密 级_____

桂林電子科技大學
GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY

硕 士 学 位 论 文

题目 价格预测中代价敏感的机器学习算法及优化

（英文） Cost-sensitive Machine Learning Algorithm and Optimization in Price Prediction

研 究 生 学 号: 1608202004

研 究 生 姓 名: 马超

指导教师姓名、职称: 刘振丙 研究员

申 请 学 位 门 类: 工学硕士

学 科、专 业: 控制科学与工程

论 文 答 辩 日 期: 2019 年 6 月 2 日

独创性（或创新性）声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得桂林电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：　　日期：

关于论文使用授权的说明

本人完全了解桂林电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属桂林电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为桂林电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密的论文在解密后遵守此规定）

本学位论文属于保密在_____年解密后适用本授权书。

本人签名：　　日期：

导师签名：　　日期：

摘要

P2P 汽车共享服务逐渐受到大众追捧，同时也吸引了来自学术界和企业界的关注。由于没有官方的定价标准，因此汽车共享服务中最重要的问题之一是预测汽车租赁价格。价格预测模型可以作为中间的定价模型，来避免过多的讨价还价过程。由于 P2P 的去中心化特点，预测出价格区间更为合适。

解决价格区间预测问题的传统方法是通过分类算法解决回归问题，但它的三个弊端限制了此方法直接应用于 P2P 共享汽车服务的价格预测，分别是离群值影响了 K 均值聚类结果的分布；区间的数量越多预测性能将会越差；深度森林忽略了不同错误分类情况的区别。为解决这些问题，本文做了如下研究：

1）我们分析了离群值对 K 均值聚类的影响，根据 K 均值的均匀作用和孤立森林的启发，改进了 K 均值离散化方法。改进的 K 均值算法降低了离群值对 K 均值聚类的影响，这使得离散化所产生的区间宽度更加一致。

2）对比传统算法和深度森林算法，我们发现深度森林的自适应深度可以使越差的分类性能获得更高的提升，这有助于缓解区间数量越多时分类性能越差的问题。这主要是因为，深度森林的自适应深度使得集成学习获得了足够的多样性，从而改进了集成学习的性能。这为深度森林开辟了新的应用领域；由于深度森林的基分类器可以是任意分类器，也使得深度森林的这一特点也为集成学习提供了借鉴。

3）我们将代价敏感学习引入到深度森林中，提出了代价敏感深度森林算法。相较于传统深度森林，代价敏感深度森林可以在相同准确度下，获得更低的代价，也就是说，其错误分类的区间更加接近真实区间。代价敏感深度森林不仅可以利用于通过分类算法解决回归问题中，它还可以根据不同的代价矩阵，来解决不同的代价敏感问题。

同时，整个模型也可以很容易地应用在其他领域的价格预测问题中。

关键词：P2P 汽车共享，价格预测，改进的 K 均值，集成深度学习，代价敏感深度森林

Abstract

Peer to peer (P2P) car-sharing services have gradually become popular, and attracted great attentions from researchers as well as entrepreneurs. One of the most important factors in the car-sharing service is to predict the price as there is no official price deciding rule. The price prediction model can be used as an intermediate pricing model, which avoids the complex bargaining process. Because of the decentralization feature of P2P, the prediction of price intervals is more suitable.

The traditional method for predicting price intervals is regression using classification algorithm, but three problems of it limit the direct application of this method on the price prediction of P2P car-sharing service, i.e. the outliers impact the distribution of the result of K-means clustering; the more number of intervals will lead worse prediction performance; the difference between different misclassification is ignored by deep forest. To solve these problems, the researches of this paper are as follow:

1) Inspired by uniform effect of K-means and isolation forest after we analyse the impact of outliers on K-means clustering, K-means discretization methods is modified. Modified K-means algorithm reduces the impact of outliers on K-means clustering, and get much uniform range of prediction intervals.

2) Comparing the traditional algorithms with deep forest, we find that the self-adapting depth of deep forest can make worse classification performance get better improvement, which is helpful to mitigate the problem, the more intervals, the worse classification performance. The reason is that the self-adapting depth of deep forest makes ensemble learning get enough diversity which improves the performance of ensemble learning. This opens up a new application area for deep forest. And this feature of deep forest can be referred by ensemble learning, because any classifier can be used as base classifier of deep forest.

3) Cost-sensitive learning is introduced into deep forest, and cost-sensitive deep forest algorithm is proposed. Cost-sensitive deep forest can get lower cost with same accuracy compared with traditional deep forest, which means the wrong prediction of it is closer to the real interval. Cost-sensitive deep forest can be used for not only regression using classification algorithm, but also different cost-sensitive problem by different cost matrix.

Meanwhile, the whole model can be used for price prediction problem in other domains easily.

Keywords: P2P Car-sharing, Price Prediction, Modified K-means, Ensemble Deep Learning, Cost-sensitive Deep Forest

目录

第一章 绪论

§1.1 研究背景与意义

近几年，中国“新四大发明”之一的共享单车使得共享经济蓬勃发展，也促进了众多不同领域的资源共享，例如房屋共享、拼车、共享单车、汽车共享[5]，共享经济正快速成长为我们生活中必不可少的一部分。为解决交通拥堵和环境污染，汽车共享最近正成为热点研究领域。相较于自行车，汽车成本更高，选择更加多样化，用户需求差别更大，虽然同样是为了满足人们的出行需求，但汽车共享却面临比共享单车更复杂的问题。目前针对汽车出行领域的共享主要集中在，以“滴滴”为代表的同时提供车辆和司机的“出租车”模式，以“PonyCar”为代表的随时随地借还的“共享单车”模式，以“神州租车”为代表的基于站点的“共享租车”模式，以“START”为代表的整合私人手中闲置车辆的“P2P”模式。

其中，“出租车”模式仅限于将乘客从出发地载到目的地这种方式，其本质上提供的是一种服务而不是车辆，虽然能满足大部分人的出行需求，但不能满足商务出行和家庭出行等短期租车的需求；“共享单车”模式是受共享单车启发，提供车辆给不同使用者共享使用的模式，因其没有司机参与，所以获得了更好的私密性和驾驶自由度，但其车辆普遍为低端车型，车辆寻找和停放麻烦，目前使用者主要是年轻人；“共享租车”模式有多种车型来满足不同出行模式的租车需求，但其车辆购置占据大量成本，租车也受网点限制，所以推广效率较低，目前主要在大城市中使用，同时也由于其高运营成本导致了高租车费用；“P2P”模式整合了私人车主手中闲置车辆，构建租车平台给有租车需求的租客促成双方交易，其车辆种类齐全，租车地点灵活，租车费用也比“共享租车”模式低三成左右，目前越来越成为不同出行需求租客的选择。除去“出租车”模式这种提供服务的模式，仅针对出租车辆的租车市场，“共享单车”模式集中在低端市场，“共享租车”模式选择有限，主要集中在商务出行，而“P2P”则可以以更低的价格满足不同需求、不同区域的个性化出行要求。

其中 P2P 汽车共享服务中，所共享的汽车是去中心化的，即汽车是个人所有，而不是像前两种汽车共享模式那样，汽车是由中间的经营者所有。P2P 共享经营者主要的角色是，提供在线交易平台来连接汽车所有者和潜在的汽车租赁者。与其他形式的汽车共享服务相比，P2P 共享汽车的用户通常有更加多样性的车辆选择[6]。然而，这也引发了其他问题：对于租车没有太多经验的新租赁者而言，这将会导致对在众多不同的汽车中，某个汽车租赁价格是否合理的怀疑；对于打算开始出租自己闲置车辆的车主而言，这也会面对如何定价才能获得最佳收益的问题。另外，由于共享物的属性

各不相同导致了买卖双方的心理价位不同，使得交易过程中容易出现价格争端，此时基于数据挖掘的价格预测系统可以发挥巨大的作用。该系统利于缩短讨价还价的过程，促进高效的共享交易，推动 P2P 共享租车个性化服务，促进 P2P 租车行业蓬勃发展。从而在满足大家出行需求的同时，降低汽车保有量，提高汽车利用率，节约资源。另外，由于 P2P 共享经济之间具有相似的数据分布和相同的预测目标，因此 P2P 汽车共享价格预测方法还可以简单地移植到其他的 P2P 共享经济中去。

§1.2 国内外研究现状

§1.2.1 汽车共享

在汽车共享研究领域中，大多数研究人员关注于汽车共享服务的市场形势，更进一步的研究包含考虑位置、交通行为、信息系统、电动汽车以及可持续性[8]。汽车共享是车主暂时将自己的汽车租赁给其他人，其中 P2P 汽车共享是一种新的方法。相对于直接拥有私家车，使用共享汽车的固定成本更少，可变成本更高[8]。数据显示，在中国，私家车的实际利用率仅有 7%，而汽车共享可以将利用率提升至 40%-60%，这可以减轻特大城市的交通拥堵问题。在北美洲，一辆共享汽车可以替代 9 至 13 辆私家车[9]。因此，相较于拥有私家车，共享出行服务可以降低环境污染和交通拥堵[10]。

P2P 汽车共享已经引起了学术界和工业界的浓厚兴趣。从市场的角度来看，大多数学术界研究人员关注于 P2P 汽车共享市场的可行性和潜力。Degirmenci K 和 Breitner M H[8]回顾了汽车共享的市场分析、交通行为、可持续性。从运输的角度来看，研究人员关注于降低能量消耗和温室气体排放[5]。另外，汽车保险问题也是一个重要的研究点[6]。本文，我们并不关注汽车共享的市场或者交通，而是关注于 P2P 汽车共享租赁价格预测。

§1.2.2 价格预测

房价[14 - 15]、股价[19]、油价[25]等方面的价格预测，已成为广泛研究的问题，它吸引着来自于包括经济、金融、数学、计算机科学等众多领域的研究人员。

1）股价预测：在金融和计算机科学领域，股价预测是最受欢迎的研究课题之一。大多数研究利用不同的机器学习方法来预测未来股价，例如人工神经网络。Schoneburg[11]用神经网络分析了短期的股价预测概率，并提出了预测次日股票价格涨跌的模型。不同于对次日股价的预测，几位研究者[12 - 15]预测了每日股价变化的趋势，并认为人工神经网络是最佳的预测模型。

最近的研究显示，由于股票市场数据的强噪声和高维度，人工神经网络学习特征

上有局限性，Kim 和 Han 提出了人工神经网络和遗传算法的混合模型对特征离散化，来克服以上问题。除了人工神经网络，支持向量机是另一种用于股价预测的主流模型[17 - 18]，Manish 和 Thenmozhi[17]使用支持向量机和随机森林预测股票交易量，他们采用了与人工神经网络相似的输入。Kara 等人[19]使用人工神经网络和支持向量机预测股价指数，这两种方法都表现出了良好的性能，同时模型还可作为有效的预测工具来预测股票价格。

2）油价预测：由于三分之二的世界能源消耗来源于原油和天然气，石油在全球经济中扮演着重要角色，所以很多研究者对研究油价波动很感兴趣[20]。W.Xie 等人提出了基于支持向量机的算法来预测时间序列，并应用于原油价格预测。结果显示，支持向量机在每月的油价预测上优于其他方法[21]。整体预测正确率为 81%，支持向量机的成功是由于它考虑了油价预测中的非线性问题[22]。

Kulkarni 等人提出了一种基于多层反馈神经网络的模型，可提前三天预测短期原油价格[23]。而基于小波降噪的预测会使数据投影更接近真实值，Jammazi 等人提出了融合多层负反馈神经网络和小波分解的混合模型[24]。由于石油市场是不断变化的，而神经网络只能反映之前参数的平均值，所以它预测石油市场有一定局限性。Lee 等人提出了使用贝叶斯方法来预测油价的模型，与现有方法相比有更高的准确度[25]。

3）房价预测：对于潜在屋主、投资者和其他的房地产市场参与者而言，房价预测十分重要[26]。研究房价预测模型有两个主要的研究趋势：基于 hedonic 的回归方法[14]和人工智能技术。基于 hedonic 方法有不同的模型[27]。Limsombunchai 等人凭借经验将 hedonic 价格模型的预测能力与人工神经网络模型在房价预测上进行对比，实验证明人工神经网络模型表现略优于 hedonic 价格模型[14]。然而，由于基于 hedonic 的方法可以获取房价内在变化趋势，因此其仍然应用于部分实际问题中。在房价预测的算法中，人工神经网络和支持向量机是两个基本模型，一些研究使用了人工神经网络模型[15，28 - 29]或者与支持向量机的混合模型[30 - 31]来预测房价。

在价格预测领域，股票、石油、房屋这三个常见商品受到了广泛的研究，而这三种商品在我们的市场经济中起着重要作用。由于股票、石油和房屋具有投资价值，所以目前的股价、油价和房价预测模型，主要希望预测出未来价格走势和准确价格，以便辅助投资决策。而对于 P2P 共享经济，它与其他共享经济之间最大的区别就是去中心化，交易双方将共同讨论决定具体的交易价格和方法。如果使用回归方法得到一个具体的推荐价格，则最终的交易价格通常会受限于这个价格，而这与 P2P 共享经济中的去中心化原则相违背。因此，预测一个价格区间供租赁双方参考更为合适。另外，他们之间的特点也不相同，例如 P2P 汽车共享价格预测需要考虑交易便捷程度、车况、供求关系等诸多因数。因此，P2P 汽车共享服务的价格预测与其他价格预测问题并不相同，因此不能用现有技术解决，但这些问题的研究仍对 P2P 汽车共享价格预测有一定借鉴作用。

§1.2.3 离散化方法

数据离散化是一种将连续属性转换为离散属性的数据简化方法。数据离散化用来降低连续属性的总数据量[83]。数据离散化还可以定义为一种量化连续属性的过程[84]。使用连续属性需要大量存储空间，还需要更长的规则。因此，需要通过离散化方法将连续属性变为离散属性。使用离散属性还可以增加预测准确度。虽然数据离散化方法在预处理中处于极其重要的地位，但对于这个方法的研究还很有限。离散化过程包含将连续属性划分到几个区间中。随后使用标签化的区间，而不是真实数据的连续值。离散化数据的优点是其增加了学习准确度和处理速度，并产生了比连续数据更为紧凑、简练、准确的结果[85]。离散属性通常更容易解释和理解。另外，它也更容易使用和操作[83]。从 1993 年至今，为达到这个目的，已经有了超过 70 种离散化技术。

这些数据离散化方法可以根据几个方面的区别来归类，例如，有监督和无监督、分层和不分层、动态和静态、全局和局部、积极和懒惰、单变量和多变量[63], 83 , 85 - 87]。但总体而言，这些离散化方法的思路可以分为分层的方法和不分层的方法。其中分层方法可以分为三种子方法，即，分割（自上至下）、融合（自下至上）、组合方法。每一种方法又被划分为基于有监督的和无监督的方法。

- 分割方法。分割方法的主要概念是创造区间界限点。它以一个空的界限点列表开始，在离散化过程中通过拆分或分割区间来增加新的项到列表中[84]。这个方法也叫做自上至下方法。事实上，拆分技术也分为三种技术，即，无监督、有监督、无监督和有监督的结合技术。其中，无监督技术在离散化期间不使用任何类别信息。
- 融合方法。这个方法也叫做自下至上方法。因为这个过程将连续值的完全列表视为独立的界限点开始。随后在离散化过程中，通过融合区间来降低区间数量[84]。融合离散化既可以通过无监督方法，也可以通过有监督方法。其中 K 均值聚类离散化是基于无监督的融合离散化。
- 组合方法。有些离散化技术同时使用了分割和融合方法。例如，使用分割产生一个区间，再使用融合。同时使用分割和融合的离散化技术可分为两种方法，即，有监督和无监督方法。

不分层技术指的是不使用层次的离散化技术。这些技术也包括无监督和有监督方法。在无监督方法中，区间测量是基本的离散化测量，它先确定用于离散化连续属性的区间数量来。对于等概率区间离散化方法和等宽度区间离散化方法，就是使用了区间测量的无监督不分层离散化方法。这些不分层技术的基础包括，将数据划分到固定数量的区间内，并基于均值或者边界改变数据。所有的这些技术的主要问题包括：在高准确度的情况下寻找合适数量的区间；最小化区间数量；避免数据内部的区间重叠。

