样本。如果此时直接分类为负样本，则数据集中正样本的数量更少。

表 4-3 样本特征与样本类别间的因果关系

Table 4-3. Causal relationship between sample characteristics and sample categories

	正样本q	负样本$\bar{q}$
正样本特征p	a	b
负样本特征$\bar{P}$	c	d

在样本分布过程中，样本可分为如图 4-4 的四类：符合正样本特征中的正样本x_1、符合负样本特征的正样本x_2、符合正样本特征的负样本x_3和符合负样本特征的负样本x_4。根据这四种类型的样本周围样本分布情况，采用松散对称模型判断样本特征和样本类别之间的关联程度，并将其作为样本训练的学习权重。

表 4-4 a、d的大小关系和与之对应的样本类别判断

Table 4-4. Size relationship of a and d and corresponding sample category judgment

a和d的关系	该样本附近的其他样本分布分情况
$a > d$	$d = 0$ 、$a = 1$，附近的样本均为正样本
	$d \neq 0$，附近的正样本数量多于负样本
$a < d$	$a = 0$、$d = 1$，附近的样本均为负样本
	$a \neq 0$，附近的负样本数量多于正样本

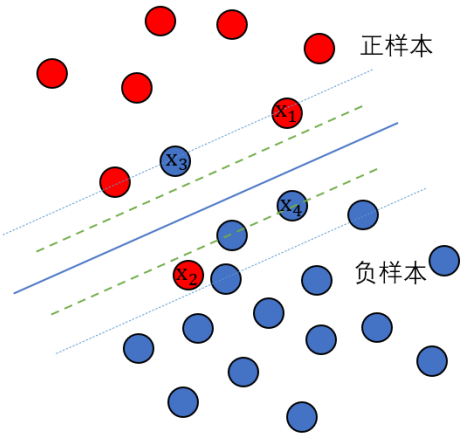


图 4-4 不平衡数据集分类边界局部样本分布可能

Figure 4-4. Unbalanced data set classification boundary local sample distribution possible

利用 K 近邻算法计算某个样本点距离最近的 k 个已知样本标签类别的样本的距离，统计这些样本中正、负样本的具体数值，其中正样本的数量用N_+表示，负样本的数量用N_-表示，通过概率公式得到在最近邻的 k 个样本中正、负样本所占的比例：

$$P(N_i) = \frac{N_i}{K} \tag{4-5}$$

$P(N_+)$表示在最近邻的 k 个样本中正样本出现的概率，以此表示表 4-3 列联表中的参数a；$P(N_)$表示在最近邻的 k 个样本中负样本出现的概率，上述列联表中的参数d用此来表示。通过 4.1.2 章节描述的排中律，我们已知$a+b=1$、$c+d=1$，也得到参数b和c的值。

接下来根据表 4-4，通过a和d的比较，对该样本附近的其他样本分布分情况进行讨论。当$a=1$、$d=0$时，该样本点附近的样本均为正样本，可以较为确定地判断该样本为正样本；当$a>d(d\neq 0)$时，该样本点附近的正样本数量多于负样本，这种情况下可能包含符合正样本特征的负样本，但为了保证少数的正样本数量，模型判断该点为正样本；为保证判断为正样本的样本点在训练学习过程中的被重视程度，对此类样本不进行模糊化处理，使其训练权重仍然保持为1。当$a=0$、$d=1$时，该样本点附近的样本均为负样本，可以较为确定地判断该样本为负样本；当$a<d(d\neq 1)$时，附近的负样本数量多于正样本，这种情况下即使存在符合负样本特征的正样本，但比起符合负样本的负样本的占比偏少，因此，模型判断这种情况下的样本点均为负样本。上述被判断为负样本的样本点，在训练之前进行模糊化处理，对不同的负样本赋予不同的模糊隶属度。综上，得到本文提出的融合人的心理认知特性的模糊隶属度计算规则：

$$\rho=\begin{cases}1, & \text{当}d<a\text{时}\\ LS(\bar{q}|p), & \text{当}d>a,d\neq 1\text{ 时}\\ LS(\bar{q}|\bar{p}), & \text{当}d>a,d=1\text{ 时}\end{cases} \tag{4-6}$$

$LS(\bar{q}|p)$是针对周围分布特点符合$a<d(d\neq 1)$条件的样本的模糊化计算方法，如式(4-7)，这样的样本点分布在靠近分类边界的负样本类别一侧，包括可能的负样本和具有负样本特征的正样本。由于整体的负样本数量明显多于正样本，位于分类边界的具有负样本特征的正样本出现的概率低于位于分类边界具有正样本特征的负样本出现的概率，将少量的处于分类边界的正样本当作负样本做模糊化处理不会加重数据集本身的不平衡程度。

$$LS(\bar{q}|p)=\frac{b+\frac{a}{a+c}c}{a+b+\frac{a}{a+c}c+\frac{b}{b+d}d} \tag{4-7}$$

$LS(\bar{q}|\bar{p})$为针对样本周围分布特点满足 $a=0$、$d=1$条件的样本模糊化处理，处理方法如式(4-8)所示，此时样本点周围全部为负样本，通过 K 近邻思想我们认为这个点一定为负样本。对负样本模糊化处理，降低其训练权重，此时能平衡负样本在样本中所占比例过高带来的负面影响。

$$LS(\bar{q}|\bar{p})=\frac{d+\frac{c}{a+c}a}{c+d+\frac{c}{a+c}a+\frac{d}{b+d}b} \tag{4-8}$$

通过人的认知偏差模型对不平衡数据中的负样本模糊化处理具体流程如下：

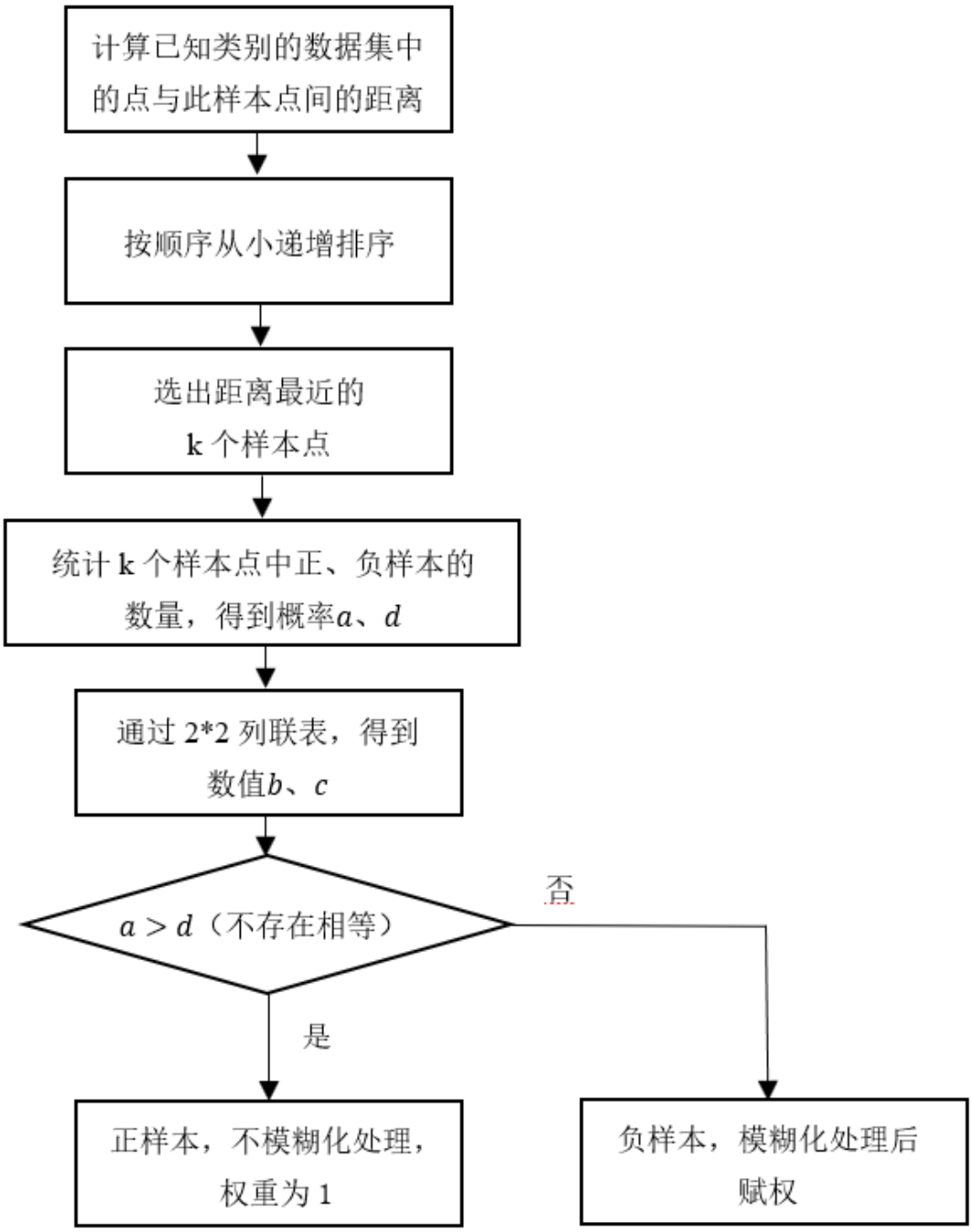


图 4-5. 利用认知偏差模型对负样本进行模糊化处理流程图

Figure 4-5. Flowchart of fuzzy processing of negative samples using cognitive bias model

4.2.2 融合人的心理认知模型的模糊支持向量机

前一节详细说明了如何使用松散对称模型对不平衡数据集中的样本进行模糊化处理。将融合认知偏差的样本模糊隶属度规则加入到支持向量机模型，得到一种融合人的心理认知的模糊支持向量机算法(Fuzzy membership support vector machine based on cognitive bias，简称 FSVM-BS)。

不平衡样本数据集中，正样本和可能的正样本学习权重为1，负样本和可能的负样本的权重根据公式（4-6）隶属度模糊化处理后的结果更新。训练集$S =$

$\{x_i，y_i，\rho_i\}_{i=1}^{N}$，其中$x_i$表示样本，$y_i \in \{+1,-1\}$表示样本标签(+1 表示正样本，-1表示负样本)，ρ_i表示公式(4-6)定义的基于松散对称模型的样本模糊隶属度。FSVM-BS 对不同的训练样本赋予不同的训练权重，使少量的正样本在训练过程中受到重视，找到一个能最大化限度分离正负样本的最有决策面：

$$\begin{gathered}\min_{w\in R^d}\frac{1}{2}\|w\|^2+\mathrm{C}\sum_{i=1}^{N}\rho_i\xi_i \\ s.t.\ \ y_i(w\cdot x_i+b)\geq 1-\xi_i, i=1,2,\ldots,N \\ \xi_i\geq 0, i=1,2,\ldots,N\end{gathered} \tag{4-9}$$

其中w是决策面的权重向量，b表示偏差，为了更方便找到最优决策边界，$\varphi(x_i)$表示非线性函数x_i映射到高维特征空间。C是正则化参数，ξ_i是x_i的约束变量。

为了解决二次优化问题，通过拉格朗日函数、KKT 条件引入拉格朗日乘子α_i对公式进行推导（推导过程参考第二章），

$$\nabla_w L(w,b,\xi,\alpha,\mu)=w-\sum_{i=1}^{N}\alpha_i y_i x_i=0 \tag{4-10}$$

$$\nabla_b L(w,b,\xi,\alpha,\mu)=\sum_{i=1}^{N}\alpha_i y_i=0 \tag{4-11}$$

$$\nabla_{\xi_i} L(w,b,\xi,\alpha,\mu)=C\rho_i-\alpha_i-\mu_i=0 \tag{4-12}$$

由此可得到，

$$w=\sum_{i=1}^{N}\alpha_i y_i x_i \tag{4-13}$$

$$\sum_{i=1}^{N}\alpha_i y_i=0 \tag{4-14}$$

$$C\rho_i-\alpha_i-\mu_i=0 \tag{4-15}$$

代入拉格朗日函数求极小极大问题，在这个过程中ρ_i被约掉，得到

$$\begin{gathered}\max_{\alpha}-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i \\ s.t.\ \ \sum_{i=1}^{N}\alpha_i y_i=0 \\ 0\leq\alpha_i\leq\rho_i C, i=1,\ldots,N\end{gathered} \tag{4-13}$$

其中，基于人的心理认知特性的模糊隶属度ρ_i作为拉格朗日乘子α_i的约束系数。得到基于人的心理认知特性的模糊支持向量机决策函数：

$$f(x)=\mathrm{sign}\left(\sum\nolimits_{i=1}^{N}\alpha_i y_i\,K(x,x_i)+b\right) \tag{4-14}$$

这个算法的具体实现流程如表 4-5：

表 4-5　FSVM-BS 算法实现步骤

Table 4-5. Implementation steps of FSVM-BS algorithm

输入： $\{x_i, y_i\}_{i=1}^{N}$，$y_i \in \{+1, -1\}$，**核函数** $\ker(x_i, y_i)$，**最近邻数** k **输出：拉格朗日乘子，偏差**b**和基于松散对称模型的模糊隶属度**$\{\rho_i\}$
步骤1：计算未知样本点到附近样本的距离；
步骤2：距离按从小到大的顺序进行排序，选取距离最近的k个样本点；
步骤3：统计k个样本点中，正、负样本的数量n_1、n_2；
步骤4：计算正、负样本在距离最近的k个样本中的概率$\frac{n_1}{k}$和$\frac{n_2}{k}$作为a、d；
步骤5：已知a和d，结合2×2列联表得到b和c的值；
步骤6：比较a和d的大小关系，得到样本类别，并基于人的心理认知特性模型计算样本点的模糊隶属度ρ；
步骤7：线性支持向量机原始问题；
步骤8：构建拉格朗日函数；
步骤 9：原始问题对偶化，则原始问题转化为公式（4-10）；
步骤 10：得到α、w和b；
步骤 11：得到融合人的心理认知特性的模糊支持向量机决策函数。

4.3 数值实验

为了证明 FSVM-BS 方法在提升解决不平衡问题的有效性，在 KEEL[56]（knowledge Extraction Evolutionary Learning）公共数据集，与标准支持向量机（SVM）、有偏支持向量机（BSVM）和模糊支持向量机（FSVM）模型进行比较：

（1）标准支持向量机（SVM）：标准支持向量机对不平衡数据集不采用任何校正方法；

（2）有偏支持向量机（BSVM）：有偏支持向量机对不平衡数据集中的正、负样本设置不同的惩罚系数；

（3）模糊支持向量机（FSVM）：模糊支持向量机根据不平衡数据集中的正、负样本的模糊隶属度设置不同的惩罚力度。

为了方便更多研究者和学生展开不平衡分类问题的研究工作，KEEL 归纳了真实环境中不同场景下的不同程度的非平衡数据集。本章节所采用的实验数

据，从 KEEL 公共数据集 Imbalanced data sets 类别下选取的 15 组不平衡程度不一的数据集，所用数据集的详细信息如表 4-6 所示，支持在 https://sci2s.ugr.es/keel/imbalanced.php 下载。

表 4-6　不平衡数据集基本信息

Table 4-6. Basic information about the unbalanced data set

数据集名称	IR	样本总数	正样本占比（%）	负样本占比（%）	特征维度
glass1	1.82	214	35.46	64.54	9
iris0	2	150	30	70	4
vehicle0	3.25	846	22.53	76.47	18
ecoli2	5.46	336	15.48	84.52	7
ecoli3	8.6	336	10.42	89.58	7
vowel0	9.98	514	9.11	90.89	8
ecoli-0-1-4-7vs 2-3-5-6	10.59	336	8.63	91.37	7
ecoli-0-1-4-6_vs_5	13	280	7.14	92.86	6
page-blocks-1-3_vs_4	15.86	472	5.93	94.07	9
glass-0-1-6vs5	19.44	184	4.89	95.11	9
yeast-1-4-5-8_vs_7	22.1	693	4.33	95.67	8
yeast5	32.73	1484	2.96	97.04	8
winequality-white-3-9_vs_5	58.28	1482	1.69	98.31	11
shuttle-2_vs_5	66.67	3316	1.48	98.52	9
poker-8_vs_6	85.88	1477	1.15	98.85	10

由表 4-6 中的正、负样本在样本总数中所占的比例可知，不平衡率 IR 和正样本率呈反比。选取的样本不同类别间的不平衡率跨度在 1.82 至 85.88 之间，用来探究不同不平衡率与 FSVM-BS 算法之间的联系以及 FSVM-BS 同其他算法的性能比较。具体的，上述 15 组数据集按不平衡程度（IR）分为：0-9、9-20、20-86 的低、中、高三个不平衡率层次的数据集，每层次 5 组数据集。

4.3.1 实验设置

在实验中，以标准支持向量机 SVM 作为基础学习模型。SVM、BSVM、FSVM 和 FSVM-BS 均采用高斯径向基（RBF）核函数：$K(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$，其中核参数$\sigma^2=\frac{1}{N^2}\sum_{i,j=1}^{N}\left\|x_i-x_j\right\|^2$，正则化参数 C 从集合$\{2^{-6},2^{-4},\ldots,2^4,2^6\}$中

选取。此外，用来概率估计的 K 近邻中的数值 k 从集合{3,7,15,33,51,67}中选择，在计算最近的数据距离时用欧式距离公式。BSVM 算法只需要调节一个参数 C；FSVM 算法由于不能忽略样本标签所指示的类，s_i规范在(0.5,1.0]区间内。所有实验都是在 Windows10，Intel（R）Core（TM）i5-7400 CPU @ 3.00GHz, 8.00GB 内存，PyCharm 2020.3 环境中进行。

4.3.2 实验结果

为了较为精确地衡量比较算法在不平衡数据集上的性能，采用 F-measure、G-mean 和 ROC 曲线下面积 AUC 三个评价指标对分类性能进行评价。在算法训练的过程中，每次试验取 9 份数据作为训练集，1 份数据作为测试集，每个数据集进行 10 次试验，取 10 次的平均值作为准确率的最终实验结果，并将每个评价指标下的最大值加粗标出。在下表中，对四种支持向量机方法进行了评价，FSVM-BS 算法在 F-measure、G-mean 和 AUC 值三个评价指标上具有明显的优势。

表 4-7　四种分类器的 F-measure 值比较（加粗为最佳结果）

Table 4-7. Comparison of F-measure values of four classifiers (best result in bold)

数据集名称	SVM	BSVM	FSVM	FSVM-BS
glass1	0.6756	0.6561	0.5682	**0.7584**
iris0	0.5613	0.5219	**0.5860**	0.5712
vehicle0	0.6424	**0.6721**	0.5655	0.6387
ecoli2	0.5601	**0.5602**	0.5565	0.5481
ecoli3	0.6105	0.5827	0.5937	**0.6354**
vowel0	0.7894	0.8264	0.8139	**0.8543**
ecoli-0-1-4-7vs 2-3-5-6	0.9037	0.8971	0.9118	**0.9126**
ecoli-0-1-4-6_vs_5	0.8815	0.7487	0.8426	**0.8911**
page-blocks-1-3_vs_4	0.7223	0.7542	0.7752	**0.7614**
glass-0-1-6vs5	0.6421	0.7328	0.7934	**0.8269**
yeast-1-4-5-8_vs_7	0.8909	0.8769	0.8597	**0.8945**
yeast5	0.8841	0.8915	**0.9364**	0.9243
winequality-white-3-9_vs_5	0.9114	0.9048	0.9231	**0.9521**
shuttle-2_vs_5	0.8427	0.8231	0.8549	**0.8772**
poker-8_vs_6	0.8689	0.8728	0.8963	**0.9011**

F-measure 的值由准确率和查全率共同决定，准确率和查全率的反比关系使准确率升高时查全率降低，设置公式（3-7）F-measure 的参数β为 1。通过表 4-7 四种分类器的 F-measure 值的统计结果可知，本文提出的 FSVM-BS 算法的 F-measure 在 15 数据集中的 12 个数据集中结果最佳，尤其是在 winequality-white-3-9_vs_5 数据集中，F-measure 的值高达 0.9521。并且，随着数据集的不平衡率的增加，FSVM-BS 的 F-measure 值没有明显的波动。当准确率和查全率都比较高时，F-measure 值会比较高，这表明 FSVM-BS 算法在不改变原本的样本信息时，仍能较为精准地进行分类任务。

表 4-8　四种分类器的 G-mean 值比较（加粗为最佳结果）

Table 4-8. Comparison of G-mean values of four classifiers (best results in bold)

数据集名称	SVM	BSVM	FSVM	FSVM-BS
glass1	0.3981	0.4605	0.5267	**0.6421**
iris0	0.5525	0.6038	**0.6153**	0.6093
vehicle0	0.6412	0.6459	0.6551	**0.6712**
ecoli2	0.7529	0.7814	0.7787	**0.7789**
ecoli3	0.7741	0.7868	0.7984	**0.8245**
vowel0	0.8021	0.8387	**0.8512**	0.8411
ecoli-0-1-4-7vs 2-3-5-6	0.8154	0.8425	0.8410	**0.8649**
ecoli-0-1-4-6_vs_5	0.7926	0.7714	0.7631	**0.7846**
page-blocks-1-3_vs_4	0.8281	0.8463	0.8684	**0.8941**
glass-0-1-6vs5	0.8542	**0.8649**	0.8559	0.8602
yeast-1-4-5-8_vs_7	0.8715	0.8653	**0.8862**	0.8734
yeast5	0.8467	0.8541	0.8346	**0.8705**
winequality-white-3-9_vs_5	0.8829	0.8934	0.9016	**0.9068**
shuttle-2_vs_5	0.6614	0.6768	**0.7129**	0.7020
poker-8_vs_6	0.8923	0.8725	0.8867	**0.8978**