在选择离散化方法时，需要根据需求特点来选择[89]。其中一个需要考虑的因素是，需要离散化的数据是否包含类别或者目标值。这个因素与有监督和无监督方法十分相关，有的数据集不提供任何类别标签。注意这个因素与数据挖掘中类别标签不可知情况下的聚类任务十分相关。这个任务的策略是，先通过专家来创造类别标签，再选择有监督离散化方法。另一个选择是通过训练数据得到的无监督方法选择区间。我们可以使用这些策略中的不同技术。

§1.2.4 代价敏感学习

传统的分类算法的设计目标是得到最小的分类错误，并通常用分类准确率来衡量分类性能，但这必须在不同的错误分类会导致相同代价的前提之下。如果错误分类代价不同，那么准确率将无法全面地反映算法的分类性能，这就是代价敏感问题。例如，对一个展厅来说，如果面部识别系统将使用者错误识别为了入侵者，不允许其进入，这将会造成不便。但如果将入侵者错误识别为了使用者，允许其进入，这将会导致严重损失。

代价敏感问题最早开始于 1974 年[65]，Elkan 在 International Joint Conference on Artificial Intelligence 2001 上所发表的综述[66]，使得代价敏感问题被人们所关注。随后开始有大量研究学者开始投身于代价敏感研究，其研究成果也不断被应用于各个领域[57 - 59 , 67 - 73]。另外，通过分类算法解决回归问题中的代价敏感学习，已经基于传统分类方法进行了初步研究，实验结果显示，代价敏感学习对于通过分类算法解决回归问题具有有效性[60]。

目前研究较多的代价敏感学习是代价敏感分类方法，很多文献中提到的代价敏感学习一般指分类器的学习[1]。在实际应用中，不同的错误分类往往会带来明显不同的错误分类损失[2]。代价敏感分类大致分为两类：直接代价敏感学习和间接代价敏感学习[3]。

直接的代价敏感学习的主要思路是，通过修改学习算法，将错误分类代价直接嵌入学习算法，代表性的工作有代价敏感决策树、代价敏感逻辑回归模型、代价敏感支持向量机。Sahin Y 等人[75]在决策树的剪枝过程中加入了最小化代价函数，并利用信用卡欺诈数据集进行模型验证，证明该方法可以明显降低经济损失。代价敏感支持向量机的实现方法主要可以分为 3 种：1）修改核函数；2）在训练过程中加入惩罚因子；3）改变 Hinge 损失函数。方法一的代表研究为 Wu 等人的研究[76]，但是其隐含了无法合理解释的假设：线性支持向量机无法具有代价敏感性。方法二面对类别区分较大的训练样本时，惩罚因子会变大导致松弛变量接近 0，使得代价敏感支持向量机退化成标准支持向量机。方法三的代表研究为 Masnadi-Shirazi 等人[77]提出的代价敏感支持向量机；Lee 等人[71]提出的用于解决多分类问题的代价敏感支持向量机；Li 等人[78]

提出的半监督代价敏感支持向量机；Zhang 等人[57]提出的多分类代价敏感核逻辑回归和多分类代价敏感 K 近邻算法等。

间接代价敏感学习可以不改变原有算法，仅通过对数据进行预处理或者对分类结果进行后处理，在传统学习算法中引入了错误分类代价。因此，间接代价敏感学习通过简单的改进就能使现有的算法代价敏感化，所以间接代价敏感学习得到了更多的关注。目前，间接代价敏感学习的研究方向主要集中在两个方面：采样法和阈值法，它们分别是对现有算法进行预处理和后处理[4]。

采样法是对样本类分布进行调整来获得代价敏感性，即增加高代价样本的数量或者降低低代价样本的数量来构建一个新训练集，利用新训练集训练出的最小错误率的标准分类器也会具备代价敏感性[58，74]。Elka[79]提出的 Elkan 定理证明了抽样法的有效性，表明将负类样本的数量调整为$(1-p)(1-p_0)/pp_0$，可以实现目标阈值为 p 时等价于给定阈值 p_0 的决策。但是抽样法有一个明显的缺点，因为抽样过程是随机选择样本，所以减少样本的过程中容易剔除重要样本，增加样本的过程中容易造成过拟合。为了改善这个问题，Chawla 等人[80]提出利用已知样本虚构高代价样本来实现增加样本的 SMOTE 算法。

阈值法利用了贝叶斯最小风险理论，通过改变后验概率的阈值减少错分代价。例如，对于一个 2 类样本分类问题，假设正确分类样本的代价为零，而将样本预测为正类的风险为 $p(0|x)C(0,1)$，预测为负类的风险为 $p(1|x)C(1,0)$。当预测为正类的风险小于负类时，贝叶斯决策会优先将样本预测为后验概率较大的标签类别中，即将样本标记为正类：

$$p(1|x) \geq \frac{C(1,0)}{C(1,0)+C(0,1)} \tag{1-1}$$

间接代价敏感不改变原有算法，因此可以对所有算法进行代价敏感化，所以现有的代价敏感算法以间接法为主。而其中的阈值法由于不改变原始数据，得到了较多关注。因此我们之后仅关注间接代价敏感中的阈值法。

在多分类代价敏感学习中，假设有 d 个类别，则其类别标签为 $G=\{G_i\}$，$i=1,2,\cdots,d$，正类的标签为 i。代价分为三种类型：接受了本应拒绝的代价为 C_{IG}；拒绝了本应接受的代价为 C_{GI}；将一类错误分类为另一类的代价 C_{GG} [57]。代价敏感学习通常将最小化错误分类代价设为目标函数。假设测试样本标签为 y，其预测类别标签为 $\bar{y}$。则标签可通过下面的目标函数得到：

$$L(y)=\underset{\bar{y}\in\{G_1,\cdots,G_d,I\}}{\arg\min}\ loss(y,\bar{y}) \tag{1-2}$$

其中 $loss(y,\bar{y})$ 由代价矩阵决定。

近年来，随着机器学习算法研究的不断深入，集成学习模型被大量研究证明优于单一分类模型，仅在单一分类模型中加入代价敏感显然限制了其分类性能的提升。Xia

等人[81]在梯度 boosting 中利用直接代价敏感方法构建了代价敏感 boosting。Sun Y 等人[82]将代价敏感学习引入到 boosting 算法中，通过训练来解决类别不平衡问题。Zhou Z H 等人[56]提出了一种借鉴深度学习优势的集成学习方法，即深度森林。它具有极少的参数，并且对参数设置都不太敏感；在无论是大规模或者是小规模数据，以及在不同领域中，它的表现都是不错的。但深度森林目前还没有代价敏感模型。

§1.3 主要工作及组织结构

§1.3.1 主要工作

本文，我们基于连续的目标值，预测出离散化的目标值区间，现有的方法是通过分类算法解决回归问题[32]。而此方法并没有考虑到离群值对预测区间宽度均匀性的影响。同时，此方法的另一个缺陷也限制了其广泛应用，即预测区间越多，分类性能越差。另外，此方法也并没有考虑不同错分情况的区别。因此，此算法并不适合直接应用于价格预测的任务中。针对这些问题，本文通过改进的 K 均值算法和深度森林对其进行改善，为使错误分类的区间更接近真实区间，我们还提出了代价敏感深度森林算法，来进一步提高算法的分类性能。本文的主要贡献如下：

- 离群值严重影响了 K 均值的聚类结果分布。受 K 均值的均匀效应和孤立森林的启发，我们改进了 K 均值算法，使得离散化得到的区间宽度更加均匀。
- 离散化步骤产生的区间越多，分类器的性能就会越差。而我们发现，由于深度森林的自适应深度，它可以使准确度越差的基分类器获得更好的改进。由于任意分类器均可用作深度森林中的基分类器，因此，深度森林的自适应深度可以促进通过分类算法解决回归问题的发展，并为集成学习提供借鉴。
- 传统的深度森林忽视了不同错误分类情况造成的后果是不同的。本文首次提出了代价敏感深度森林，通过不同区间中心之间的距离设置错误分类代价，可以让错误的预测区间更加靠近真实价格区间。还可以通过不同的代价矩阵，将代价敏感深度森林应用到其他代价敏感问题中。

§1.3.2 组织架构

本文按如下方式组织：

第一章是绪论，首先介绍了论文的研究背景与意义，再对汽车共享、价格预测、离散化方法、代价敏感学习的国内外研究现状进行了分析，最后阐述了论文的主要工作及组织结构；

第二章是相关背景知识介绍，其中包括：通过分类算法解决回归问题、目标值离

散化方法、传统分类方法、集成深度学习，并展示了本文所用到的评价指标；

第三章是关于我们所提出的离散化方法，即改进的 K 均值算法，首先详细阐述了 K 均值的缺陷，然后展示了我们方法的灵感，再具体介绍我们所提出的方法，最后是实验结果及分析；

第四章是关于深度学习有效地提高了通过分类算法解决回归问题的性能及其原因，首先讲解了深度森林的自适应深度特点，然后介绍了实验所用到的数据集及预处理，再展示了实验结果，最后分析了影响分类性能的因素和深度森林提升的原因；

第五章是关于深度森林的代价敏感学习，首先介绍了其实现方法，随后展示了实验结果及分析；

文章以第六章的总结和展望结尾。

第二章 相关背景知识介绍

§2.1 通过分类算法解决回归问题

通过分类算法解决回归问题首先要把连续的目标值离散化为不同的区间，然后将每个区间视为一类，再用分类算法得到区间的预测。因此，此方法可以分为两个主要步骤：1）将连续的目标值离散化为区间值（§2.2），2）将一个区间视为一类进行分类（§2.3）。这个方法已成功应用于不同领域，例如估计软件缺陷[33]，学习回归规则[34]。

§2.2 目标值离散化方法

离散化步骤将连续的目标值离散成区间值，进而将回归任务转化为分类任务，在这个过程中必须解决两个重要问题：1）如何将连续数值划分到区间中，2）确定离散化需要划分的区间数量。现有的离散化方法大多数是针对特征值的，因此很多技术都是有监督方法。而通过分类算法解决回归问题中需要离散化的是目标值，因此只能选择无监督的离散化方法。同时，得到的离散值还需要用于后面的分类，因此离散化方法的选择还必须考虑到分类性能。

基于搜索的离散化技术使用错误分类代价作为平均值，来说明区间顺序，因此适合于通过分类算法解决回归问题。这一技术使用等概率区间、等宽度区间、K 均值聚类技术来选择界限点[64]。它还使用封装方法来寻找最终的区间数量[88]。具体的：

- 等概率区间：将连续值划分到具有相同数量样本的区间中。此方法产生了相等的类别频率，这可以完全避免类别不平衡对于分类器性能的影响。需要注意的是，如果同一数值的不同样本被分到了两个相邻的区间里，则将此数值的所有样本重新划归到前一个区间中。
- 等宽度区间：将整个目标值区间划分到具有相同宽度的区间中。太不均匀的区间宽度会影响价格预测系统的实际使用效果。例如，预测系统对一个未知样本的分类结果的范围是 10，而对另一个则是 1000，则认为这个模型预测的宽度十分不均匀。而等宽度区间方法产生了相等的区间宽度，这使得所离散化的区间宽度更加均匀。
- K 均值聚类：先通过等概率区间方法或等宽度区间方法来初始化区间中心，再将每个样本到其区间中心的总距离最小化。这种离散化方法反映了数据分布，因此其对于不同的数据集来说更为鲁棒。

其中前两种方法是简单的基于数据统计的离散化方法，对不同数据分布的鲁棒性差；最后一种方法是基于分类的离散化方法，能反映数据之间的类别关系，有利于后面的分类模型，但对离群值敏感。基于这几种离散化方法，RECLA 系统专门用来将回归问题转化成分类问题，从而可以直接使用现有的分类方法来解决这个新的问题[64]。

另外，还有一种新的集成方法，其在均匀分布的数据集上达到了比等概率区间方法更好的性能[35]。这一方法通过极端随机离散化方法来离散化目标值，极端随机离散化方法的通过随机集成来获取区间边界。由于其仅适用于均匀分布的数据集，所以并不能视为一种基本方法。又由于本实验的数据并非均匀分布，因此此方法在本实验中并不适用，之后将不再讨论。

在之前的离散化方法中，我们假设区间数量是已知的。因此，在选择了离散化方法之后，作为每个离散化方法的超参数，区间数量必须确定。增加区间数量可以降低每个区间的宽度，以及每个区间之间的距离，但这增加了分类困难度，反之亦然。因此，对区间数量的选择是平衡离散化步骤性能和分类步骤性能的结果。合适的离散化方法及区间数量可以降低类间距离，从而降低总体代价。同时，更优的分类模型可以将更多的样本正确分类，从而降低总体代价。因此，拥有较低的总体代价的模型，其离散化结果和分类结果均可令人满意。通常来说，使用封装方法来确定区间数量，通过尝试不同数量的区间，来选择出最低总体代价的模型[32]。

§2.3 传统分类方法

§2.3.1 支持向量机

支持向量机的基本思想是利用“支持向量”找到划分样本集类别的最优超平面。对于可分的样本空间，其最基本的思路是找到一个超平面，将不同类别的样本尽可能地区分。支持向量机的基本函数是如下式所示，它使得“支持向量”距离超平面的和最大。

$$\begin{aligned} &\min_{\omega,\mathrm{b}} \frac{1}{2}\|\omega\|^2 \\ &s.t.\, \mathrm{y}_i(\omega^T x_i + b) \geq 1, i = 1,2,...,m \end{aligned} \tag{2-1}$$

其中 ω=（$\omega_1;\omega_2;...;\omega_d$）为法向量，$b$ 为位移项。为了更简单的求解上式，根据 KKT 定律将其转换为其对偶问题，推导过程如式（2-2）至式（2-4）所示。

$$L(\omega,\mathrm{b},\alpha) = \frac{1}{2}\|\omega\|^2 + \sum_{i=1}^{m}\alpha_1(1-y_i(\omega^T x_i + b)) \tag{2-2}$$

$$\begin{aligned}\omega&=\sum_{i=1}^{m}\alpha_i y_i x_i\\ 0&=\sum_{i=1}^{m}\alpha_i y_i\end{aligned} \tag{2-3}$$

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j\\ &s.t.\sum_{i=1}^{m}\alpha_i y_i=0,\\ &\alpha_i\geq 0,\quad i=1,2,...,m\end{aligned} \tag{2-4}$$

求解如式（2-4）的对偶式可得α后，可知ω和b，由此得出超平面对应模型，如下式所示。

$$f(x)=\sum_{i=1}^{m}\alpha_i y_i x_i^T x+b \tag{2-5}$$

核函数对于构建超平面至关重要，有多种不同的核函数可供选择，其中最常用的有径向基核函数和线性核函数：

- 径向基核函数有两个超参数，即内核系数、核惩罚参数，因此使用径向基核函数的支持向量机模型必须要调参。
- 线性核函数仅有惩罚参数，但不同的超参数设置产生基本相同的性能，因此使用线性核函数的支持向量机模型不需要调参。

另外，径向基核函数可以近似模拟线性核函数，因此线性核函数的表现通常略低于最优的径向基核函数，但线性核函数的计算量却少得多。

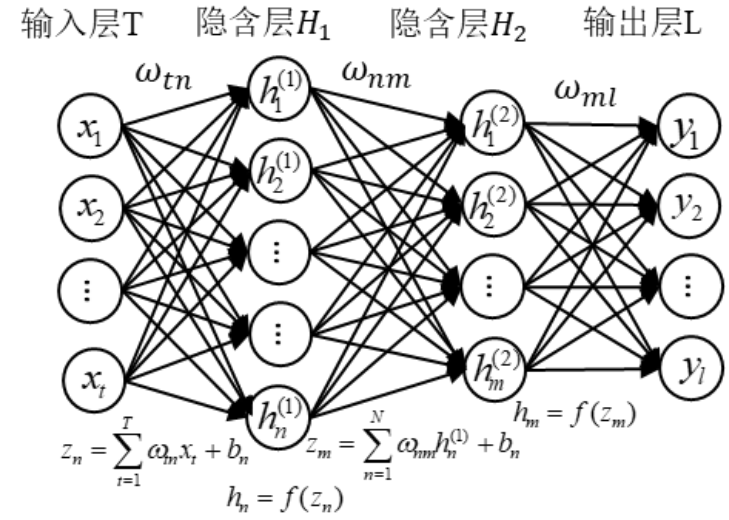
支持向量机是设计用来进行二分类的，同时利用投票框架扩展后还可用于多分类[38 - 40]。多分类问题中扩展的支持向量机方法有很多，例如一对其他、一对多[41 - 42]，其中一对其他方法的计算量小于一对多。