通过对以上数据的观察可知，四种分类器中 FSVM-BS 算法在 10 个数据集中的 G-mean 结果为最佳，次之为 FSVM 算法，SVM 算法在所有算法模型中的表现最差。在 glass1 数据集，FSVM_BS 算法的 G-mean 较 FSVM 算法提升了 11.54%，在 winequality-white-3-9_vs_5 数据集中 FSVM-BS 的 G-mean 数值最高。已知，当被分类为正、负样本中真正、负样本占比越多时，G-mean 的值越高。

至此可得到结论，FSVM-BS 算法在解决不平衡数据学习问题时，正样本、负样本均被正确分类的能力优于一些其他的相关方法。

表 4-9　四种分类器的 AUC 值比较（加粗为最佳结果）

Table 4-9. Comparison of AUC values of four classifiers (best results in bold)

数据集名称	SVM	BSVM	FSVM	FSVM-BS
glass1	0.7661	0.7715	**0.8032**	0.6948
iris0	0.7461	**0.7659**	0.7543	0.7606
vehicle0	0.7462	0.8012	**0.8387**	0.7839
ecoli2	0.7834	0.7675	**0.8751**	0.8409
ecoli3	0.7761	0.7806	0.7779	**0.7801**
vowel0	0.9088	0.9372	0.9319	**0.9458**
ecoli-0-1-4-7vs 2-3-5-6	0.8875	**0.8964**	0.8805	0.8769
ecoli-0-1-4-6_vs_5	0.6492	0.7161	0.8081	**0.8176**
page-blocks-1-3_vs_4	0.6955	0.5567	0.7938	**0.8462**
glass-0-1-6vs5	0.6678	0.7812	0.7457	**0.8713**
yeast-1-4-5-8_vs_7	0.7317	0.7924	**0.8905**	0.8863
yeast5	0.6288	0.6987	0.8354	**0.8457**
winequality-white-3-9_vs_5	0.5936	0.6749	0.6771	**0.6980**
shuttle-2_vs_5	0.5782	0.6749	0.8066	**0.8417**
poker-8_vs_6	0.7495	0.7326	0.8015	**0.8532**

在 AUC 这部分结果中，AUC 值越高表示正样本错误预测率越低、正样本正确预测越高，FSVM-BS 算法在 9 个不同程度不平衡数据集中的正样本正确预测率较高于其他 3 种模型。其中，在 vowel0 数据集中 FSVM-BS 的 AUC 值在 15 组数据集中最好，达到 0.9458。即使在不平衡率达到 85.88 的 poker-8_vs_6 数据集中，FSVM-BS 的 AUC 值在 4 种模型中仍保持领先，保证了少量的正样本在分类边界中的作用。由此可得，本文提出的方法随着样本的不平衡率的增加，性能更为突出。

4.4 本章小节

本章主要介绍了人的互斥偏差和对称偏差两种认知偏差模型，阐述了以此为基础的松散对称认知模型，并提出了基于人的松散对称认知模型的模糊支持

向量机模型（FSVM-BS）。本章 4.1 节所述的人的心理认知模型是基于“人”宏观概念下的一种普遍性的认知能力，不受场景、人群认知水平、情绪等非理性因素的影响。4.2 节所提出的融合人的认知模型的模糊支持向量机算法结合了 K 近邻规则、人的心理认知特性和支持向量机算法，是一次新的尝试。根据 FSVM-BS 模型的算法流程设置数值实验，与其他解决不平衡问题的算法进行实验对比。通过 15 组不同程度不平衡率的数据集的实验介绍和结果分析，FSVM-BS 算法在 F-measure、G-mean 和 AUC 值三个评价指标的分析结果中由于 SVM、BSVM 和 FSVM 算法，证明了 FSVM-BS 算法在解决不平衡问题的有效性。

第五章 基于 FSVM-BS 算法的具体场景应用及其分析

在上一章中，提出的融合人的认知模型的模糊支持向量机(FSVM-BS)面对不平衡分类问题性能优于其他的算法。考虑到算法在实际场景中应用的有效性，结合在国内外新冠疫情风波还未全面平稳的大环境下，许多行业和人群的正常收入受到影响，部分信贷人员难以或无法履行还款承诺对银行、信贷机构造成负面影响的现象。本章内容利用真实的银行信贷风险预测数据集，通过 FSVM-BS 方法对申请信贷人员进行预测，检验 FSVM-BS 方法在真实环境中的表现能力。

本章的 5.1 节介绍了信用风险评估的研究背景和研究现状，表明信用风险评估在实际生活中的重要性；5.2 节针对实验所用的真实信用风险预测数据进行介绍和处理；5.3 节说明了 FSVM-BS 在信用风险预测数据集的数据实验，包括参数的选择和实验结果的对比，得出方法在真实的信用风险预测场景中的有效性。第 5.4 节对本章的内容进行总结。

5.1 信用风险评估的研究背景及研究现状

随着先进消费理念被越来越多的人所接受，以及先进消费平台的不断发展，申请信贷的人数也在不断增加。如果成功申请贷款的群体不能按时偿还贷款，将给银行和其他机构带来不可避免的损失。有必要在贷款申请者中有效地筛选违约人群。在本节中，试图用 FSVM-BS 方法在一个真实的不平衡银行信贷风险预测数据集中正确区分正常人群和失信人群。

5.1.1 背景介绍

信用风险，又称为违约风险，表示交易对方不履行到期债务的风险[56]。随着生活节奏的加快和物质需求的增加以及互联网平台技术的发展，越来越多的人加入“超前消费”的队伍中，用来购房居住、旅游增长见识、求学提升自身、综合消费或者用于创业投资，尤其受房价不断追高的影响，渐渐形成了以购房贷款为主题的贷款体系。网络平台的支付宝提出的蚂蚁花呗、借呗和京东白条以及微信出现的微粒贷开通的信用消费等信贷平台的出现，使加入超前消费的人群组成和借款形式越来越多样化。

中国人民银行于 2019 年 10 月 21 日发布的《中国普惠金融指标分析报告（2018 年）》中指出，我国的个人消费贷款呈现持续快速增长的趋势。截至 2018 年年底，我国全国人均个人消费贷款余额为 27089.4 元，同比增长 19.54%。

在中国人民银行2021年统计数据中可发现，截至2021年7月，金融机构信贷收支统计已达186万亿人民币，同比增长可达12%。

在信贷日渐受人追捧的大环境下，信用消费达成后，交易对方不愿意或者无能力偿还贷款金额（可能包含的利息），势必对银行、投资方或者其他交易方造成一定程度的损失。交易达成前能对交易对象进行合理准确的信用风险评估，针对借款人的个人信息、信用记录、生活状态、受教育程度和工作收入等个人信息进行评估，预估借款人的偿还能力和违约概率，这一定程度上有效避免了借款人不能按期还款对银行业、商业投资、互联网金融平台等金融机构的影响。突尼斯是受信贷风险影响较大的国家之一。根据世界银行(2015)的统计数据，突尼斯的银行不良贷款率比例从2009年的13.2%升至2014年的16.2%，这是一个很高的国际标准比例。因此，鉴于银行信贷是突尼斯的支柱在经济方面，突尼斯银行业被要求实施良好的风险管理实践。在这一视角下，突尼斯背景下的信用风险研究就显得尤为重要[57]。

在2008年全球金融危机以来以及现有新冠病毒疫情全球化的影响，信用风险评估越来越被重视。不管是实体银行或者是网络平台的贷款业务，都需高效、合理的信用评估体系。目前，一方面，我国缺乏完善的个人信用记录体系，查询个人信用记录的途径有限且查询到的个人征信记录信息有限；另一方面，我国在关于信贷风险的控制和管理上，针对贷款用户的信用评估方法乏善可陈，没有统一的评估标准和操作规范。在当下的信用评级机构中容易出现这些问题：基础工作不到位导致的信贷档案资料漏缺、审理贷款和申请贷款的分离机构没有严格分离导致贷款金额和日期与审批内容有出入、贷款前的调查审批流于形式使对借贷人的个人资产信用情况了解不实等现象。这给贷款申请人的贷款通过审核后期没有按期偿还带来了便利。

合理的信用风险预测系统不仅有效减少银行等金融机构的经济影响，同时避免借贷人可能出现的未按时还款带来的征信问题，影响其正常的生活、学习。

5.1.2 研究现状

近几十年来，随着信贷行业的快速发展，信贷风险评估越来越受到人们的关注。为了避免潜在的损失，信用风险评价模型必须具有较高的准确度来决定是否给予借款人借贷。从根本上来，信用风险评估可以归纳为一个二元分类问题。信用风险预测数据集通常是不平衡的，由于信用记录对以后借贷甚至就业等其他方面的影响后果，借贷人中其中正常还款的人数远高于到期违约的人数。如何通过少量的违约借贷人的信息对后续申请借贷人进行评估，将贷款申请人划分为能按时偿还的“好”贷款人或到期违约的“坏”贷款人是一个重要的问题，不少研究学者对此做出了努力。

风险评估模型一般分为三类：主观判断模型、统计模型和因果模型。主观判断模型，主要靠信用专家的主观判断，采用典型问卷、专家判断、专家系统

和模糊逻辑系统[58]。统计模型包括参数统计和非参数统计，参数统计模型常用回归分析、判别分析、Logistic 模型等方法；非参数统计模型主要使用决策树、神经网络、最近邻结点算法等。自 20 世纪 80 年代以来，我们进入了人工智能分析阶段，现代数据挖掘和机器学习技术的健康快速发展，它克服了传统分析方法在运用科学方法方面存在的问题，在准确评估信贷风险预测能力做出正确的决策方面提供了有效的帮助。目前，信用评估风险预测方法主要有基于分类或回归的方法包括：逻辑回归、神经网络评估系统、随机森林、支持向量机或集成方法等[58]。

James A.Ohlson 提出将 Logistic 回归方法应用到商业银行信用风险评估任务中，以预测公司破产概率事件作为研究证明，结果表明这种方法对于预测信用风险是有效的[59]。Chen Yan 等人采用 z-score 消去维数对所选风险指标标准化，并采用 Logistic 回归分析模型对标准化指标数据进行评价，得到企业违约风险的概率[60]。Jiexin Lu 等人通过 excel 和 Matlab 软件将原始信用记录和风险等级处理和整合，接着从中提取具有代表性的主成分因子作为自变量，进行二元 Logistic 回归分析，得到企业是否违约的评价，建立违约筛选模型，然后采用多元有序 Logistic 回归分析对未违约企业进行分析[61]。李淑锦利用羊群效应和性别、年龄等指标，共同建立一个关于个人的借贷人信用风险评估指标体系，通过 Lasso-Logistic 模型预测借贷人的违约风险[62]。实验结果表明，羊群效应在 Lasso-Logistic 模型中发挥了重要影响，使得准确性高于 Logistic 回归。