§2.3.2 多层感知器

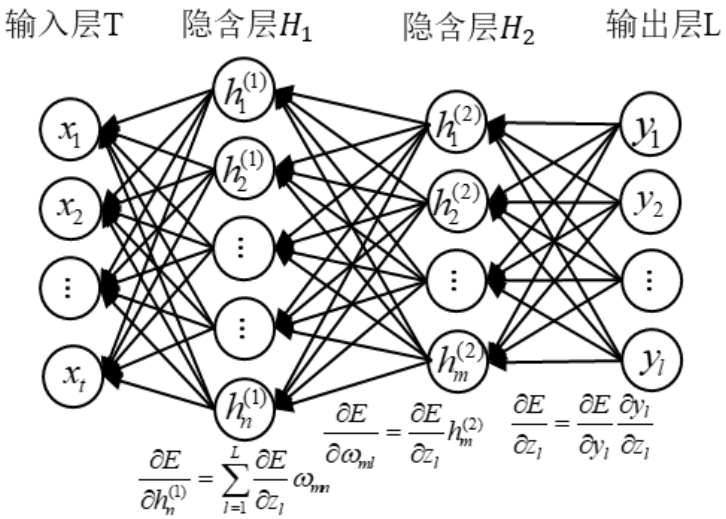
虽然神经网络的计算量很大，但它能在很多情况下得到比其他算法更好的结果。但对于神经网络模型，不同的任务需要构建不同的结构，这个过程十分繁琐。另外，神经网络的模型复杂度对于模型性能十分重要，其中深度比宽度更为重要[43]。

多层感知器算法是一种全连接的神经网络，也是最简单的神经网络模型。它由多个单层感知器组合而成，按照每层感知器的功能不同，可以分为输入层、隐含层和输出层，其中输入层和输出层均有且仅有一层，隐含层数量不限。每层网络包含若干个神经元，神经元的激活函数可以根据应用场景进行选择。包含两层隐含层的多层感知器结构示意图如图 2-1 所示，其中 2-1(a)表示网络的前向传播过程，2-1(b)表示网络的

反向传播过程。



（a）前向传播过程



（b）反向传播过程

图 2-1 包含两层隐含层的多层感知器

其中 x 表示输入样本，ω 表示层与层的权重，y 表示输出结果，z 表示节点输入和偏置的和，$f()$ 表示激活函数。

多层感知器输出结果可以与真实值对比计算误差，当输出结果等于真实值时，误差为 0，相差越大误差越大。最小化训练样本上的总损失是多层感知器的优化目标，在利用数学方法求解的过程就是网络的“学习”过程或者利用样本对网络的“训练”过程。在网络训练阶段，首先将信号在网络节点中前向传播，每个节点的操作可以分为 2 步：1）将上一层节点的输出做线性组合，2）对中间值进行非线性变换作为输出值。然后在网络预训练完成以后利用反向传播算法，通过使用链式求导方法计算期望结果与输出结果的误差，来反向调整权重，直到误差达到停止条件。

§2.3.3 随机森林

随机森林分类器是一种基于决策树的集成分类器[44]。在分类问题中，每个决策树都是一个分类器，然后汇总所有的结果，得到最多投票的类别视为最终的输出。由于其具有样本随机性和特征随机性，所以可以在预测精度提高的同时减少过拟合现象。具体的：

- 样本随机性。首先随机森林可以采用有放回抽样和无放回抽样这两种随机采样方式创建子数据集，保证了子数据集中数据的随机性。其次，每个子数据集均会训练一颗决策树，产生一个分类结果。最后，最终预测结果是由所有的决策树结果按照集成学习中多数投票法得出的。
- 特征随机性。随机森林在构建决策树时会进行一个重要过程：选择随机特征变量，即按照一定规律随机选择部分特征属性，再从中选择最优特征属性构建决策树模型，而产生随机特征变量的方式主要通过对输入变量进行随机选择和随机组合。这样可以降低决策树间的相似度，提高子模型的多样性，增强随机森林的泛化能力。

图 2-2 所示为生成随机森林的流程图。主要包括以下几步：

1) 利用有放回或者无放回的抽样方式从训练集中选择一定数量的样本，形成样本子集；
2) 从属性集 A 中按随机选择或者随机组合的方式选择特征属性；
3) 根据选择的样本子集和特征属性按照一定方法构建决策树，如 ID3、C4.5 和 CART；
4) 重复以上 3 步，构建足够数量的决策树学习模型；
5) 随机森林学习模型构建完成，将每颗决策树生成的预测结果按照多数投票方式确定最终预测结果。

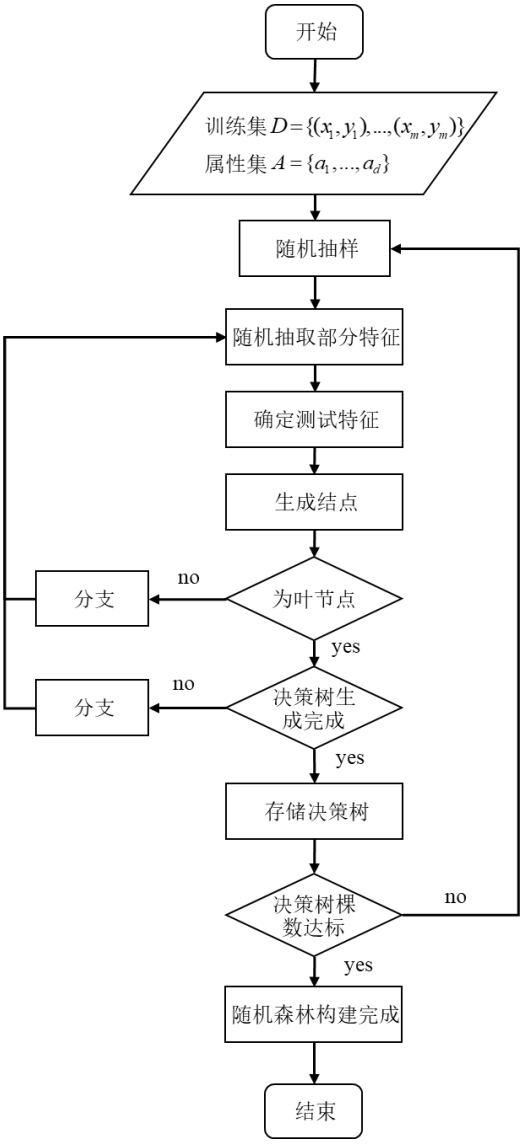


图 2-2 随机森林流程图

完全随机树是随机森林的一个变型，它在树的每个节点随机地选择一个特征来划分，并生长到每个节点上的叶子都是同一类别的[45]。

随机森林虽然在预测精度和过拟合现象上相比决策树有很大的改进，但是随机森林仅仅是对于决策树在同一层次上的集成。但相较于深度森林而言，随机森林还是有所不足。深度森林可以在不同层次上集成，而且可以根据深度的不同，自适应地调节多样性程度，还能在同一层集成各种不同的基分类器。深度森林是集成深度学习的一种，它结合了集成学习和深度学习的优势。

§2.4 集成深度学习

§2.4.1 集成学习

对于大多数分类任务而言，集成学习模型通常可以得到比单个模型更好的性能[46]。图 2-3 所示为集成学习的结构图，先产生一组“基分类器”，再通过某种策略将其结合，既可以包括同种类型的分类器也可以包含不同类型的分类器，其中随机森林是同种类型的分类器的典型代表。

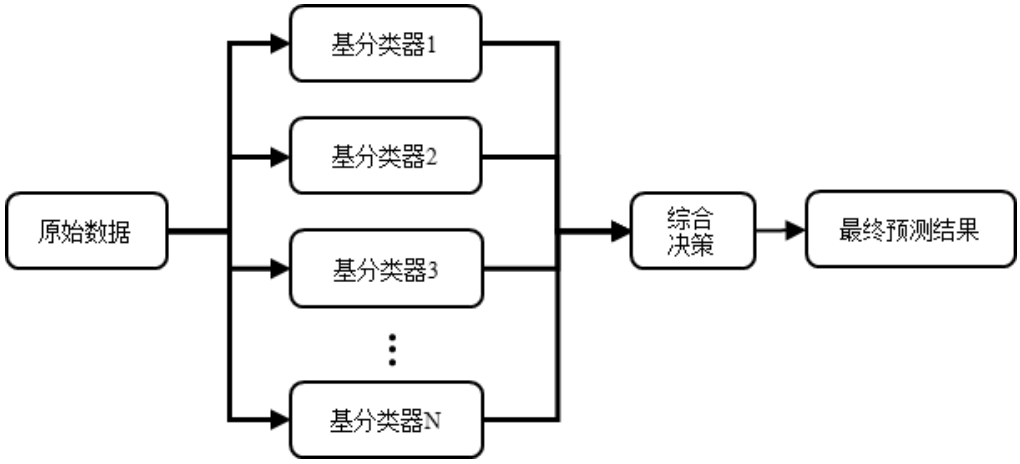


图 2-3 集成学习结构图

集成学习的结合策略可以分为三类：1）平均法，对预测值取平均；2）投票法，根据预测值进行投票；3）学习法，对预测值进行学习。平均法主要应用于数值输出型的基分类器，又可以分为两种方式，如式（2-6）和式（2-7）所示。投票法主要适用于直接输出预测标签的学习器，又可以分为 3 种方式，如式（2-8）至式（2-10）所示。学习法则是对基分类器输出标签进行再学习，其典型代表是 Stacking 方法。

$$H(x)=\frac{1}{T}\sum_{i=1}^{T}h_i(x) \tag{2-6}$$

$$\begin{aligned} &H(x)=\sum_{i=1}^{T}\omega_i h_i(x) \\ &s.t.\omega_i\geq 0,\sum_{i=1}^{T}\omega_i=1 \end{aligned} \tag{2-7}$$

$$H(x)=\begin{cases} c_j, & if\ \sum_{i=1}^{T}h_i^j>0.5\sum_{k=1}^{N}\sum_{i=1}^{T}h_i^k(x) \\ reject, & otherwise \end{cases} \tag{2-8}$$

$$H(x)=c_{\arg\max_j \sum_{i=1}^{T} h_i^j(x)} \tag{2-9}$$

$$H(x)=c_{\arg\max_j \sum_{i=1}^{T} \omega_i h_i^j(x)} \tag{2-10}$$

集成分类器的性能是由基分类器的性能和集成学习的多样性决定，整体而言，基分类器的性能越好，集成学习的多样性越大，集成分类器的性能通常会越好。对于分类性能还可以接受的基分类器，在适度提高集成学习的多样性后，集成学习可以得到令人满意的性能。在拥有足够的多样性，而且基分类器性能至还可以接受的情况下，继续提高多样性或者基分类器性能，并不会使得集成学习的性能得到较大提高[47]。集成学习中有四个策略用于提高多样性[48]：

- 数据采样操作，即采集不同的数据来训练不同的子分类器。例如 Bagging 方法[49]所使用的 bootstrap 采样[50]，AdaBoost 方法[51]所使用的 sequential importance 采样。
- 输入特征操作，即产生不同的特征子空间来训练不同的子分类器。例如，随机子空间方法[52]随机地选择一个特征子集来训练不同的子分类器。
- 学习参数操作，即为不同的子分类器设置不同参数。例如，每个神经网络使用了不同的初始权重[53]，每个决策树使用不同的划分选择[45]。
- 输出表示操作，即使用不同的输出表示来产生不同的子分类器。例如，EOOC 方法[54]使用误差纠正输出编码，Flipping Output 方法[55]随机改变一些训练样本的标签。

另外，不同的策略可以同时采用。

§2.4.2 深度学习

深度学习作为机器学习的一个重要分支，近年来在大量领域都取得了卓越成效。神经元是深度神经网络的基本组成部分，其思想起源于生物模型中的神经细胞，基本模型如图 2-4 所示。

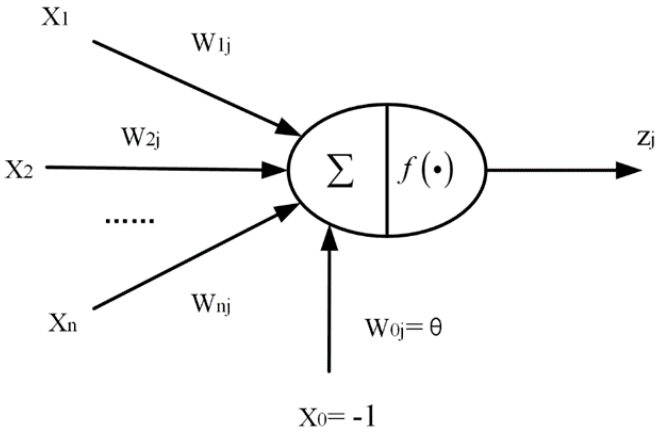


图 2-4 神经元的基本模型

其中，$x_0,x_1,x_2,\cdots,x_n$ 为输入，$W_{0j},W_{1j},W_{2j},\cdots,W_{nj}$ 表示每个输入对应的权重，$f(\bullet)$ 为神经元激活函数。数学表达式如下式所示：

$$Z_j = f\left(\sum_{i=1}^{n} W_{ij}X_i - \theta\right) \tag{2-11}$$

其中，常用的激活函数有：sigmoid 函数、Tanh 函数、ReLu 函数和 LeakReLu 函数，数学公式如式（2-12）至式（2-15）所示。

$$S(t) = \frac{1}{1+e^{-t}} \tag{2-12}$$

$$\tanh(t) = \frac{e^{t}-e^{-t}}{e^{-t}+e^{t}} \tag{2-13}$$

$$R(t) = \max(0,t) \tag{2-14}$$

$$f(t) = \begin{matrix} \alpha t & (t<0) \\ t & (t\geq 0) \end{matrix} \tag{2-15}$$

误差逆传播[90]是神经元构成的最简单的神经网络模型，目前主要的神经网络算法，如递归神经网络[91]、深度信念网络[48]、卷积神经网络[92]，均可以受到 BP 网络的影响。

在深度森林被提出以前，大多数人认为，深度学习近似等于深度神经网络[56]。其原因是深度学习在计算机视觉、语音识别、自然语言处理、强化学习等任务中获得了巨大成功，且其中几乎所有的深度学习应用都是基于深度神经网络。深度神经网络的成功主要是基于其三个重要特点：

- 逐层处理，借此逐渐提取出高层次特征，这对表示学习十分重要，而表示学习又对深度神经网络十分重要。
- 模型内特征转换，深度神经网络借此从上一层提取出信息，并传递给下一层作为特征，决策树和 Boosting 方法并不存在这一特点。
- 足够的模型复杂度，其宽度和深度可以随意增加来提高模型复杂度，其中增加深度来提高模型复杂度的效果比宽度更加明显[43]，而决策树和 Boosting 方

法并不满足这一特点。

§2.5 评价标准

本文主要采用了 3 个评价指标对模型进行评估：变异系数、准确度、代价。

变异系数（CV）是用来衡量不同值之间的均匀水平的评价标准，它等于均值除以标准差。给定一系列的数据 $X=\{x_1,x_2,\cdots,x_n\}$，则有：

$$CV=\frac{s}{\bar{x}} \tag{2-16}$$

其中

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{2-17}$$

$$s=\sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n-1}} \tag{2-18}$$

如果所有的值都相同，则变异系数为零。变异系数越大，说明数据越不均匀。

对于一个分类问题，准确度（accuracy）是最常用的模型评价指标，它等于被正确分类的样本占总样本的比例。它所代表的是样本被正确分类的情况，但是在面对代价敏感问题来说，错误分类的情况也需要考虑，所以准确度无法完全地评价代价敏感问题模型的优劣。