Jiboning Zhang 采用一种脱离了传统的基于历史数据回归分析的 Logistic 评价模型，人工智能的模糊神经网络[58]，这种模型适用于信息披露相对不完善、评级机制不全面的投资者进行信用风险评估，此模型随环境变化的学习能力强。杨月提出一种新的 RBF 神经网络模型预警方法[63]，同时与 BP 神经网络算法进行比较，实验证明起到了优化了信用风险模型的作用，合理预测了行业发展的信用风险状况。Mohammad Mahbobi 等人采用人工深度神经网络与重采样技术结合[64]，对不平衡数据进行信用风险评估，能较为准确地预测出其中地违约客户。

Shuang Pan 和 Sheng Zhou 提出一种基于随机森林和可视化图模型的 P2P 借贷信用风险评估方法，该方法对不同等级风险样本的识别准确率达到 98.63%[65]。Nisha Arora 等人提出用 Boostrap-Lasso 从特征池中选取出相关且一致的特征，所谓一致的特征即在不同的数据集中有较强的鲁棒性，这些特征应用于随机森林、支持向量机、朴素贝叶斯和 K 近邻算法中，其中在随机森林中的预测精度最高[66]。陈倩等 3 人构建了一种基于随机森林的 Elastic Net-Logistic 个人风险违约风险评估模型[67]，在南德信贷数据和澳大利亚信贷数据中都有很好的分类效果。

Paweł Pławiak 等人提出一种基于不同支持向量机分类器的深度遗传级联集成方法，在 Statlog Australian data 数据集中预测精确度高，可以用来银行系统评

估申请人的银行信用[68]。Jian Luo 等 3 人开发了一种新的无监督无核二次曲面支持向量机(QSSVM)模型，避免了对核和相关的 kernel 参数的选择，并设计了黄金分割算法来生成平衡和不平衡数据的合适分类器[69]。在基准个人信用数据和实际企业信用数据上的数值结果表明了该方法的有效性、效率和可解释性。Lean Yu 及其团队提出了一种基于模糊集理论(FST)和近端支持向量机(PSVM)的双权模糊近端支持向量机模型，模糊集理论的加入使 FPSVM 模型具有更好的通用性。在公开的信用数据集中进行测试，实验结果表明所提出的 FPSVM 模型优于文中所列的其他 SVM 模型[70]。

5.2 银行信用风险评估数据集介绍和处理

在银行信用风险评估的数据集中，按时还款的优质借贷人及其信息是多数的，违约不能按时还款的危险借贷人及其信息是少数的。银行信用风险评估的数据集符合不平衡数据集的特征。

5.2.1 数据说明

本部分实验所用的数据集为公开数据集：credit risk analysis，可用于分类、数据清洗、数据分析等任务。该数据集的创建者 Ramesh Mehta，AI 方面的数据专家，具有 15 年从事金融科技领域的工作经验。该数据集包含某金融机构 2007 年-2015 年发放的所有贷款的完整贷款数据，包括当前贷款状态（已通过，正在审核，已放款等）和最新的支付信息，当借款人不能及时还款、错过还款、逃避或停止还款时，则认定为违约，数据集中的“违约”和“不违约”分别用“0”和“1”表示两种标签类别表示。该数据集共包含 855969 条数据，包括目标变量在内的 73 个特征。该数据集违约贷款占比约为 5%，属于不平衡数据集。详细数据可以搜索网址 https://www.kaggle.com/rameshmehta/credit-risk-analysis?select=data.csv 进行登录下载。

5.2.2 数据描述和处理

分析数据的特征可以发现，数据集 credit risk analysis 中的数据分为分类型变量和数值型变量，其中 grade、homeowner 等数据类型为分类型。数据集共包含 855969 个样本标签，其中标签为违约样本(标签用 1 表示)有 46467 个，不违约样本(标签用 0 表示)有 809502 个，分别占总数据集的比例如图 5-1 所示，数据集的整体不平衡率约为 19，属于不平衡数据集，适用于不平衡问题分类任务。

因为数据来自于真实环境，数据集中的 emp_title、desc、mths_since_last_delinq、mths_since_last_record、initial_list_status、last_pymnt_d、next_pymnt_d、mths_since_last_major_derog、annual_inc_joint、dti_joint、verification_status_joint、tot_coll_amt、tot_cur_bal、open_acc_6m、open_il_6m、

open_il_12m 、 open_il_24m 、 mths_since_rcnt_il 、 total_bal_il 、 il_util 、 open_rv_12m、open_rv_24m、max_bal_bc、all_util、total_rev_hi_lim、inq_fi、total_cu_tl、inq_last_12m 等 28 个特征的相关数据可能由于涉及隐私问题等因素出现不同程度的缺失。本文意图构建一个在真实环境中能有效应用的模型，为了保证数据的真实性，不对这部分特征相关的数据进行“盲目”补充，而是对这部分特征进行剔除。

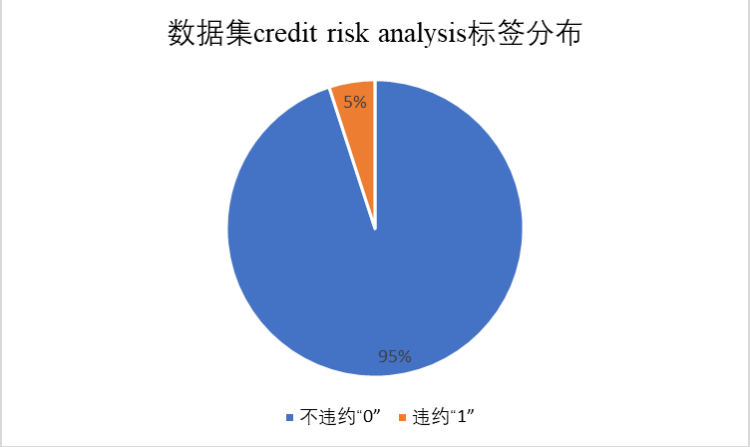


图 5-1 数据集中的违约样本、正常样本数量比例

Figure 5-1. Ratio of default samples to normal samples in the data set

剩余的数据集中的样本特征多达 45 个，其中不乏与预测结果不相关的特征。为了提升算法的效率，通过 python 构建随机森林模型，采用 feature_importances_对原始数据集中的特征进行重要性排序处理，确定一个阈值，按特征的重要性选择出与预测是否“违约”关联性高的特征作为最终的数据集特征。最终选出表 5-1 所示的八个主要特征的相关数据进行实验。

表 5-1 数据集 credit risk analysis 重要的特征信息

Table 5-1. Data sets Credit Risk Analysis Important characteristic information

特征	特征信息
loan_amnt	可使用额度
int_rate	借款利率
grade	借款等级
annual_inc	借款人年收入
purpose	借款用途
installments	选择的每月还款金额
term	贷款还清期限
default_ind	样本标签，0：不违约，1：违约

对于数据被定义为离散型的 grade、purpose、term 特征属性转化为数值型数据，如：借款用途的几种记录：债务合并、信用卡或其他，将以上 3 种情况分别定义为数字 0、1、2。之后对这些除标签以外的所用数据进行归一化处理。

5.3 FSVM-BS 算法模型的实验设计及结果分析

小微银行或贷款机构，无法获得大量的可靠的真实数据集对申办贷款者进行信用风险预测。为了检验本文提出的 FSVM-BS 算法在少量的、不平衡数据集下的银行信用风险评估效果，考虑到类似符合不违约特征而标签为违约的数据对预测结果的影响，本节设计了如下的真实数据的例子进行论证与分析。

在实验开始前，我们把处理过的数据集分为三部分，一部分作为训练集用来构建模型的训练，一部分作为测试构建模型的性能的测试集，还有一部分作为参数选择的验证集。三者的数量比例设置比 6：2：2。这样一来，选择出来的参数比人工自主选择出来的数更合理、客观，另外也避免了测试集“信息泄露”带来的困扰。

5.3.1 参数的设置

基于此数据集构建模型需要确定模型中的参数值，关于模型中相关参数的选取，做出以下说明：

（1）以标准支持向量机为基础的惩罚变量C_i：$C_{不违约}$设值为 10，$C_{违约}$设值为 1000；

（2）K 近邻规则中的最近邻数 k 的选择。k 值定义为奇数，测试区间设置在$[5,7,9,11,13,15]$ 。这样设置的原因有二：一是 k 值为偶数时，可能会出现近邻的 k 个样本中正负样本出现概率相等的情况，影响分类器的判断；二是 k 值较大时，距离训练样本较远的样本也会对结果产生干扰，且易受样本均衡影响程度大，使预测结果出现误差。此外，要保证尽可能大的区间统计到的概率准确性，计算距离的成本也会较高。

（3）另外，K 近邻规则中所用距离公式为欧氏距离不变，SVM、BSVM、FSVM 和 FSVM-BS 模型所用的核函数为高斯核函数不变。

以上各参数确定好以后，按照算法流程进行训练、评估、测试流程。

5.3.2 模型实验结果与分析

在银行信用风险预测的不平衡数据集中，通常把无风险人群判定为有风险比有风险人群判定为无风险的接受成本更低，因此，考虑到减少银行信用风险等级，应当对有风险人数的预测精确度要求更高。在同样的测试集测试的前提下，选用标准支持向量机 SVM、有偏支持向量机 BSVM、模糊支持向量机 FSVM 和认知偏差的模糊支持向量机 FSVM-BS 四种算法模型建模。其中，标准

支持向量机作为预测问题的常用方法，易于使用和解释。有偏支持向量机模型较标准支持向量机算法在解决不平衡分类问题上的泛化性较优。选用模糊支持向量机，意在比较一般的模糊隶属度计算方法和基于人的认知特性的模糊隶属度计算方法两种方式的差异。

为避免正负样本出现概率相等影响判断，通过验证集对定义区间[5,7,9,11,13,15]内的不同 k 值进行测试。为避免数据的偶然性，每个 k 值试验 20 次，并记录每次对应的正、负样本数量和样本权重。图 5-2 至图 5-7 表示不同 k 值下的正、负样本数量分布及其样本权重 LS 随之产生的变化：

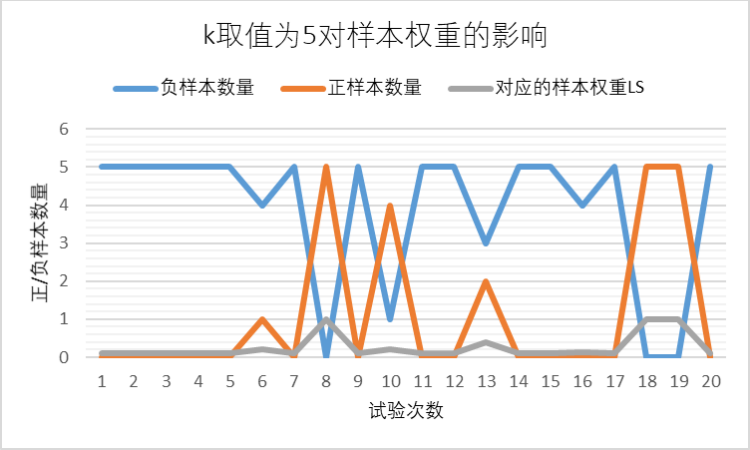


图 5-2 k 取值为 5 对样本权重的影响

Figure 5-2. The influence of k is 5 on the sample weight

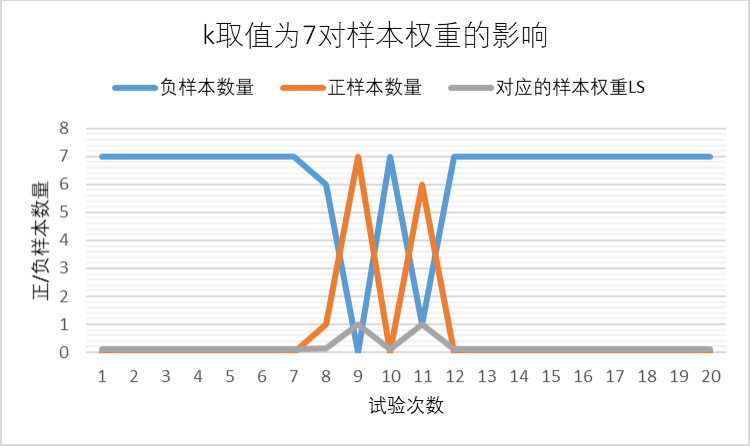


图 5-3 k 取值为 7 对样本权重的影响

Figure 5-3. The influence of k is 7 on the sample weight

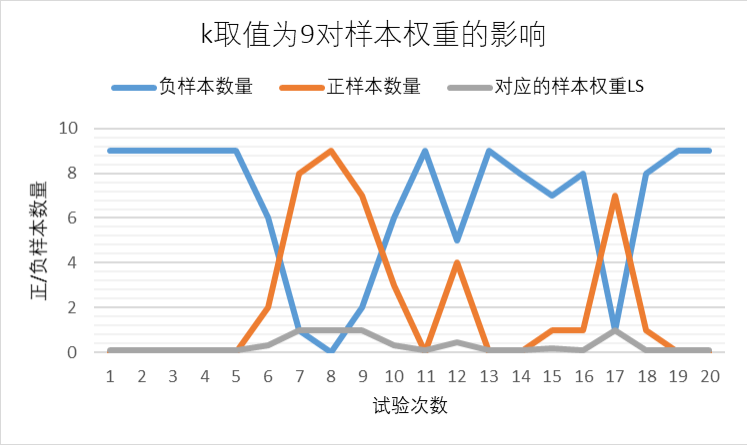


图 5-4 *k* 取值为 9 对样本权重的影响

Figure 5-4. The influence of *k* is 9 on the sample weight

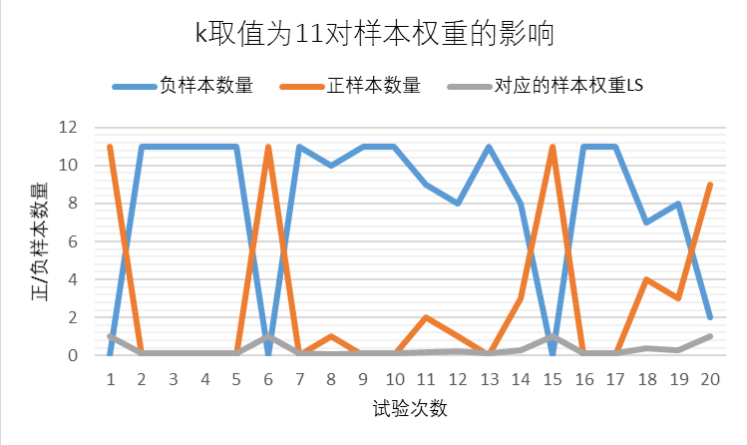


图 5-5 *k* 取值为 11 对样本权重的影响

Figure 5-5. The influence of *k* is 11 on the sample weight

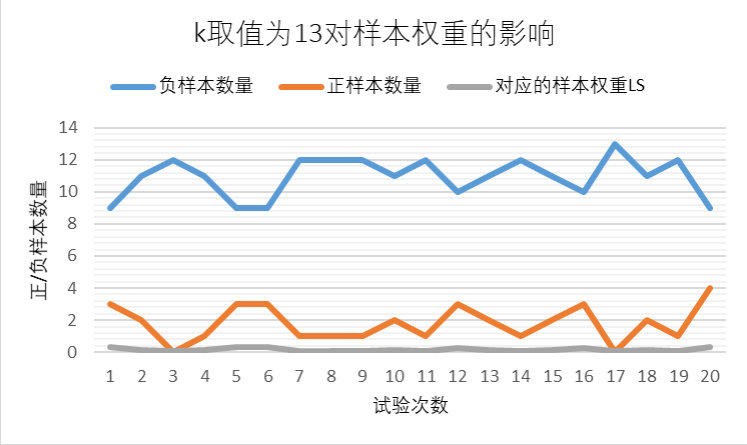


图 5-6 *k* 取值为 13 对样本权重的影响

Figure 5-6. The influence of *k* is 13 on the sample weight

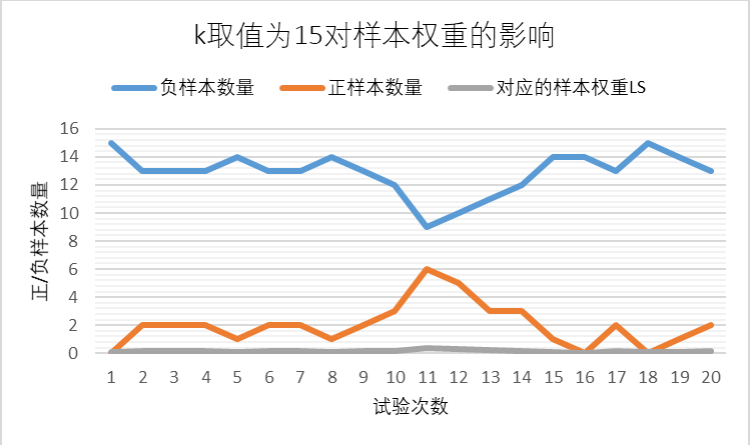


图 5-7 k 取值为 15 对样本权重的影响

Figure 5-7. The influence of k is 15 on the sample weight

由图可知，样本权重 LS 与正样本数量成正比例关系，与负样本数量成反比关系，区间内的 k 值均符合上述标准。当 k 值为 5、7 时，最近邻样本取样结果中全为负样本的次数约为总次数的一半，远多于全为正样本的情况，取样结果倾向于样本多的负样本一类，因此选取 5 或 7 作为 k 的具体数值均不恰当。当 k 值分别为 9、11 时，在各自的 20 次试验中，最近邻样本中全为正样本、全为负样本、正样本多于负样本或负样本多于正样本这四种情况均被测试到，此时的样本权重 LS 可以根据不同的正、负样本分布情况依据流程图 4-5 进行调整。当 k 值为 13、15 时，最近邻样本取样的 20 次结果中，整体上负样本数量均明显多于正样本，导致样本权重较低且没有明显变化，这样的结果虽然符合方法的预期，但没有覆盖正样本占比高于负样本占比这一情况，使取样涵盖情况不全面，此时的 k 值不是合适的取值。

通过上段内容的分析，认为k取值为9或11相较于其他四种取值更为合适。现进一步分析，确定 k 的取值。仔细观察图 5-4 和图 5-5 不难发现，图 5-4 中 k 为 9 时，第 6 次、第 14 次中的正、负样本数量总和等于 8；图 5-5 中 k 为 11 时，第 12 次正、负样本数量总和等于 9，第 18、19 次和第 20 次试验结果中正、负样本数量总和等于 10，这四次的统计结果均不等于 11。在其它的 k 取值时，也有不同程度的错误统计情况。模型输出值和真实值之间的差异，究其原因，是数据欠拟合的表现，这种 k 取值情况下未能充分利用数据中的有效信息。图 5-8 表示定义区间内不同 k 取值下的 20 次试验统计结果中的错误统计及错误出现的概率，正负样本数量统计错误次数表示正负样本数量加起来的总和不等于总数 k 的次数，错误的概率表示 20 次试验中正负样本数量总和不等于总数 k 的占比情况。由于统计结果只有 20 次，错误统计次数对出现错误的概率影响较为明显，错误率为 0、5.00%和 10.00%对应的错误次数分别为 0、1 和 2，与之相应的 k 取值分别为 5、7 和 9。

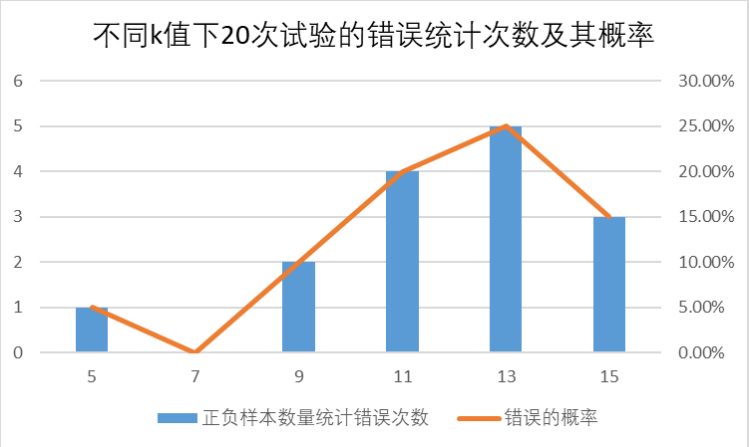


图 5-8 不同 *k* 值下 20 次试验的错误统计次数及其概率

Figure 5-8. Error statistic times and probability of 20 tests under different K values

考虑到最近邻数 k 下正、负样本数量分布情况涵盖的全面性，k 取值为 9、11 时较为合适；考虑到不同 k 值下的错误统计情况，为保证实验结果的准确性，k 取值为 5、7、9 时较为合适。涵盖情况、准确性两方面综合分析，最近邻数 k 选为 9 最为合适。

为了更直观、准确地表示不同模型的性能好坏，利用 ROC 和 AUC 的值作为评价指标。已知 ROC 曲线下的面积 AUC 越大，此时的证明模型针对不平衡数据集的预测效果越好。通过实验，四种模型的 ROC 曲线及其线下面积结果如下图所示：