对于一般的代价敏感分类问题而言，主要有四种分类代价：正类预测为正类（CTP）、负类预测为负类（CTN）、正类预测为负类（CFN）和负类预测为正类（CFP）。通常情况下，正确分类的代价为零，即 $CTP=CTN=0$，并且假设正类样本是更受关注的样本，即 $CFN\gg CFP$。由此根据错误代价可以得到平均代价（cost）的计算公式如下所示：

$$\cos t=\frac{C_{10}FP+C_{01}FN}{n} \tag{2-19}$$

其中 n 为样本数量。

§2.6 本章小结

本章首先介绍了通过分类算法解决回归问题的步骤，然后详细描述了其中的目标值离散化方法和传统分类方法，再解释了集成深度学习的原理，最后展示了论文中用到的评价标准。

第三章 改进的 K 均值算法

在离散化步骤中，作为一种无监督学习方法，K 均值不仅可以挖掘数据之间的分布关系，同时还对不同数据集更为鲁棒，因此得到了广泛的应用[32，62 - 63]。但 K 均值对离群值较为敏感，容易得到区间宽度较不均匀的离散化结果（§3.1）。为解决这一问题，我们受 K 均值的均匀作用（§3.2.1）和孤立森林（§3.2.2）的启发，提出了改进的 K 均值算法（§3.3），得到了更为均匀的区间宽度（§3.4）。

§3.1 问题

商品的价格分布情况通常是：大多数价格位于小范围的高性价比区间，少数价格位于很大范围的高价格区间。换句话说，高性价比商品的价格分布较为密集，而高价格商品的价格分布则较为稀疏。图 3-1 展示了 P2P 汽车共享数据集中，价格和对应商品的数量关系。我们可以看到，几乎所有的价格都低于 1000 元，高于 1000 元的价格非常少，却分布在非常大的区间中。其他商品的价格也通常遵从类似分布。

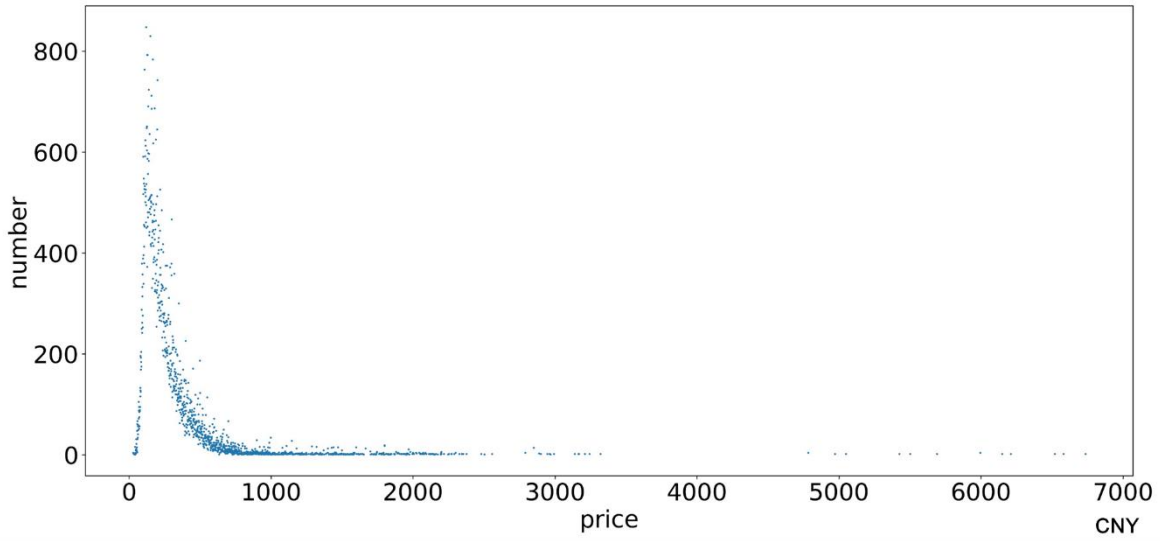


图 3-1 数据集的价格分布

等概率区间方法可以避免类别不平衡问题，但不能很好地反映目标值的分布。等宽度区间方法会将几乎所有的数据划分到前几个区间中，其余的区间将会仅有极少数据。因此，等宽度区间方法并不适合价格数据。在中国（其他国家可能也一样），为商品定价时，通常有一个有趣的规则，即许多商品的价格定为比整数少一元，也就是几十九元或者几百九十九元。这是因为，由于消费者更为关注高数位上的数字，所以

低数位上的数字通常就尽可能大。对于我们的数据，中位数为 199，上四分位为 299，这也在某些程度上证明，对消费者而言，每一百元之间有很大差距。因此，我们将价格按每一百元的区间来划分，最后一个区间则包含了剩余的价格。例如，如果区间数量为 4，则区间为$[0-99]$，$[100-199]$，$[200-299]$和$[300-+\infty]$。这种离散化方法称为每百元区间方法，它用来替换等宽度区间方法。考虑到中位数为 199，上四分位为 299，这种方法的最小区间数量设为 4。

对于数据集中的价格，我们使用土耳其检验来粗略观察离群值[61]。最小估计值 $E_{\min}$ 和最大估计值 $E_{\max}$ 如下：

$$E_{\min} = Q_1 - k(Q_3 - Q_1) \tag{3-1}$$

$$E_{\max} = Q_3 + k(Q_3 - Q_1) \tag{3-2}$$

其中 Q_1 和 Q_3 分别是下四分位和上四分位；高于 $E_{\max}$ 或者低于 $E_{\min}$ 的样本即为离群值；k 用来控制离群程度，若 k 为 1.5，选择出中度离群值，若 k 为 3，选择出重度离群值。图 3-2 展示了重度离群值的箱型图。我们发现，其中大约 97%的数据集中在仅有 9.7%宽度的范围内。换句话说，仅有 3%的数据是重度异常数据，但它们却占据了超过 90%宽度的范围，这一情况将严重影响离散化结果。

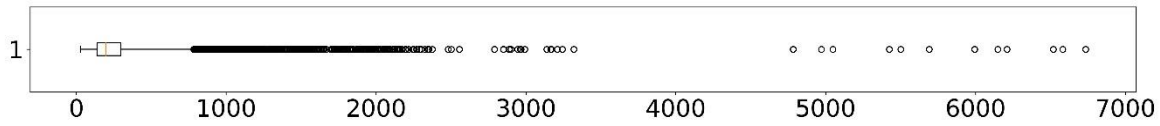


图 3-2 重度离群值的箱型图。其中箱子内的数据为正常数据，箱子右侧的粗线实际上是大量离群值的重叠，它们的点其实类似于图片右侧的系数点。

图 3-3 展示了等概率区间方法（EPI）、每百元区间方法（EOH）和 K 均值聚类方法（KM）中，区间内样本数量的变异系数（CVN）和区间宽度的变异系数（CVR）。EPI 的 CVN 最低，接近于零，但其 CVR 却是最高的。EOH 的 CVN 和 CVR 均低于 EPI 的 CVR，但整体上仍高于 KM 的 CVN 和 CVR。KM 的 CVN 整体上是下降的，直到区间数量高于 15，KM 的 CVN 才略低于 0.85。而在没有太多离群值的数据集中，由于 KM 的均匀作用（§3.2.1），CVN 通常介于 0.09 和 0.85 之间[36]，所以，KM 在此数据集上受到了离群值的影响。高 CVN 会导致类别不平衡问题，仅在区间数量很多的时候可以避免高 CVN，但这又会导致高 CVR。而高 CVR 又会使得不同区间之间的宽度产生很大差异，为更直观地解释高 CVR 的影响，我们展示了区间数量为 9 和 14 时，KM 所划分得到的区间，如表 3-1 和表 3-2 所示。可以看到，当区间数量为 9 时，KM 所划分得到的区间中，最宽的区间宽度是最窄区间的 24.7 倍；而当区间数量为 14 时 KM 所划分得到的区间中，最宽的区间宽度是最窄区间的 57.4 倍。当区间数量为 14 时，KM 的 CVR 高于区间数量为 9 时。也就是说，CVR 越高，不同区间

之间的宽度差异就越大。最宽区间宽度相对于最窄区间的倍数仅能反映这两个极端区间的宽度差异情况，而 CVR 可以反映所有区间之间的宽度差异情况，所以虽然不够直观，但我们仍使用 CVR 来而衡量不同区间之间的宽度均匀情况。为解决 CVN 和 CVR 的矛盾，使得在少量区间上即可得到低 CVN 和 CVR，我们改进了 KM 方法。

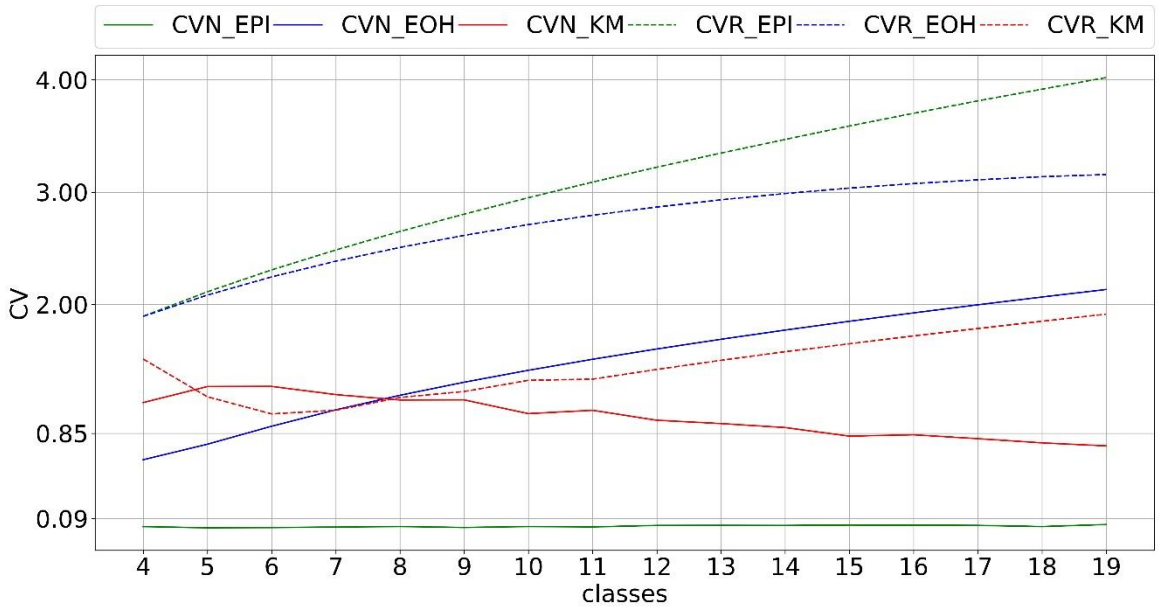


图 3-3 不同的传统离散化方法的 CVN 和 CVR

表 3-1 区间数量为 9 时，KM 所划分得到的区间

类别	最小值	最大值	均值	样本数量	区间宽度	密度
0	28	155	119.679	37410	127	294.567
1	156	235	192.102	32703	**79**	413.962
2	236	333	279.728	21456	97	221.196
3	334	459	388.639	12268	125	98.144
4	460	649	531.031	6904	189	36.5291
5	650	1040	769.723	2303	390	5.90513
6	1042	1701	1313.12	581	659	0.881639
7	1709	3320	2104.69	315	1611	0.195531
8	4783	6735	5741.64	28	**1952**	0.0143443

表 3-2 区间数量为 14 时，KM 所划分得到的区间

类别	最小值	最大值	均值	样本数量	区间宽度	密度
0	28	109	92.8531	11692	81	144.346
1	110	144	126.694	19951	**34**	586.794
2	145	183	163.157	19141	38	503.711
3	184	227	204.35	16599	43	386.023
4	228	280	251.914	14123	52	271.596
5	281	344	309.816	11516	63	182.794
6	345	423	380.355	8168	78	104.718
7	424	516	466.842	5954	92	64.7174
8	517	639	567.489	3416	122	28
9	640	855	711.334	1944	215	9.04186
10	857	1263	1002.77	797	406	1.96305
11	1277	1918	1550.5	448	641	0.698908
12	1928	3320	2296.08	191	1392	0.137213
13	4783	6735	5741.64	28	**1952**	0.0143443

§3.2 动机与灵感

§3.2.1 K 均值的均匀作用

K 均值方法有一个特点，即其所产生的聚类中样本的数量倾向于相对均匀分布[36]。Junjie Wu 等人[36]的研究表明，在没有离群值的数据集中，K 均值所产生的聚类中样本数量的变异系数，其 95%的置信区间在$[0.09, 0.85]$。换句话说，K 均值的均匀作用通常会使得其产生的聚类中样本数量的变异系数低于 0.85。如果其变异系数远大于 0.85，其原因是数据集中有大量的离群值，它们影响了聚类中的样本数量。因此，需要在离散化步骤之前，检测出离群值，并进行相应处理。

§3.2.2 孤立森林

在通过分类算法解决回归问题中，目标值是连续的，因此离群值的异常程度也是

相对的。具体来说，当区间数量较小时，不同异常程度的目标值被 K 均值方法分类到一个聚类中去，这将会提高聚类中样本数量的变异系数。随着聚类数量的增加，不同异常程度的目标值被分类到不同的聚类中去，这会降低聚类中样本数量的变异系数。因此，我们可以说，正常值和离群值之间没有明显的区别。正是如此，常用的通过划分出正常值轮廓来检测离群值的方法，并不适用于通过分类算法解决回归问题中。

Fei Tony Liu 等人[37]提出一种新的异常检测思路，即孤立森林，其将所有的样本彼此孤立来寻找离群值，而不是划出正常值的轮廓。这个方法的提出是基于离群值的两个特点：1）离群值的数量较少，2）离群值的特征值与正常值的特征值区别很大。因此，当基于特征值来随机地孤立所有的样本，离群值更容易从其他样本中划分出来，而正常值更为困难。孤立一个样本的困难程度，是由孤立这个样本所需要的分割次数决定的。相对于正常值，孤立离群值需要的分割次数更少。

§3.3 方法

在没有太多离群值的数据集上，K 均值产生k个区间，由于 K 均值的均匀作用，如果增加一个区间，K 均值所产生的新的$k+1$个区间将没有一个与之前的k个区间相同，但如果存在太多的离群值将会影响这一规则。对于有很多离群值的数据集，如果增加一个区间，新的区间将更加关注于正常样本，而忽视异常样本。这可能会导致，当区间数量增加时，有较多离群值的区间不发生变化。正如表 3-1 和 3-2 所示，虽然增加了五个区间，但区间数量为 14 时，KM 的最后一个区间和区间数量为 9 时的最后一个区间一样。孤立森林中定义离群值有一个特点，即离群值更容易孤立，并因此有更少的分割次数[37]。受此启发，我们定义，当区间数量增加时，不改变的区间为异常区间。异常区间的数量可能不唯一，为此，每次只改进最为异常的区间。我们为异常区间提出了三种不同的定义，即拥有最大范围的区间、拥有最少样本数量的区间、密度最小的区间。基于这三种不同定义的改进的 K 均值方法，分别叫基于区间范围的改进 K 均值（KMR）、基于区间中样本数量的改进 K 均值（KMN）、基于区间密度的改进 K 均值（KMD）。其整体步骤如下：

1） 将区间数量设为最小，进行 K 均值聚类；
2） 逐个增加区间数量，直至找到最为异常的区间；
3） 将异常区间中的所有价格重置为，朝向整体密集区域方向上的最近价格；
4） 基于（3）生成的新的价格，重复（1）（2）（3）步骤，直到步骤（2）中的区间数量等于想要的数量；
5） 步骤（4）最终所产生的区间，就是改进 K 均值方法的结果。

算法 1 展示了改进的 K 均值算法的细节。

算法 1：改进的 K 均值

输入：原始数据，P；

改进的 K 均值中的区间数量，k；

最小区间数量，s

输出：改进的 K 均值所产证的k个区间，I^+

初始化之前的区间数量，$N^- = s$；

初始化之前的价格，$P^- = P$；

基于P^-通过 K 均值得到N^-个区间，I^-

while $N^- < K$ **do**

得到新的区间数量，$N^+ = N^- + 1$；

基于P^-通过 K 均值得到N^+个区间，I^+

if I^-和I^+有相同的区间 **then**

选择最为异常的区间，并将此区间中所有的价格重置为，朝向整体密集区域方向上的最近价格，来得到新的价格，P^+；

重新初始化，$P^- = P^+$；

重新初始化，$N^- = s$；

基于P^-通过 K 均值得到N^-个区间，I^-；

else

$N^- = N^+$；

$I^- = I^+$

传统方法通过代价确定区间数量，这限制了不同区间范围的选择。对于价格预测模型，预测区间范围需要由用户调节。因此，我们提出了选择区间数量和离散化方法的新的规则。首先，根据所需区间范围，选择区间数量；然后选择 CVN 和 CVR 均低的离散化方法。低 CVN 有助于更好地分类，低 CVR 有助于更均匀地预测区间宽度。

§3.4 实验结果及分析

为更好地对比不同离散化方法，应为所有的离散化方法设立最小和最大区间数量。考虑到每百元区间方法，最小区间数量设为 4。通过 K 均值方法将所有的价格离散化至 19 个区间中，如果忽略异常区间，仅考虑低于 1000 元的区间，最小区间范围为 18，最大为 200，这对于预测区间来说已经足够小了。因此，所有离散化方法的最大区间主观上设定为 19。之后的所有实验对比，都将在区间数量从 4 至 19 上进行。

图 3-4 展示了传统的 K 均值方法和改进的 K 均值方法的 CVN 和 CVR。我们可以看到，当聚类数量增加时，所有改进的 K 均值方法的 CVR 整体上都是在下降。

KMR 的 CVR 快速下降，在聚类数量为 9 时，CVN 和 CVR 均较低，但直到聚类数量增加至 16，这两个值均不再明显变化。这反映出 KMR 有能力快速降低区间样本数量和宽度的差异程度。其中 CVR 降低的原因是，KMR 的改进作用降低了离群值对 K 均值方法的影响。对 KMD 来说，KMD 通过区间密度来降低离群值的影响，其中区间密度连接了区间的宽度和样本数量，然后，K 均值方法的均匀作用一同降低了 CVN 和 CVR。由于 KMD 的改进作用和 K 均值方法的均匀作用同时影响了 CVN 和 CVR，所以当聚类数量为 14 时，KMD 的 CVN 和 CVR 均达到最低。由于 KMD 的改进作用并不直接作用于区间样本数量或者宽度，CVN 和 CVR 降低得较为缓慢。对于 CVN 和 CVR，KMN 与 KMD 的整体趋势相近，但当区间数量较高时，KMN 则略差于 KMD。所以，对于有异常值的数据集，KMR 在少量区间的情况下，可以降低离群值的影响，此时的预测区间宽度较宽；KMD 在大量区间时，可以得到区间宽度较均匀的预测区间，此时的预测区间宽度较窄。对于 P2P 共享汽车数据集，如果只需要较宽的预测区间，则选择 KMR 以及 9 个区间；如果需要精确的区间，则选择 KMD 以及 14 个区间。为更直观地来看改进情况，我们展示了区间数量为 9 时，KMR 所划分得到的区间，如表 3-3 所示；以及区间数量为 14 时，KMD 所划分得到的区间，如表 3-4 所示。可以看到，当区间数量为 9 时，KMR 所划分得到的区间中，最宽区间的宽度是最窄的 6.8 倍，而 KM 是 24.7 倍；当区间数量为 14 时，KMD 所划分得到的区间中，最宽区间的宽度是最窄的 9.8 倍，而 KM 是 57.4 倍。可以说明，改进的 K 均值方法对 CVR 的改进，使得不同区间之间的宽度更加均匀。

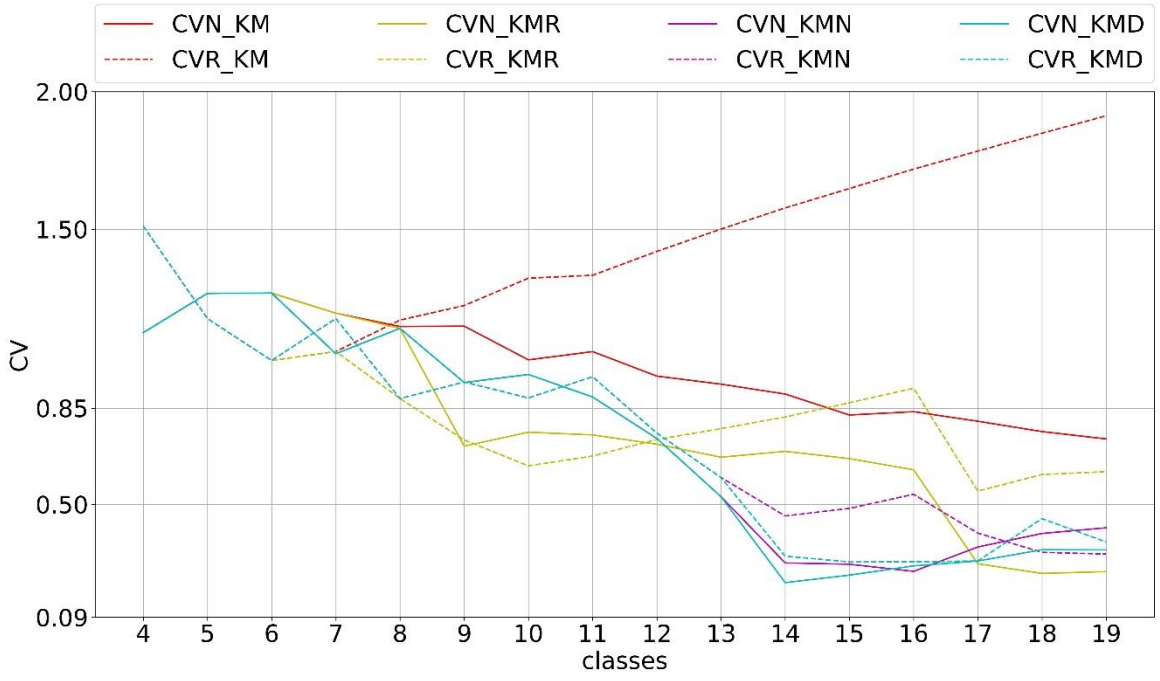


图 3-4 传统 K 均值和改进的 K 均值离散化方法的 CVN 和 CVR。由于一些方法在开头有相同的离散化结果，所以有些线的前面部分是重叠的。

表 3-3 区间数量为 9 时，KMR 所划分得到的区间

类别	最小值	最大值	均值	样本数量	区间宽度	密度
0	28	126	104.363	21572	98	220.122
1	127	174	149.493	25194	**47**	536.043
2	175	229	200.396	21362	54	395.593
3	230	295	259.417	16272	65	250.338
4	296	382	332.352	13054	86	151.791
5	383	494	432.353	8406	111	75.7297
6	495	659	557.223	5057	164	30.8354
7	660	979	762.834	1978	**319**	6.20063
8	980	1263	1196.52	1073	283	3.79152

表 3-4 区间数量为 14 时，KMD 所划分得到的区间

类别	最小值	最大值	均值	样本数量	区间宽度	密度
0	28	96	82.5935	5869	68	86.3088
1	97	124	111.368	14448	**27**	535.111
2	125	153	138.591	16187	28	578.107
3	154	186	169.3	15451	32	482.844
4	187	224	204.573	14534	37	392.811
5	225	267	244.477	12014	42	286.048
6	268	316	290.931	10198	48	212.458
7	317	375	342.951	8028	58	138.414
8	376	445	408.152	6100	69	88.4058
9	446	529	483.393	4809	83	57.9398
10	530	639	575.239	2922	109	26.8073
11	640	812	703.696	1826	172	10.6163
12	814	1078	923.575	682	**264**	2.58333
13	1079	1263	1232.93	900	184	4.8913

§3.5 本章小结

本章中，我们针对 K 均值对于离群值较为敏感的问题，根据 K 均值的均匀作用和孤立森林的启发，提出了改进的 K 均值方法，得到了区间宽度更为均匀的离散化结果。对于 P2P 共享汽车的数据集，KMR 在少量区间的情况下，可以降低离群值的影响，此时的预测区间宽度较宽；KMD 在大量区间时，可以得到区间宽度较均匀的预测区间，此时的预测区间宽度较窄。同时他还能应用在其他类似的价格预测模型中。

第四章 深度森林的自适应深度研究

对于传统的通过分类算法解决回归问题，通常使用封装方法来确定区间数量，即尝试不同数量的区间，选择出最低总体代价的模型[32]。但对于价格预测问题，则希望可以根据预测区间宽度的需要，来选择不同的区间数量。这样会在区间数量越多时，传统分类方法的分类性能越差。然而，深度森林却可以改进这一问题。本章中，我们先对分析了深度森林的自适应深度特点（§4.1），再通过实验（§4.2，§4.3），发现了影响分类性能的因素（§4.4.1）和深度森林提升的原因（§4.4.2）。

§4.1 深度森林的自适应深度

深度森林可以视为一种基于决策树的集成算法，这与随机森林相似，但深度森林的集成不仅是在同一层次上。深度森林几乎利用了§2.4.1 中所提到的所有策略来提高多样性，另外，深度森林还利用了§2.4.2 中所提到的深度学习的特点，使得集成学习借鉴了深度学习的优势。深度森林包含两个独立的部分，即级联森林和多粒度扫描[56]。级联森林用于构建集成分类器，多粒度扫描用于从具有空间或时间关系的原始数据中提取信息。因此，对于没有空间或时间关系的数据集，仅需要级联森林来构建深度森林模型。而对于有空间或时间关系的数据集，需要先利用多 粒度扫描来从原始数据中提取信息，然后将提取的信息传递到级联森林来训练分类器。考虑到所使用的数据集的特征较少，且没有空间或时间关系，因此本文只使用了级联森林，下文所指的深度森林仅指代级联森林。

深度神经网络由很多层组成，而每层网络是由大量的计算单元组成，深度森林的结构与之类似。不同于深度神经网络的计算单元仅是做简单的加权求和，深度森林的计算单元是一个完整的分类器。深度森林的每一层网络是估计器的集合，每个估计器都可以产生一个概率向量。因此，深度森林的每一层都将产生一系列概率向量。为了层数可以随意增长，不同层之间的估计器具有相同数量和类型。为提高集成学习的多样性，每一层网络内部可以包含不同类型的估计器，所有的估计器以不同的初始化状态开始。并且对每个估计器进行 K 折交叉验证，来提高多样性并降低过拟合的风险。在每个估计器中，每个样本均被训练了$k-1$次，这$k-1$个概率向量的平均作为此估计器在训练集上的输出。在一层新的网络训练完成后，使用这个层中的所有估计器在验证集上的概率向量的平均来评估这一层的性能。如果性能改善明显，将此层中的所有估计器在训练集上得到的概率向量与原始数据连接起来，作为下一层的输入。如果性能改善不明显，训练过程将会停止，停止明显改进的那一层的预测就是深度森林的预

测。因此，与大多数深度神经网络不同，深度森林的深度是自适应的。

从本质上来讲，深度森林是一种结合了深度学习优势的集成算法。由层和估计器构建的网络，使得深度森林和深度神经网络有相似的结构。因此，深度森林与其他集成算法最大的区别是深度，而深度是可以明显提高深度学习的模型复杂度[43]。对深度森林来讲，每一层中的每个单元是一个分类器，因此，网络越复杂，集成学习的多样性越高。而足够的多样性可以使得集成学习得到令人满意的性能，即使基分类器的性能仅仅还可以接受[47]。深度森林的这一特点，有效地改进了通过分类算法解决回归问题中，区间数量越多，分类效果越差的问题。

§4.2 数据集及预处理

本实验所使用数据集来自中国最大的 P2P 汽车共享平台——START 共享有车生活平台，涵盖了从 2017 年 10 月 16 日至 2017 年 10 月 29 日（不包含公共假期）国内三个主要城市的 116,145 条 P2P 共享汽车租赁信息，其中北京 52,700 条、广州 31,478 条、上海 31,967 条。之所以选择这两周的数据，首先是因为分析了更广泛时间段内的租赁价格和时间之间的关系后发现，通常来说节假日的出行需求大，租赁价格高，但节日的出现无明显规律，而周末则是周期性出现，因此为排除节日的影响，我们选择了不包含公共假期的数据，并将周几作为一个特征加入到模型中；又因为对于两周的数据量来说，我们进行一次完整的实验，就需要有 40 个核的工作站运行一周左右，因此我们目前拥有的运算力不足以支持更广时间的数据量。

我们选择了每个汽车的 19 个属性作为特征，具体信息说明如表 4-1 所示；租赁价格作为目标值，货币单位为元。原始数据集包含数字信息，但也有大量的文本信息，因此使用独热编码方法来将文本信息转化为数字信息。为避免在输入时较小值被较大值淹没，对所有数字特征进行归一化。模型的输入是特征和目标值，输出则是目标值的区间。我们随机且分层地选择数据集的 70%作为训练集，30%作为测试集。

表 4-1 模型所用到的特征

特征	解释
租赁次数	汽车已完成的交易次数
回复速度	车主的回复速度
车龄	汽车的车龄
里程数	汽车已行驶里程数，按区间展示， $<2,2-4,\cdots,>20$ 由数字 $1-7$ 来表示
排量	汽车排量， 大多数是小数，若是区间则用中位数来表示， 大多数单位是 L，通过 $1T=1.4L$ 来将 T 转化为 L
额外费用	超过三百公里后每公里的收费
座位数	汽车的座位数
GPS	汽车是否有 GPS 导航
MP3	汽车是否可以连接 MP3
当面交易	是否需要当面交易
变速器	汽车的变速器，自动挡或者手动挡
推荐	汽车是否被平台推荐
自动接单	是否可以自动接单
驾龄	对租赁者驾龄的要求
是否可出城	是否可以驾车离开城市
周几	发生交易在周几，用数字 $1-7$ 来表示
城市	交易发生的城市，包括北京、广州、上海
性别	车主的性别，包括男、女、未知
国家	汽车品牌的国家，包括中国、德国、美国等

§4.3 实验结果

§4.3.1 传统分类方法

我们对比了支持向量机、多层感知器、随机森林这三种传统的分类方法的性能。六种离散化方法分别将目标值离散化到 16 个不同的类别数量，即 4 至 19。所以，每种分类方法将会有 96 个分类任务。另外，为降低随机因素的影响，每个分类任务将会训练和测试多次，一般来说重复次数一般是 30、50、100 次。通过单一方法的实验，我们发现，重复到 30 次时，结果已趋于拟合，增加至 50、100 次会使结果更加稳定；

但考虑到即便仅重复 30 次，我们进行一次完整的实验，就需要有 40 个核的工作站运行一周左右，因此我们目前拥有的运算力不足以支持更多次的重复实验。因此，我们实验的重复次数定为 30 次。由于每种方法需要对多个分类任务进行多次重复实验，这极大地增加了计算量，为了尽可能地减少计算量，当性能相似时，我们选择计算量较小的超参数。另外，在多层感知机中对任务的调参也非常困难，对一个任务的参数进行调整都需要耗费大量时间，所以对所有任务调参几乎是不可能的。幸运的是，这些任务有相似的数据和目标，所以对于使一个任务达到最优性能的超参数，通常也会使另外一个任务接近最优性能。因此，我们在相同分类方法的不同任务上使用相同的超参数。

对于支持向量机来说，由于一对其他策略可以降低计算量，因此被用来进行多分类。支持向量机的超参数是核函数、内核系数和惩罚参数。先使用 K 均值离散化出 9 个和 14 个区间，并基于此进行调参。首先使用径向基核函数，惩罚参数从 $\{1,10,100,1000\}$ 中选择，内核系数从 $\{0.0001,0.001\}$ 中选择。当惩罚参数为 1000，且内核系数为 0.001 时，基于径向基核函数的支持向量机达到最佳。然后使用线性核函数，不同的惩罚参数产生相同的性能。线性核函数的准确度比径向基核函数的最佳值低 2% 左右，但线性核函数的计算速度却要比径向基核函数快得多。所以，支持向量机选择了线性核函数。