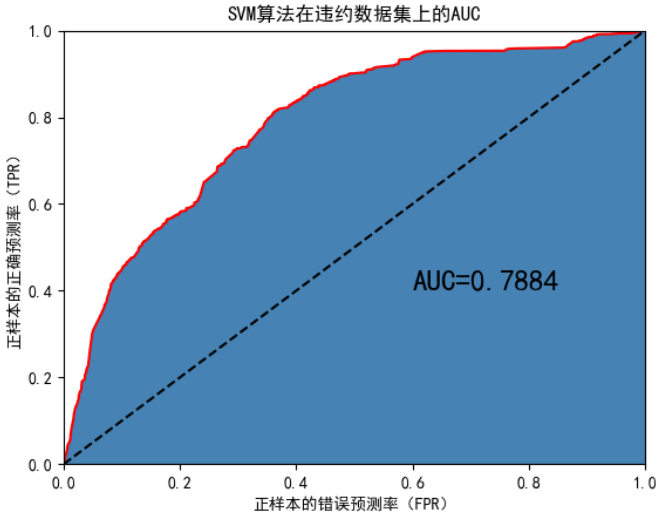


图 5-9 SVM 算法在违约数据集上的 AUC

Figure 5-9. AUC of SVM algorithm on default data set

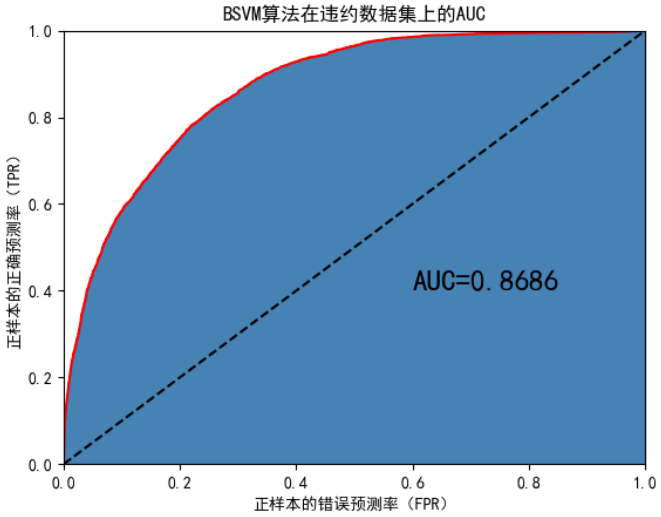


图 5-10 BSVM 算法在违约数据集上的 AUC

Figure 5-10. AUC of BSVM algorithm on default data set

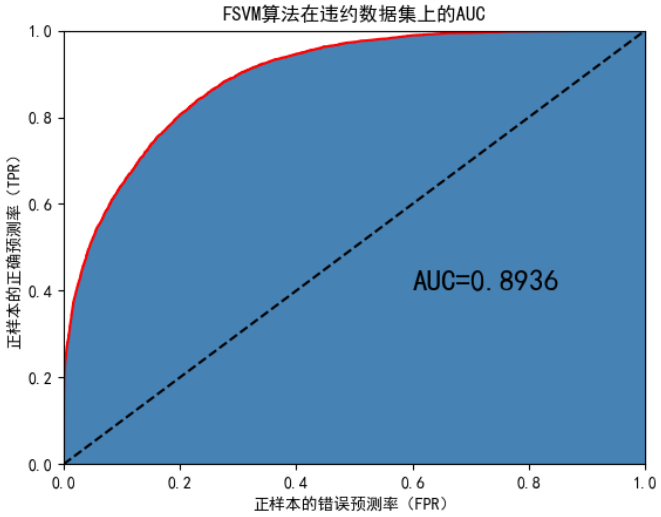


图 5-11 FSVM 算法在违约数据集上的 AUC

Figure 5-11. AUC of FSVM algorithm on default data set

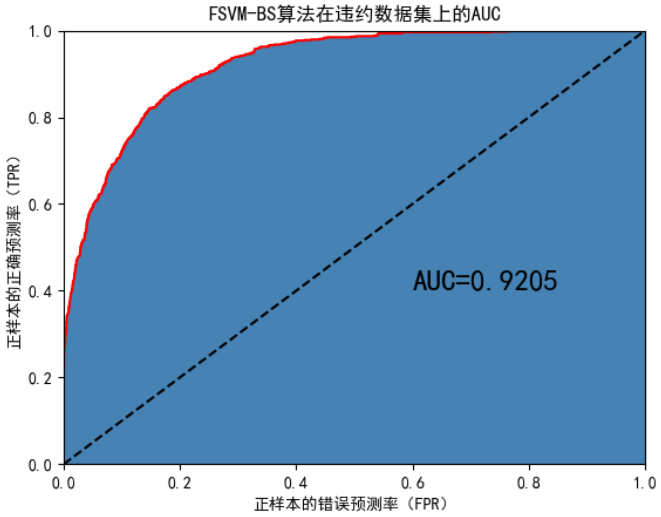


图 5-12 FSVM-BS 算法在违约数据集上的 AUC

Figure 5-12. AUC of FSVM-BS algorithm on default data set

由以上运行结果可知，SVM 模型在银行信用风险预测任务中的 AUC 值为 0.7884，模型结果与其他三个相比性能最差；BSVM 算法在 SVM 算法的基础上，性能稍有提升，AUC 值达到 0.8686；FSVM 算法和 FSVM-BS 算法的 ROC 线下面积 AUC 分别为 0.8936 和 0.9205，FSVM-BS 算法的精确度最高,把 AUC 转换成百分数制来看比其他三种模型提高了至少 2.6 个百分点。这也说明，FSVM-BS 算法正确预测无风险人群的同时把少数类的、有违约风险的样本错误预测的概率最低，即对有违约风险的人群正确预测的概率高于其被预测为无违约风险的人群，更高程度地帮助银行规避风险。

在以上基础下，SVM、BSVM、FSVM 和 FSVM-BS 四种模型在训练集和测试集上的表现能力以及花费时间也不同，如柱状图 5-13 所示。从某一单一的性能来看，SVM 算法模型在训练集的准确率最高为 1，其次是 BSVM、FSVM，FSVM-BS 模型最低；从测试集的准确率角度进行对比，FSVM-BS 模型的准确率最高，其次是 FSVM、BSVM 和 SVM。通过图 5-13 不难发现，FSVM-BS 模型在训练集和测试集的准确率差距最小，即 FSVM-BS 模型在不论是银行信用风险预测数据集的训练还是测试过程中，表现较其他三种模型更为稳定。四种模型总花费时间最少的是 SVM 模型，其次是 BSVM 和 FSVM-BS，FSVM 模型总花费时间最多约 16s，远超过其他三种模型所花费的时间。

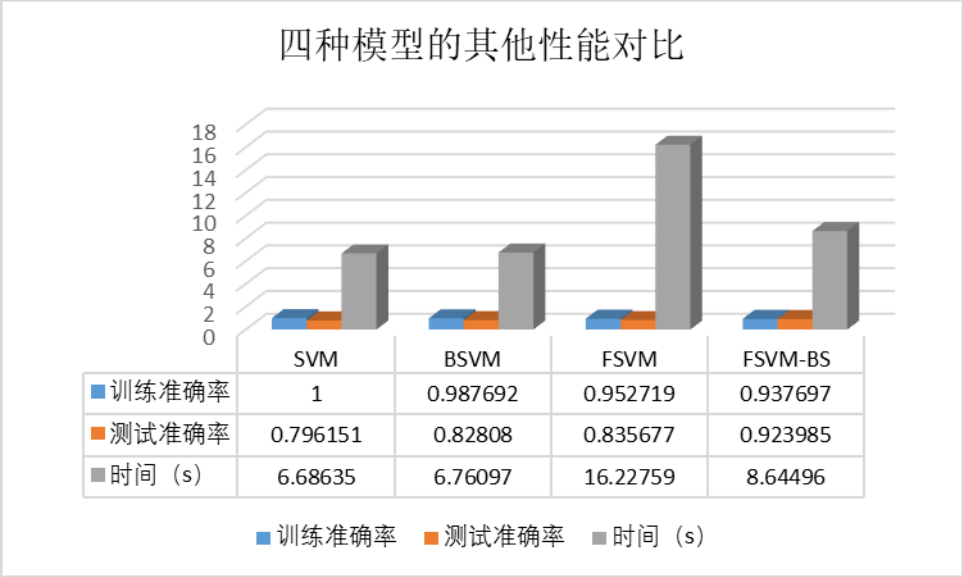


图 5-13 四种模型的其他性能对比

Figure 5-13. Other performance comparisons of the four models

综合 ROC 曲线下面积 AUC、模型在训练和测试准确率的稳定性以及花费时间三个方面的比较和分析，发现与其他 SVM 模型相比，FSVM-BS 算法在该银行信用风险预测相关的数据集中的分类精度提升得更明显，稳定性更强，花费的时间相对较少，整体性能更好。

5.4 本章小结

在本章中，介绍了银行信用风险预测的发展和研究现状和提高风险预测的必要性。在实际银行信用风险预测的应用背景下采用真实的数据集，对数据集中有缺失的特征进行过滤，采用随即森林对数据集中的特征进行重要性排序，选取出重要性最高的特征进行训练学习，经过训练集、验证集、测试集一系列流程后，本文提出的方法 FSVM-BS 算法在银行信用风险预测应用中的 AUC 结果明显高于其他三个模型，另外从稳定性和花费时间进行对比，进一步证明了本方法在实际应用中的有效性。

第六章 总结与展望

6.1 总结

随着机器学习的不断发展，机器智能越来越贴近于人们的日常生活、服务于人们的日常生活。可围绕人们生活场景中的真实数据特征不满足于传统机器学习方法要求的大量的、平衡的特点，导致数据“理想”状态下训练出来的学习模型容易忽略数据集中的样本较少的一类，分类边界偏向数据集中样本较多的类别。类如医疗诊断中诊断出某种疾病的病患样本数量远小于非患病样本数量，而这些少量的病患样本所包含的样本信息恰好是我们分类关注的重点。传统的机器学习方法遇到类似医疗诊断、风险预测、故障检测等难以大量获取平衡样本的真实情境，不能切合实际地应用到现实生活中解决问题。如何在面对不平衡数据集时抑制多类样本影响确定分类边界的“强势”地位，追求样本整体分类精度的同时更保证了少类样本的准确率，是提高现有机器学习算法泛化性能和鲁棒性的关键。

本文对不平衡数据学习问题和基于支持向量机方法解决不平衡问题的基本理论模型进行探讨，分析不平衡数据的类间不平衡中相对不平衡、绝对不平衡两种因素对标准支持向量机模型性能的影响，介绍了基于标准支持向量机解决不平衡分类问题存在的不足进一步优化得到的有偏支持向量机（BSVM）算法和模糊支持向量机（FSVM）算法的理论基础和模型。同时，结合具体研究举例说明人的生理认知特性和心理认知特性在机器学习方面的有效应用。本文围绕以上内容展开研究并提出了一种融合人的认知模型的模糊支持向量机方法（FSVM-BS）来提升不平衡分类问题的精度，并通过具体的数值实验验证本文方法的可行性和实际应用的有效性，文章各章节的研究内容相互紧密连接，主要贡献有以下：

（1）分析数据不平衡对支持向量机算法的影响，得出类间绝对不平衡对算法性能影响更明显的结论。根据人面对不平衡的、少量的信息快速、准确地学习特点总结出人的基于生理、心理的认知特性，以其与机器学习方法结合的有效案例。

（2）基于人的心理认知特性，借助 K 近邻算法，设计了一种融合人的认知模型的模糊支持向量机方法，对分类边界附近的正、负类样本依据其模糊隶属度给予不同的学习权重，强化少数类的正样本在训练过程中的作用，来提升二分类问题中的不平衡分类问题。

（3）通过不同程度的不平衡率的数据集的数值实验和基于银行信用风险预测的实际背景下的不平衡数据集实验，证明了融合人的认知模型的模糊支持向量机算法在提升不平衡分类问题精度的有效性。本文提出的融合人的认知模型的模糊支持向量机算法理论上适用于存在因果关系的不平衡二分类场景，不局限于银行信用风险预测任务中。

整体上来说，随着认知科学、脑神经科学和机器学习的多学科融合发展，融合人的认知模型的模糊支持向量机算法为提升不平衡分类问题精度提供了新的思路。将人的智能和机器智能相融合，利用人的智能促进机器智能的发展，机器智能再服务于人，形成人与机器智能的良好循环发展。

6.2 展望

融合人的心理认知特性到支持向量机算法中解决不平衡分类问题，涉及到认知科学、机器学习等因素非常多，受到时间、研究条件的限制，本文的很多研究工作有待进一步扩展和提高。日后将在以下几个方面更深层次地进行思考与研究：

（1）针对不平衡分类问题，分析类间不平衡情况对支持向量机算法的影响程度时，应该探究不同的不平衡率的变化下，与之对应的分类精度的具体数值变化，找到不平衡率与算法性能之间的定量数据关系，并根据这种定量关系修正新提出算法的参数。

（2）本文中所做的工作都是基于二分类的不平衡问题研究。在解决多分类的不平衡问题时，样本中各类别数据的不平衡对多分类分类器同样带来干扰，日后，应尝试把自己的方法扩展到多分类的不平衡问题研究上，使其满足多分类任务的分类需求。

（3）了解更多的关于认知科学和脑神经科学相关的研究进展，寻找人或动物的其他认知特性与人工智能结合的可能性。针对不同的应用场景，尝试不同的认知特性与现有机器学习方法的有效结合解决现实问题，也将是具有研究价值的方向。

最后，由于本人学识有限，论文难免存在不足之处，恳请各位专家、同仁批评指正，本人在此表示感谢。

参考文献

[1] Fan Q, Wang Z, Li D D, et al. Entropy-based Fuzzy Support Vector Machine for Imbalanced Datasets[J]. Knowledge-Based Systems, 2016, 115(JAN.1): 87-99.

[2] Chawla N V, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets[J]. Acm Sigkdd Explorations Newsletter, 2004, 6(1): 1-6.

[3] Cardie C., Howe N.. Improving minority class predication using casespecific feature weights. In Proceedings of the fourteenth International Conference on Machine Learning, 1997. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. , 1997: 57-65.

[4] Ezawa K., Singh M., Norton S.W.. Learning goal oriented Bayesian networks for telecommunications risk management. In Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 1996. Morgan Kaufmann, 1996: 139-147.

[5] Merlin P M, Farber D J. Department of Information and Computer Science, University of California[J]. Computers IEEE Transactions on, 1975, c-24: 96-98.

[6] Kubat M, Holte Robert C, Matwin S. Machine Learning for the Detection of Oil Spills in Satellite Radar Images[J]. Machine Learning, 1998, 30(2-3): 195-215.

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

[8] Kubat M, Matwin S. Addressing the Course of Imbalanced Training Sets: One-sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann, 1997: 179-186.

[9] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.

[10] Hui H, Wang W Y, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[C]. //Huang De-Shuang, Zhang Xiao-Ping, Huang Guang-Bin. Advances in Intelligent Computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 878-887.

[11] 谢子鹏, 包崇明, 周丽华, 等. 类不平衡数据的 EM 聚类过采样算法[J]. 计算机科学与探索, 2021, 15: 1-14.

[12] 陈刚, 郭晓梅. 基于时间序列模型的非平衡数据的过采样算法[J]. 信息与控制, 2021, 49: 1-10.

[13] Douzas G, Rauch R, Bacao F. G-SOMO: An oversampling approach based on self-organized maps and geometric SMOTE[J]. Expert Systems with Applications, 2021, 183(2): 115230.

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

[15] Nathalie J. Concept-Learning in the Presence of Between-Class and Within-Class Imbalances. In Advances in Artificial Intelligence, Proceedings of the 14th Biennial Confer-ence of the Canadian Society on Computational Studies of Intelligence, Berlin, Heidelberg, 2001. Eleni Stroulia, Stan Matwin: Springer Berlin Heidelberg, 2001: 67-77.

[16] 孟东霞, 李玉鑑. 基于特征边界欠采样的不平衡数据处理方法[J]. 统计与决策, 2021, 37(11): 30-33.

[17] 崔彩霞,曹付元,梁吉业.基于密度峰值聚类的自适应欠采样方法[J].模式识别与人工智能, 2020, 33(09): 811-819.

[18] 何云斌, 冷欣, 万静. 不平衡数据加权边界点集成欠采样方法[J]. 西安电子科技大学学报, 48(4): 9.

[19] Shahabadi M, Tabrizchi H, Rafsanjani M K, et al. A combination of clustering-based under-sampling with ensemble methods for solving imbalanced class problem in intelligent systems [J]. Technological Forecasting and Social Change, 2021, 169(1): 120796.

[20] Zhu Y W et al. EHSO: Evolutionary Hybrid Sampling in overlapping scenarios for imbalanced learning[J]. Neurocomputing, 2020, 417: 333-346.

[21] Feng S, Zhao C H, Fu P. A cluster-based hybrid sampling approach for imbalanced data classification[J]. The Review of scientific instruments, 2020, 91(5): 055101.

[22] Li D D, et al. Entropy‐based hybrid sampling ensemble learning for imbalanced data[J]. International Journal of Intelligent Systems, 2021, 36(7): 3039-3067.

[23] 于艳丽, 江开忠, 盛静文. 不平衡数据中基于异类 k 距离的边界混合采样[J]. 计算机应用与软件, 2021, 38(02): 299-304+310.

[24] Zadrozny B., Elkan C.. Learning and making decisions when costs and probabilities are both unknown. In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, San Francisco, California, August 2001. New York, NY, USA: Association for Computing Machinery, 2001: 204-213.

[25] Yi L, Lee Y, Wahba G. Support Vector Machines for Classification in Nonstandard Situations [J]. Machine Learning, 2002, 46(1-3): 191-202.

[26] Wu G., Chang E.Y.. Class-boundary alignment for imbalanced dataset learning. In ICML 2003 Workshop on Learning from Imbalanced Data Sets, Washington, DC, August 2003. 2003: 49-56.

[27] Liu B, Ma Y, Wong C K. Improving an association rule based classifier. In Principles of Data Mining and Knowledge Discovery, the 4th European Conference on Principles of Data Mining and Knowledge Discovery, Lyon, France, September 2000. Djamel A Zighed, Jan Komorow-ski, Berlin, Heidelberg: Springer Berlin Heidelberg,2000: 504-509.

[28] Yao C F, et al. A Class-Incremental Learning Method Based on One Class Support Vector Machine[J]. Journal of Physics: Conference Series, 2019, 1267: 70-76.

[29] Kumar B, et al. A fast learning algorithm for One-Class Slab Support Vector Machines[J]. Knowledge-Based Systems, 2021, 228.

[30] Gao L, Zhang L, Liu C, et al. Handling imbalanced medical image data: A deep-learning-based one-class classification approach[J]. Artificial Intelligence In Medicine, 2020, 108:101935.

[31] Xie H X, Liu B, Xiao Y S. Transfer learning-based one-class dictionary learning for recommendation data stream[J]. Information Sciences, 2021, 547 : 526-538.

[32] Liu B, Xie H, Xiao Y. Multi-task analysis discriminative dictionary learning for one-class learning[J]. Knowledge-Based Systems, 2021, 227(99): 107195.

[33] Sun Y, Kamel M S, Wong A, et al. Cost-sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007, 40(12): 3358-3378.

[34] Veropoulos K., Campbell C., Cristianini N. Controlling the Sensitivity of Support Vector Machines. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden (IJCAI99), 1999.1999: 55-60.

[35] Xu J, Cao Y B, Li H, et al. Cost-Sensitive learning of SVM for ranking. In Machine Learning: ECML 2006, the 17th European conference on Machine Learning (ECML'06). Tobias Scheffer, Myra Berlin Spiliopoulo, 2006. Heidelberg: Springer-Verlag, 2006: 833-840.

[36] Fong R C, Scheirer W J, DD Cox. Using human brain activity to guide machine learning[J]. Scientific Reports, 2018, 8(1):5397.
[37] 张春霞,张讲社.选择性集成学习算法综述[J].计算机学报,2011,34(08):1399-1410.
[38] Bhagat R. C., Patil S. S.. Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest. In 2015 IEEE International Advance Computing Conference (IACC). 2015: 403-408.
[39] 陈圣灵,沈思淇,李东升.基于样本权重更新的不平衡数据集成学习方法[J].计算机科学, 2018, 45(07): 31-37.
[40] Fan X, Lin H, Diao Y, et al. An Integrated Biomedical Event Trigger Identification Approach with a Neural Network and Weighted Extreme Learning Machine[J]. IEEE Access, 2019, PP(99): 1-1.
[41] 杨昊天, 顾乾晖, 王嘉璐, 等.基于混合采样和集成学习的软件缺陷预测[J]. 网络安全技术与应用, 2021(05): 59-60.
[42] Mahadevan A, Arock M. A class imbalance-aware review rating prediction using hybrid sampling and ensemble learning[J]. Multimedia Tools and Applications, 2021, 80(5): 6911-6938.
[43] Choudhary R, Shukla S. A clustering based ensemble of weighted kernelized extreme learning machine for class imbalance learning[J]. Expert Systems with Applications, 2021, 164 : 114041.
[44] Vladimir N. V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995: 1-188.
[45] Williams P, Li S, Feng J, et al. A Geometrical Method to Improve Performance of the Support Vector Machine[J]. IEEE Transactions on Neural Networks, 2007, 18(3): 942-7.
[46] Veropoulos K., Campbell C. and Cristianini N. Controlling the Sensitivity of Support Vector Machines. In Proceedings of the International Joint Conference on AI, Stockholm, Sweden. 1999: 55-60.
[47] 关天民,赵德竹,梅钊.基于模糊神经网络的脊柱侧弯矫形器设计模型[J].大连交通大学学报, 2021, 42(04): 51-56.
[48] 郭志军,刘帅.基于卷积神经网络的数字图像模糊增强算法[J].吉林大学学报(工学版),2021,65:1-6.
[49] Tatsuji T, Masahiro N, Shuji S. Cognitive Symmetry: Illogical but Rational Biases[J]. Symmetry Culture & Science, 2010, 21(1): 1-3.
[50] Richhariya B, Tanveer M. A Fuzzy Universum Support Vector Machine Based on Information Entropy[M]. Singapore: Springer Singapore. 2019: 569-582.
[51] Metz C E. Basic principles of ROC analysis[J]. Seminars in Nuclear Medicine, 1978, 8(4): 283-298.
[52] Bradley P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997, 30(7): 1145-1159.
[53] Hidetaka T, Hiroshi S, Tomohiro S. Implementation of Human Cognitive Bias on Neural Network and Its Application to Breast Cancer Diagnosis[J]. The Society of Instrument and Control Engineers, 2019, 12(2): 56-64.
[54] Taniguchi H, Sato H, Shirakawa T. A machine learning model with human cognitive biases capable of learning from small and biased datasets[J]. Scientific Reports, 2018, 8(1).