对于多层感知器，其超参数指的是网络结构，即层数和每个层中的单位数，由于不同层之间的单位数可以不同，所以这些对于一个任务调参已经太多了。考虑到提高深度神经网络的性能中，深度比宽度更为重要[43]。为了降低多层感知器的超参数，将每一层的单位数根据经验设为 100，所以分类的唯一超参数就是层数。多层感知器的层数从 1 到 7 进行测试，区间数量从 $\{4,9,14,19\}$ 中选择，对于所有离散化方法进行测试。结果显示，当层数为 4 时，大多数任务达到最优结果；当层数为 3 时，一些任务达到最优结果；当层数为 5 或 6 时，仅有个别任务达到最优结果。对于当层数为 5 或 6 时达到最优结果的任务，在层数为 4 时接近最优结果。对于当层数为 3 时达到最优结果的任务，增加一层通常仍然是最优结果。所以我们选择四层，且每层有 100 个单位为多层感知器的统一网络结构。

对于随机森林，唯一的超参数是决策树的数量。用 KM 离散化方法和 9 和 14 个区间进行测试。10 个树的准确度仅比 100 个树的低不到 1%，但却更快。所以树的数量设定为 10。

图 4-1 展示了支持向量机（SVM）、多层感知器（MLP）、随机森林（RF）在 $\{EOH,EPI,KM,KMR,KMN,KMD\}$ 离散化方法上的准确度。结果显示，随机森林得到了最佳性能，随后是多层感知器，与其他方法相比，支持向量机的结果很糟糕。但每一个分类方法有相同的问题，即区间数量越多，准确度越差，这严重限制了用分类算法解决回归问题的应用。所幸，我们发现深度森林可以改进这个缺陷。

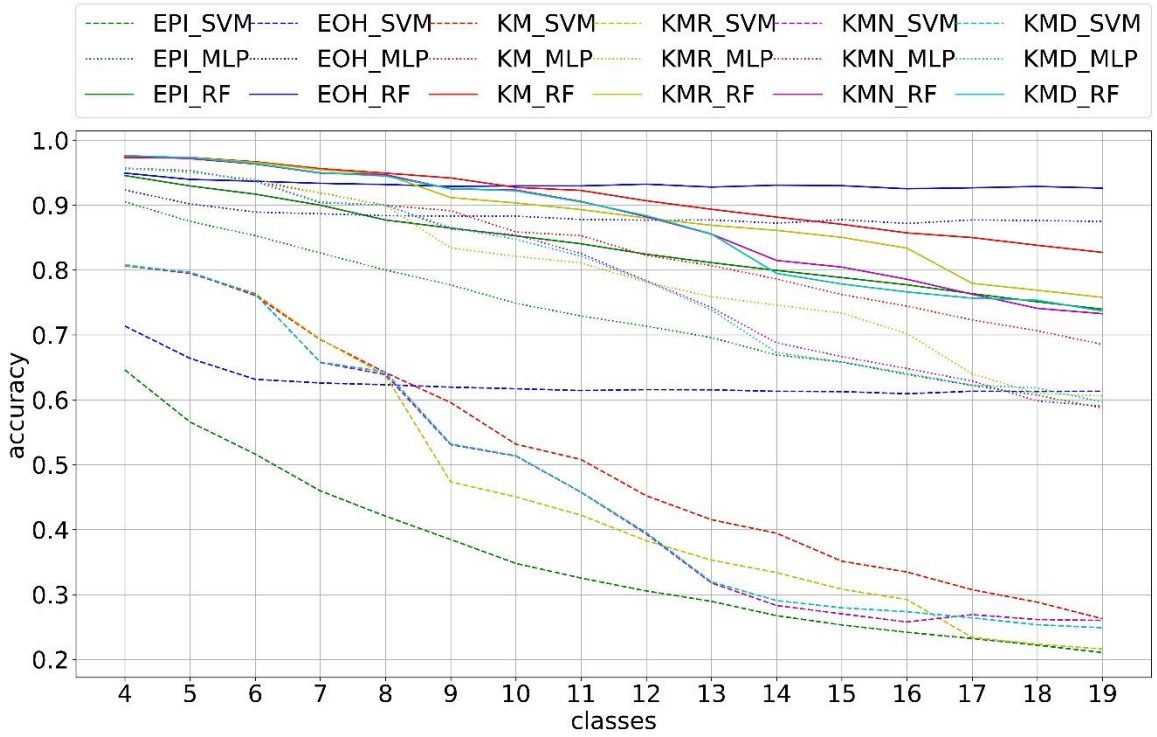


图 4-1 支持向量机、多层感知器、随机森林在不同离散化方法上的准确度

§4.3.2 深度森林

对于深度森林，它的超参数是每一层中基分类器的种类和数量，估计器中的折叠数量，改进程度，停止迭代的数量。因为深度森林对超参数不敏感，我们使用了文献[56]中所举例说明的超参数设置。即每一层使用了两个随机森林和两个完全随机树森林。每个森林使用了十个树，每个估计器中用到了 5 折交叉验证。改进程度为 0.01，这意味着少于 1%的改进将被视为不明显的。停止迭代的数量为 3，这意味着如果有三个连续的层没有明显改进，深度森林就会停止迭代。图 4-2 所示为随机森林和深度森林（DF）的准确度。当区间数量较小时，随机森林和深度森林的准确度相似。然而，除了 EOH 外，随着区间数量的增加，随机森林和深度森林的差距增加。具体而言，随机森林中最好和最差准确度差距是 21.3%，而深度则是 15.7%。我们可以说，深度森林分类器缓解了使用分类算法解决回归问题的缺陷，即区间数量越多，准确度越差。

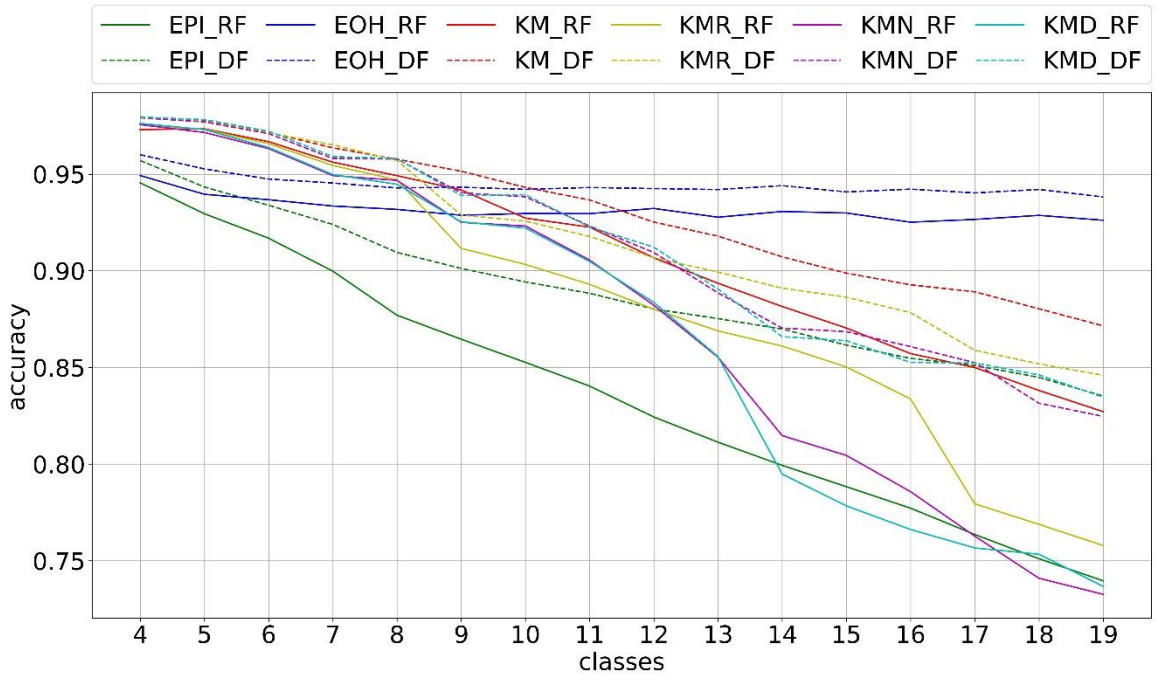


图 4-2 随机森林和深度森林的准确度

§4.4 结果分析

§4.4.1 影响分类性能的因素

对于同一分类方法，不同离散化方法和区间数量会导致不同的分类性能，为分析出影响分类性能的因素，我们单独研究了随机森林的准确度，如图 4-3 所示。分析不同离散化方法的准确度变化，我们可以发现影响准确度的主要原因。EPI 的准确度接近线性，随着区间数量的增加，每个区间的宽度会以一定比例下降，所以准确度以某一固定斜率下降。当增加一个区间时，EOH 将最后一个区间的前一百的宽度划分成一个新的区间。新区间与除最后一个区间外的其他区间有相同的宽度，因此新区间的分类困难度与它们相同。因为最后一个区间的宽度降低了一百，EOH 的准确度略有降低。总的来说，区间宽度影响分类的困难程度，区间宽度越大，越容易分类。例如，汽车包含巴士，对于一个计算机视觉系统，从不同的图片中分类出巴士比汽车更难。

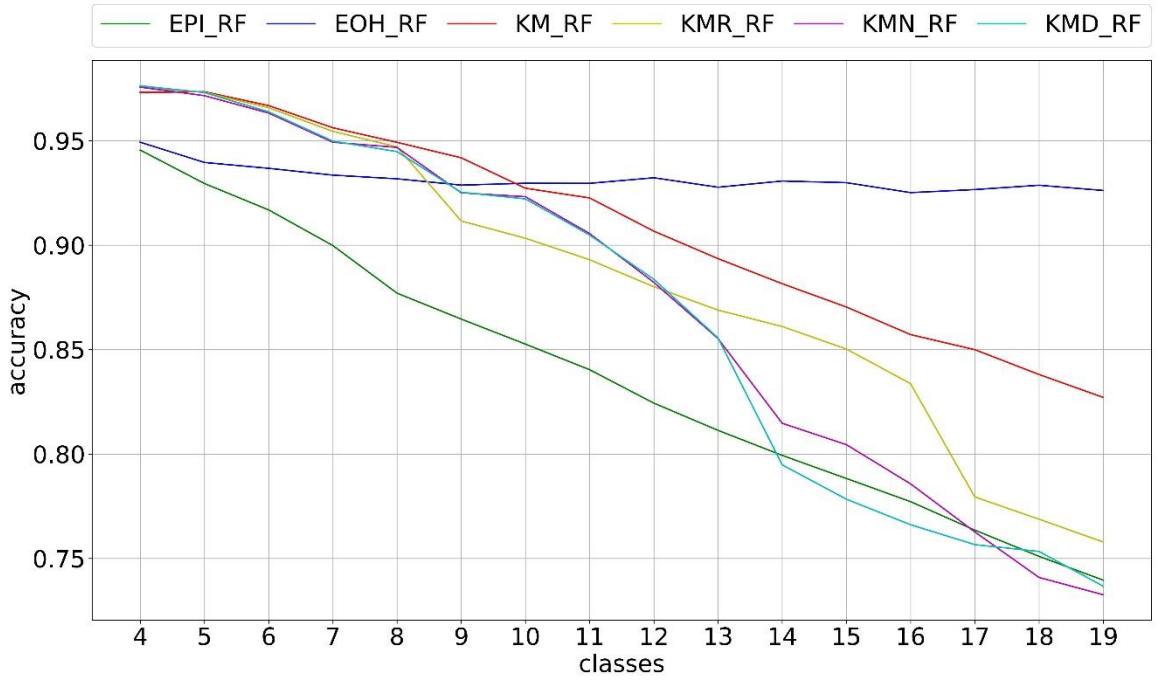


图 4-3 随机森林的准确度

§4.4.2 深度森林提升的原因

为进一步说明深度森林的提升，我们单独展示了深度森林相对于随机森林在准确度上的改进程度，如图 4-4。深度森林对于所有不同离散化方法，有不同程度上的准确度改进，最大的改进接近 10%。与图 4-3 相比，这两张图大致互为镜像关系。换句话说，随机森林的准确度越差，深度森林的改进就越大。因此可以说，深度森林有助于通过分类算法解决回归问题中，区间数量越多，分类性能越差的问题，这也为深度森林开辟了新的应用领域。

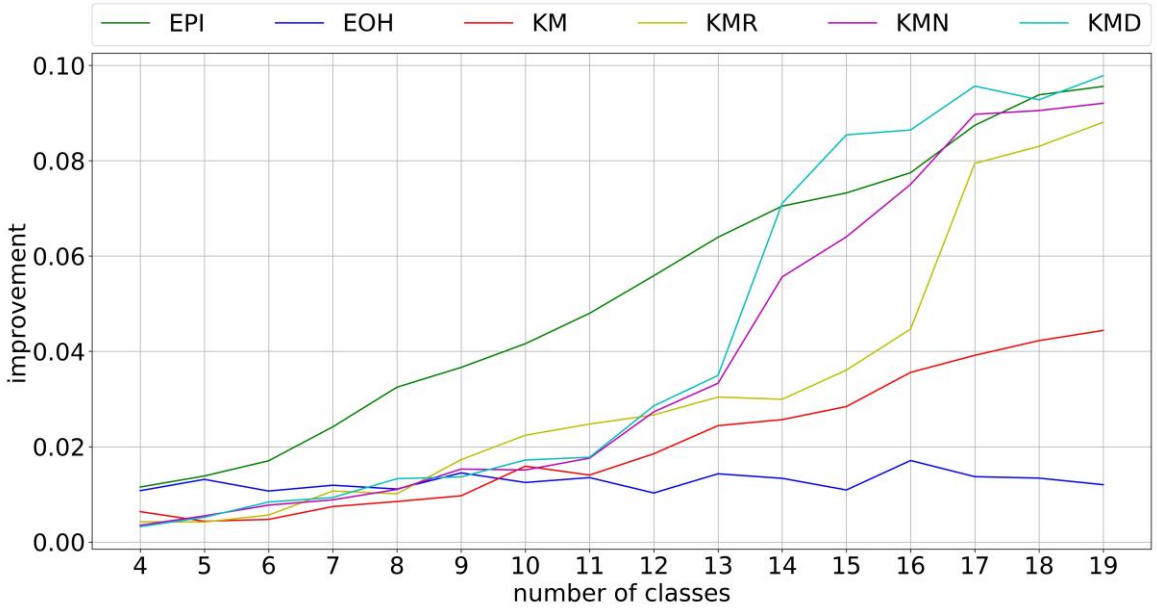


图 4-4 深度森林相对于随机森林在准确度上的改进

由于深度森林的基函数是随机森林，且超参数设置相同，所以深度森林相对于随机森林有较大改进的原因是，集成学习和深度学习所带来的多样性的提升，以及深度学习的模型内部特征转换。考虑到不同深度森林模型之间层的多样性水平是相同的，所以多样性水平的差异主要来自深度森林层数的不同。图 4-5 所示为深度森林达到最佳结果的层数。对比图 4-4 和图 4-5 中，准确度改进较小的深度森林的层数是 2，其原因是有助于深度学习的模型内部特征转换开始于第二层。对比准确度改进较大的部分，发现深度森林的改进程度和层数有相似的趋势，即层数越多，改进越高。所以，我们可以说，深度森林自适应的层数使得集成学习获得了足够的多样性，从而影响改进程度。从本质上来讲，是集成学习的多样性影响了深度森林的改进。