[55] Alcalá-Fdez J., Sánchez L., García S, et al. KEEL: a software tool to assess evolutionary algorithms for data mining problems[J]. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 2008, 13: 307-318.

[56] 李仁真.国际金融法[M].武汉: 武汉大学出版社,2011.

[57] Sihem K, Fatma B S, Younes B. Credit risk assessment for unbalanc-ed datasets based on data mining, artificial neural network and support vector machines[J]. Journal of Modelling in Management, 2018, 13(4): 932-951.

[58] Zhang J. Investment risk model based on intelligent fuzzy neural network and VaR[J]. Journal of Computational and Applied Mathematics, 2020, 371: 112707.

[59] Ohlson. Financial Ratios and the Probabilitistic Prediction of Bankruptcy[J]. Accounting Research,1980,18(1):109-131.

[60] Yan Chan, Wang Shufu, Li Xiang. Credit Risk Assessment of Small, Medium and Micro Enterprises Based on Logistic Regression Model[J]. International Journal of Higher Education Teaching Theory, 2020, 1(1).

[61] Lu Jiexin, Tong Yongzhen. Research on credit risk of commercial banks based on multiple logistic model[J]. Academic Journal of Business & Management, 2021, 3(6): 83-87.

[62] 李淑锦,嵇晓佳.基于 Lasso-Logistic 模型的个人信用风险评估——来自微贷网的数据分析[J].杭州电子科技大学学报(社会科学版),2020,16(06): 8-15.

[63] 杨月, 袁宇. Research on Credit Risk Early Warning Based on RBF Neural Network-Taking Internet Financial Trading Platform as an Example[J]. 建模与仿真, 2021, 10(02) : 257-267.

[64] Mahbobi M, Kimiagari S, Vasudevan M. Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks[J]. Annals of Operations Research, 2021:1-29.

[65] Pan S, Zhou S. Evaluation Research of Credit Risk on P2P Lending based on Random Forest and Visual Graph Model[J]. Journal of Visual Communication and Image Representation, 2019: 102680.

[66] Arora N, Kaur P D. A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment[J]. Applied Soft Computing, 2019, 86:105936.

[67] 陈倩, 贺兴时, 杨新社. 基于 RF 的 Elastic Net-Logistic 个人信用违约风险评估[J]. 西安工程大学学报, 2021, 35(03):116-122.

[68] Pawiak P, Abdar M, Acharya U R. Application of New Deep Genetic Cascade Ensemble of SVM Classifiers to Predict the Australian Credit Scoring[J]. Applied Soft Computing, 2019, 84(2019): 105740.

[69] Luo Jian, Yan Xin, Tian Ye. Unsupervised quadratic surface support vector machine with application to credit risk assessment[J]. European Journal of Operational Research, 2020, 280(3): 1008-1017.

[70] Lya C, Xiao Y B, Xz A, et al. A novel dual-weighted fuzzy proximal support vector machine with application to credit risk analysis[J]. International Review of Financial Analysis, 2020, 71.

致　谢

2019 年至 2021 年，两年半的时光稍纵即逝。仅仅的两年半，时间在走、世界在变。于国家而言，2019 年末、2020 年初新旧交替之际，武汉发现“不明原因肺炎”之后全国大范围内爆发“新冠病毒”疫情。国家整体调度各方力量火力组建“火神山医院”“方舱医院”，医护人员义无反顾抢救病人，警察、社区等基层人员为民众做好各方生活保障，民众自觉足不出户、居家自我隔离。于我的家乡而言，2021 年 7 月经历罕见暴雨，一时间城市内涝严重、村庄里家园和良田被洪水冲泡，国家人民子弟兵、全国各省份志愿者奔赴现场，对被困人员积极救援；多方救援、生活物资蜂拥而至。于我而言，来到异乡踏上了新的求学之路。这一路上，我能按原计划顺利完成学业，首先要感谢祖国面对突发疫情、灾情快速而精准的反应，让我对灾难不再恐惧和焦虑，取而代之的是对未来充满希望和期待。感谢那些舍己为人、舍小家为大家的“逆行者”，恰因他们的流泪流血流汗、义无反顾和无私奉献，我才得以在安全、稳定、正常的生活轨迹上按计划做自己的事情，如期毕业。

其次，我要郑重感谢我的导师赵云波教授，感谢他在我科研道路上的循循善诱、谆谆教导，为我的科研道路指明了方向。在与老师相处的时光里，他严谨治学、好学深思的态度深深影响了我，使我在学习和生活中不再浮于表面，更喜欢思考隐于现象或事件背后的根本。

感谢实验室里的许德衡、蒋传鹏、梁启鹏、唐敏等师兄和朱巧慧、王岭人师姐，感谢你们平时对我学习上的指导和日常生活上的帮助。感谢同课题组的吴芳同学，在我入校之初尽显“地主之谊”，帮助我尽快适应、融入新的校园生活和实验室氛围。感谢我的室友花婷婷、袁李婷两位同学在作息时间、生活习惯方面对我的包容，学习生活之余一起的集体活动为我日常紧张的学习生活增添了很多色彩。感谢同实验室的郝小梅、闫文晓、卢子轶、卢帅领同学，感谢你们在课题研究上给予我的帮助。

感谢我的男朋友苏振岭对我各方面的关心、理解和包容。两年半的异地学习、科研生活，我们相互督促、鼓励，互相因为彼此愿意付出努力成为更优秀的自己。感谢我的多年好友赵荟荟，为了理想的实验结果而泡在实验室的执着精神一直激励着我，希望她接下来的读博之路，多出成果少熬夜。

感谢我的父母多年来含辛茹苦的付出，养我长大育我成人。虽然没有什么成就，但感谢父母一直以我为傲。父母对我的期望，是我一直不懈的动力。也感谢小晨阳在家里作为“开心果”的存在，为我们家庭生活增添了很多欢乐。

最后，由衷地感谢评阅专家、老师和学者，在百忙之中审阅我的论文！

作者简介

1 作者简历

1996 年 8 月出生于河南省新乡市；

2015 年 9 月--2019 年 6 月，黄淮学院信息工程学院电子科学与技术专业学习，获得工学学士学位；

2019 年 9 月--2022 年 1 月，浙江工业大学信息工程学院控制工程专业学习，攻读专业型工程硕士学位。

2 参与的科研项目

[1] 人机系统中人与机器的自主性边界及其切换策略研究. 军科委国防科技创新特区项目（合同号：18-163-11-ZT-004-009-01）.

[2] 复杂环境下非完全信息博弈决策的智能基础模型研究理论研究. 科技创新 2030-“新一代人工智能”重大专项课题(项目编号：2018AAA0100801).

3 学术论文

[1] Min Tang, Fang Wu, Li-Li Zhao, Qi-Peng Liang, Jian-Wu Lin, Yun-Bo Zhao. Detection of Distracted Driving Based on MultiGranularity and Middle-Level Features[C]// 2020 Chinese Automation Congress (CAC). 2020.

4 发明专利

[1] 赵云波, 唐敏, 赵丽丽. 一种人机系统中人的状态的识别方法[P]. 浙江省：CN110598616 A, 2019-12-20.

[2] 赵云波, 唐敏, 赵丽丽, 吴芳. 一种程序员的疲劳程度的检测方法[P]. 浙江省：CN110786869A, 2020-02-14.

[3] 赵云波, 唐敏, 吴芳, 赵丽丽. 一种基于图像的驾驶员注意力检测方法[P]. 浙江省：CN11155 3190A, 2020-08-18.

[4] 赵云波, 唐敏, 花婷婷, 赵丽丽. 一种基于多粒度特征与中层特征的分心驾驶检测方法[P]. 浙江省：CN111695535A, 2020-09-22.

[5] 赵云波, 赵丽丽, 花婷婷, 苏振岭. 一种基于人的生理认知特点对苹果精细化分拣的分类方法[P]. 浙江省：CN112668645A, 2021-04-16.

[6] 赵云波, 花婷婷, 赵丽丽, 崔奇. 一种基于多算法集成的分歧介入珍珠分拣方法[P]. 浙江省：CN1 13159150A, 2021-07-23.
[7] 赵云波, 赵丽丽, 花婷婷, 苏振岭. 一种基于人的心理认知模型的青菜病害分类检测方法[P]. 浙江省：CN113283470A, 2021-08-20.

学位论文数据集

密 级*	中图分类号*	UDC*	论文资助
公开	TP181	681.5	
学位授予单位名称	学位授予单位代码	学位类型*	学位级别*
浙江工业大学	10337	工程硕士	全日制专业型硕士
论文题名*	融合人的认知模型的模糊支持向量机算法及其应用		
关键词*	不平衡分类问题，人的认知特性，支持向量机，银行信用风险预测		论文语种*
并列题名*	无		中文
作者姓名*	赵丽丽	学 号*	2111903287
培养单位名称*	培养单位代码*	培养单位地址	邮政编码
浙江工业大学 信息工程学院	10337	杭州市潮王路 18 号	310032
学科专业*	研究方向*	学 制*	学位授予年*
控制工程	人工智能	2.5 年	2022 年
论文提交日期*	2022 年 1 月		
导师姓名*	赵云波	职 称*	教授
评阅人	答辩委员会主席*	答辩委员会成员	
盲评	杨东勇	石崇源,洪榛,阮中远,施朝霞	
电子版论文提交格式：文本（√）图像（ ）视频（ ）音频（ ）多媒体（ ）其他（ ）			
电子版论文出版（发布）者	电子版论文出版（发布）地	版权声明	
论文总页数*	63 页		
注：共 33 项，其中带*为必填数据，为 22 项。			