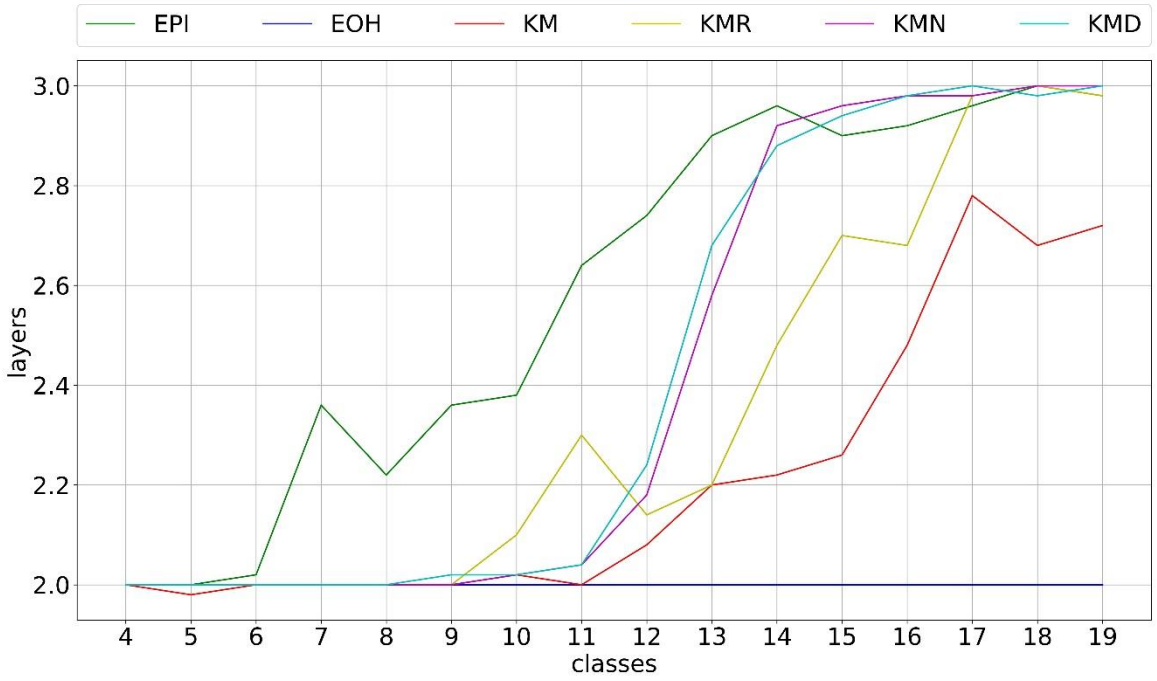


图 4-5 深度森林达到最佳结果的层数。注意，所有的任务训练并测试了 30 次，所以层数并不是整数。

因为除了随机森林外，深度森林还可以使用不同类型的基分类器，所以深度森林不仅可以对随机森林进行改进，还可以改进其他传统的分类器。例如，用不同的核函数和超参数的支持向量机作为基函数，或者使用不同结构的多层感知器作为基函数。可以说，深度森林的成功，为集成学习提供了新的思路。

§4.5 本章小结

本章中，我们将价格预测问题视为多类别分类问题，讨论了几个基本的传统算法，即支持向量机、神经网络、随机森林，以及最新的深度森林。结果发现，传统算法在价格预测问题中，区间数量越多，分类性能越差。这主要是因为，区间宽度影响预测的困难程度，区间宽度越大，越容易预测。还发现，随机森林的准确度越差，以其为基分类器的深度森林就会得到更高的改进。这主要是因为，深度森林的自适应深度使得集成学习获得了足够的多样性，从而改进了集成学习的性能。这为深度森林开辟了新的应用领域，同时为集成学习的集成方法提供了借鉴。

第五章　代价敏感深度森林

由于不需要太多的数据、计算量、调参，深度森林目前在一定程度上弥补了深度神经网络的不足，甚至在一些问题上取得了更好的分类效果，但目前它还没有用来解决代价敏感问题的代价敏感模型。对于通过分类算法解决回归问题，将真实价格区间“500-599”错误分类为“400-499”，相较于错误分类为“100-199”，在准确度上并没有什么不同。但是，考虑到价格预测的特点，“400-499”比“100-199”更为合理，因为前者更接近真实价格区间，这样即使是错误的预测，也不会与真实价格相差太多。所以，通过分类算法解决回归问题也是代价敏感问题。我们基于每个区间之间的距离来构建代价矩阵，并提出了代价敏感深度森林（§5.1）。实验结果显示，与深度森林相比，它可以在保持相似的准确度情况下，得到更低的代价，即驱使错误分类更为接近真实价格区间；还能相对于其他代价敏感方法，在正确分类情况上获得较大提高（§5.2）。另外，通过不同的代价敏感矩阵，代价敏感深度森林还能用来解决其他代价敏感问题。

§5.1　方法

代价敏感深度森林方法（§5.1.3）是基于代价敏感基分类器（§5.1.2）的，对于代价敏感学习，需要先定义代价矩阵（§5.1.1）。

§5.1.1　区间中心及代价矩阵

对于通过分类算法解决回归问题，为得到代价矩阵，首先要找到一个可以代表整个区间的值。而最能反映一个区间内所有样本整体分布的值是区间的中心，常用的指标是均值和中位数。均值可以反映一个区间内所有样本的分布，但对于离群值较为敏感；中位数对于离群值较为鲁棒，但在反映所有样本的分布上，相对于均值来说较弱。为避免离群值的影响，大多数的研究都使用了中位数作为区间中心[32 -[34]。在价格预测问题中，由于离群值的存在，这里也选择区间的中位数作为中心。

在找到所有的区间中心后，将这两个区间中心的距离视为这两个区间的类间距离，而类间距离则可视为将一个类别错误分类为另一个类别的错误分类代价。由此，我们定义 c_{ij} 是将类别 i 错误分类为类别 j 的错误分类代价，代价矩阵 C 的定义如下所示：

$$C=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \tag{5-1}$$

其中，$c_{ij}=c_{ji}$，$c_{ii}=0$。可以看出，代价矩阵是基于错误分类代价构建的一个对角线为零的对称矩阵。

§5.1.2 代价敏感基分类器

基于所构建的代价矩阵，我们先设计出代价敏感基分类器。给定一个测试样本的真实类别标签 y，预测类别标签 $\bar{y}$ 通过优化下面的目标函数获得：

$$\bar{y}=\underset{\bar{y}\in\{I_1,\cdots,I_n\}}{\arg\min}\, loss\left(y,\bar{y}\right) \tag{5-2}$$

$$loss\left(y,\bar{y}\right)=\sum_{i=1}^{n}P\left(\bar{y}_j\,|y\right)c_{ij} \tag{5-3}$$

其中，I_j 指的是预测标签为 j，而真实标签为 i 的这种情况；$P\left(\bar{y}_j\,|y\right)$ 指的是当真实标签确定时，预测标签的后验概率。

评价分类性能所用到的代价，是所有测试数据的平均代价。相比之下，准确度仅反映了正确分类的水平，而代价不仅仅反映正确分类情况，还反映了错误分类的水平。因此，我们使用代价来评价代价敏感的模型性能。

§5.1.3 代价敏感深度森林

基于上面提到的代价敏感基分类器，我们提出了代价敏感深度森林，其不仅关注了正确分类的情况，还关注了错误分类的情况，最终不仅能得到与深度森林相近的准确度，而且付出的代价还低于深度森林。代价敏感深度森林的整体结构如图 5-1 所示。代价敏感深度森林与传统的深度森林有相似的结构，其包含自适应深度的层，每个层又包含不同的估计器，每个估计器又基于基分类器进行了 K 折交叉验证。

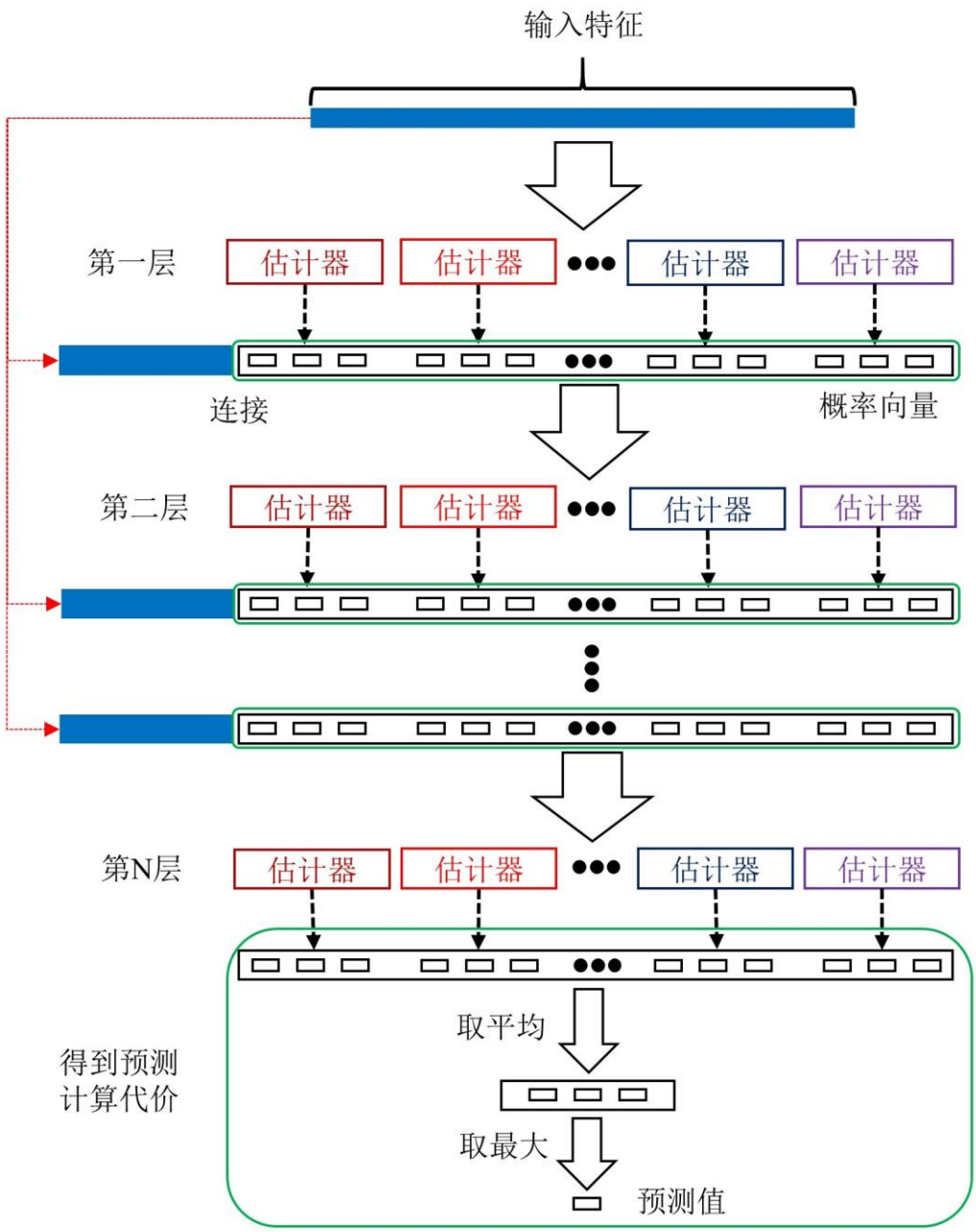


图 5-1 我们所提出的代价敏感深度森林的整体结构。其中估计器的不同颜色表示不同的基分类器。得到预测和计算每一层代价的方法如绿色框内所示。

在估计器层面，基分类器是代价敏感的，其倾向于提供代价更低的结果，所得到的概率向量中，越高的概率意味着越低的代价。估计器中使用到了 K 折交叉验证，所以将k个基分类器的所有结果连接起来之后，估计器预测出了$k-1$个在训练集上的概率向量和一个在验证集上的概率向量。平均这$k-1$个在训练集上的概率向量，可以得到这个估计器在训练集上的结果。算法 2 详述了如何通过 K 折交叉验证得到估计器

的概率向量。

算法 2：通过 K 折交叉验证得到估计器的概率向量

输入：数据，其中 X_k 是均匀且分层地划分的子数据集， $X=\{X_1, X_2, \cdots, X_k\}$；

标签， $y=\{y_1, y_2, \cdots, y_k\}$；

代价矩阵， C；

分类器， Cl

输出：训练集上的概率向量， P_{tr}；

验证集上的概率向量， P_{cv}；

估计器， E

for i 从1至 k **do**

使用除第 i 个子数据集外的所有子数据集来训练一个基分类器， Cl_i，并得到 P_{tr_i}；

使用第 i 个子数据集来测试基分类器，并得到 P_{cv_i}

连接所有的基分类器得到 E；

连接 P_{tr_k}，为每个样本得到 $k-1$ 个概率向量， P_{tr}'；

平均 P_{tr}' 得到 P_{tr}

连接 P_{cv_k} 得到 P_{cv}

对于每一层，每个估计器产生一个在训练集上的概率向量和一个在验证集上的概率向量。平均一个层中所有估计器在验证集上的概率向量，再找到其中的最大值得到预测，最后基于代价矩阵计算这个层的代价。所有估计器在训练集上的概率向量和原始特征连接后，作为下一层的输入。算法 3 详述了如何得出每一层的特征向量集合并计算代价。

算法 3：得出每一层的特征向量集合并计算代价

输入：数据，X；

标签，y；

代价矩阵，C；

一系列分类器，$Cl=\{Cl_1,Cl_2,\cdots,Cl_l\}$

输出：下一层的新特征，F；

层的代价c；

层，L

for j从1至l **do**

通过算法二训练并测试第j个估计器，得到F_j，及验证集上的概率向量，P_{cv_j}

连接所有估计器的F_j得到F；

平均P_{cv_j}再找出最大值作为预测，$\bar{y}$；

对比$\bar{y}$与y，根据C得到c；

通过训练所得到的分类器，Cl，构成L

对于整个代价敏感深度森林，将原始数据输入第一层，层中验证集的概率向量用来计算这个层的代价。如果测试集的代价明显降低，将所有的估计器在训练集上的概率向量与原始数据相连接，作为下一层的输入。代价停止降低的那一层的结果，就是代价敏感深度森林的结果。算法 4 详述了整个代价敏感深度森林的细节。

算法 4：代价敏感深度森林

输入：原始数据，X_{row}；

标签，y；

代价矩阵，C；

层中的一系列分类器，$Cl=\{Cl_1,Cl_2,\cdots,Cl_n\}$；

折叠次数，k；

停止迭代次数，s；

改进等级，p

输出：代价敏感深度森林

初始化层数，$n=1$；

初始化不明显改进的层数，$m=0$；

将 X_{row} 均匀且分层地划分进 k 个子数据集中，$X=\{X_1,X_2,\cdots,X_k\}$

while do

if $n>1$ **then**

将 F 与 X_{row} 连接得到新的数据 X

通过算法三得到 F，c，L

if $n=1$ **or** $c'-c>p*c'$ **then**

$n'=n$

$c'=c$

else

$m++$

if $m=s$ **then**

break

$n++$

连接前 n' 个层构成代价敏感深度森林

§5.2 实验结果及分析

在代价敏感深度森林中，我们所有超参数都与深度森林相同，只是在评估层是否需要增长时，使用的是代价是否下降。本实验中使用的所有数据也都与第四章相同。图 5-2 显示了代价敏感深度森林和传统深度森林的代价，明显地，代价敏感深度森林在几乎所有的离散化方法和区间数量上，所得到的代价都低于深度森林，其代价之和比深度森林低 5.6%。图 5-3 显示了代价敏感深度森林和传统深度森林的准确度，它们之间没有太大的不同。所以，与深度森林相比，代价敏感深度森林可以在相同准确

度下，达到更低的代价。换句话说，代价敏感深度森林的正确分类情况和深度森林相同，但对于错误分类的情况，代价敏感深度森林的结果更加接近真实区间。另外可以看到，EOH 的准确度虽然一直保持很高，但其代价却是最高的，这也是之前的方法中不推荐这一离散化方法的原因。

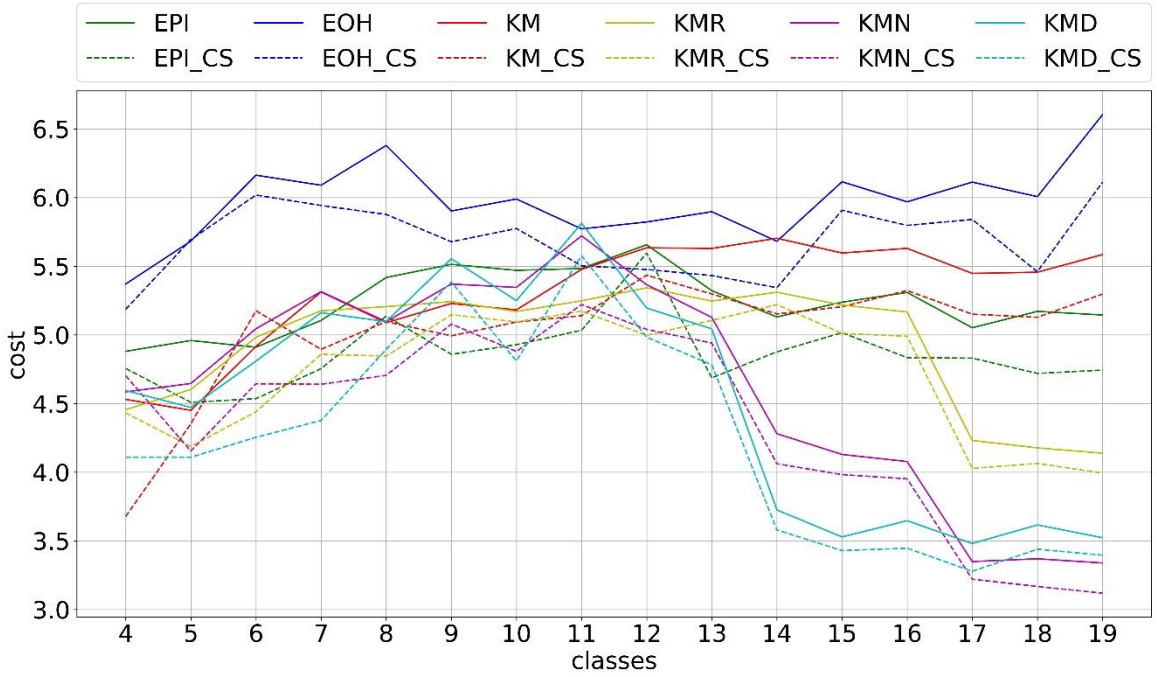


图 5-2 代价敏感深度森林和传统深度森林的代价。其中，虚线是代价敏感深度森林的代价。

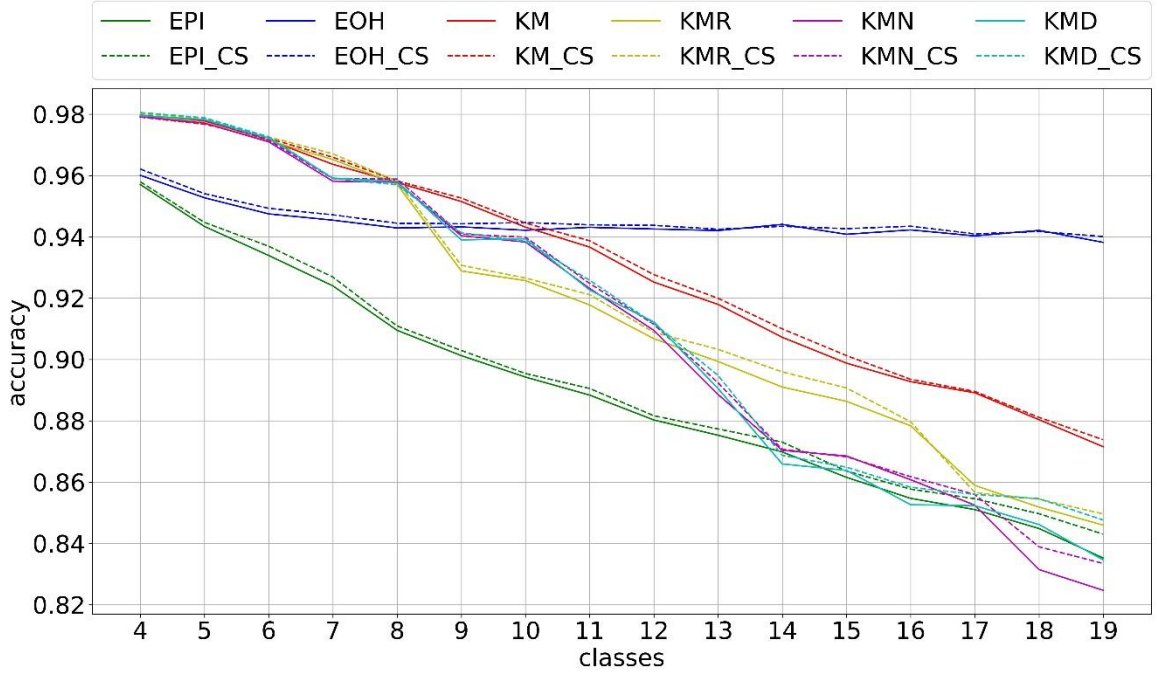


图 5-3 代价敏感深度森林和传统深度森林的准确度。其中，虚线是代价敏感深度森林的准确度。

针对代价敏感深度森林相对于传统深度森林在代价上的改进，我们详细对比了不同离散化方法和不同区间数量对于改进的区别，如表 5-1。计算同一离散化方法在不同区间数量上的和发现，EOH 的改进最高，KMN 的改进最低。由于传统深度森林只关注正确分类情况，而代价敏感深度森林既关注正确分类情况，又关注错误分类情况。这说明，EOH 离散化方法使得深度森林的错误分类造成了较高的损失，而 KMN 造成的损失则较低。

表 5-1 代价敏感深度森林相对于传统深度森林在代价上的改进。其中列代表不同的离散化方法，行代表不同的区间数量，最后一列是相同区间数量在不同离散化方法上的和，最后一行是同一离散化方法在不同区间数量上的和。

区间数量	EPI	EOH	KM	KMR	KMN	KMD	SUM
4	0.183872	0.124861	0.852291	0.485236	0.023703	-0.11749	1.552478
5	-0.01111	0.449951	0.098071	0.362563	0.414912	0.491588	1.805976
6	0.144697	0.374474	-0.25972	0.554507	0.542281	0.400816	1.757052
7	0.147389	0.350785	0.417538	0.784712	0.317099	0.674196	2.691719
8	0.502189	0.281239	-0.00744	0.201524	0.360054	0.391169	1.72874
9	0.224182	0.654725	0.23547	0.168611	0.095512	0.290813	1.669311
10	0.214791	0.541327	0.088467	0.438803	0.076128	0.466543	1.82606
11	0.268619	0.449611	0.336878	0.236427	0.075264	0.500038	1.866837
12	0.34589	0.060848	0.199854	0.212274	0.345905	0.325107	1.489877
13	0.46485	0.634545	0.33122	0.261613	0.140742	0.186624	2.019593
14	0.337813	0.254026	0.551076	0.145218	0.09046	0.218889	1.597482
15	0.207195	0.222133	0.391512	0.099679	0.206225	0.14679	1.273534
16	0.171545	0.474451	0.306933	0.200269	0.176944	0.125779	1.455921
17	0.272091	0.221336	0.296818	0.202562	0.203665	0.12855	1.325021
18	0.546427	0.451789	0.328221	0.177058	0.113065	0.201501	1.818061
19	0.491996	0.401021	0.287763	0.127025	0.14445	0.219961	1.672216
SUM	4.512437	**5.947122**	4.454951	4.658081	**3.326408**	4.650879	

接下来我们又对比了代价敏感深度森林和其基函数——代价敏感随机森林的分类性能。图 5-4 是代价敏感深度森林和代价敏感随机森林的准确度，发现它们与深度森林和随机森林的准确度（图 4-2）相似。图 5-5 是代价敏感深度森林和代价敏感随机森林的代价，表 5-2 是代价敏感深度森林相对于代价敏感随机森林在代价上的改进。可以发现 EPI 的改进最高，EOH 的改进最低，同时，整体而言，随着区间数量的增加，对不同离散化方法改进之和也越高，这些特点与深度森林相对于随机森林在准确度上的改进相似。说明代价敏感深度森林相对于代价敏感随机森林的改进，主要是由于深度森林算法在正确分类情况上的改进。总的来说，代价敏感深度森林相对于传统方法，在正确分类情况和错误分类情况上都有较大改进。

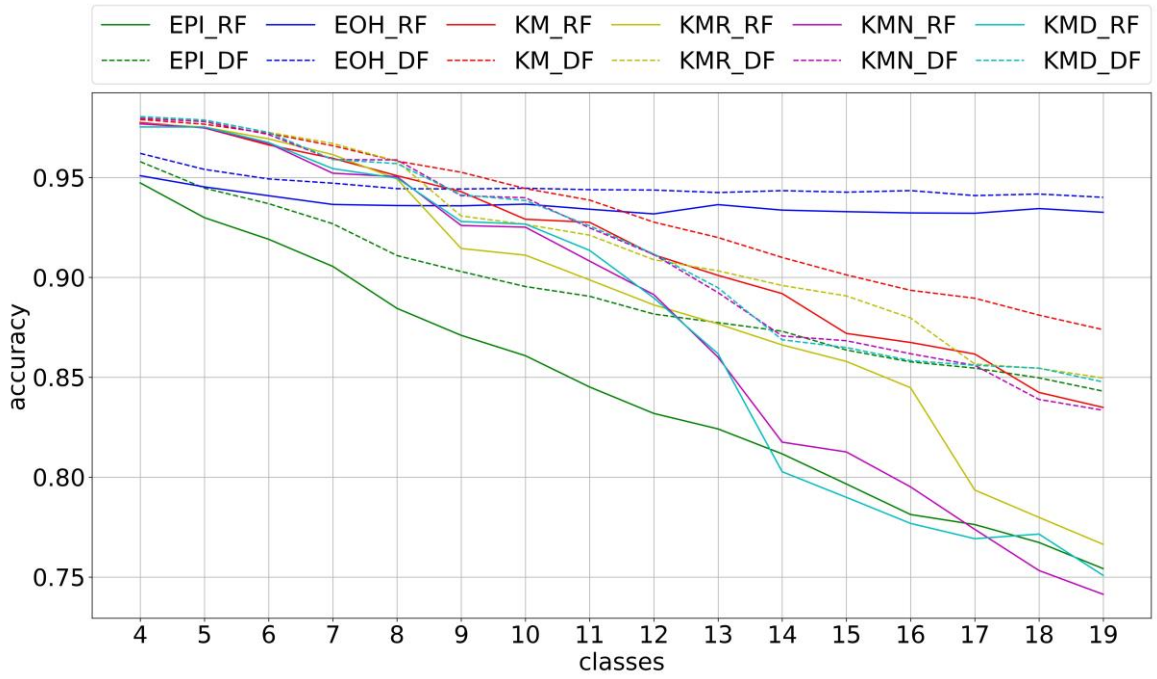


图 5-4 代价敏感深度森林和代价敏感随机森林的准确度。其中，虚线是代价敏感深度森林的准确度。

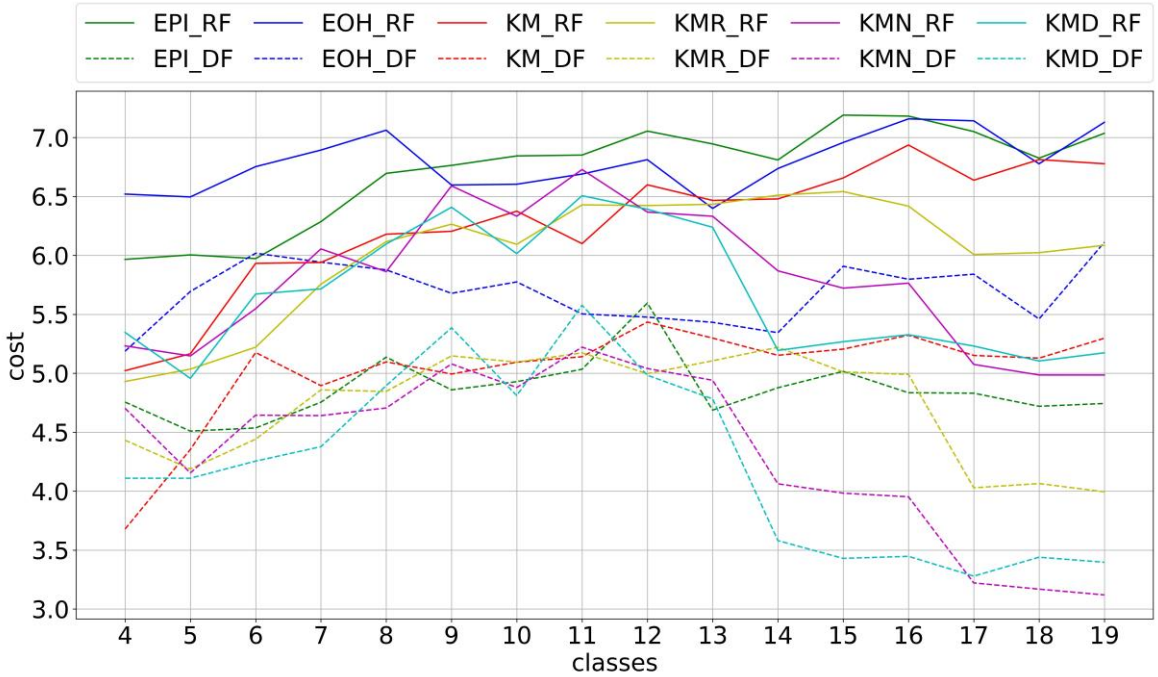


图 5-5 代价敏感深度森林和代价敏感随机森林的代价。其中，虚线是代价敏感深度森林的代价。

表 5-2 代价敏感深度森林相对于代价敏感随机森林在代价上的改进。其中列代表不同的离散化方法，行代表不同的区间数量，最后一列是相同区间数量在不同离散化方法上的和，最后一行是同一离散化方法在不同区间数量上的和。

区间数量	EPI	EOH	KM	KMR	KMN	KMD	SUM
4	1.210582	1.333159	1.343855	0.497542	0.529874	1.236681	6.151693
5	1.494946	0.801978	0.812263	0.847588	0.992259	0.847255	5.79629
6	1.436667	0.734349	0.756486	0.78021	0.903118	1.417294	6.028124
7	1.531764	0.951034	1.044901	0.896706	1.414055	1.338341	7.176801
8	1.558822	1.185041	1.083147	1.270758	1.156438	1.201015	7.455222
9	1.904353	0.918822	1.210822	1.117385	1.510436	1.022791	7.68461
10	1.914862	0.828855	1.280958	0.999231	1.452574	1.204214	7.680693
11	1.817088	1.18681	0.95957	1.255337	1.507298	0.929483	7.655587
12	1.458876	1.335568	1.164215	1.425585	1.326891	1.407623	8.118758
13	2.258188	0.966368	1.167708	1.327562	1.3913	1.455873	8.567
14	1.932508	1.392774	1.326385	1.290491	1.806971	1.613404	9.362533
15	2.173813	1.050009	1.452004	1.530046	1.739059	1.838224	9.783155
16	2.346689	1.360743	1.613205	1.427895	1.812282	1.880095	10.44091
17	2.219342	1.300981	1.486662	1.979429	1.853903	1.95184	10.79216
18	2.104941	1.316337	1.684511	1.95866	1.816853	1.664395	10.5457
19	2.29217	1.017854	1.480224	2.091168	1.865653	1.77695	10.52402
SUM	**29.65561**	**17.68068**	19.86692	20.69559	23.07896	22.78548	

§5.3 本章小结

在通过分类算法解决回归问题中，被错误分类到与真实区间不同距离的区间会产生不同的错误，因此很难用分类准确度这个单一评价标准来衡量算法优劣，有必要将代价敏感的机器学习算法引入此模型。本章基于上一章中深度森林的优秀分类性能，提出了代价敏感深度森林。相较于传统深度森林，代价敏感深度森林可以在相同准确度下，获得更低的代价，也就是说，其错误分类的区间更加接近真实区间。相比于其他代价敏感方法，代价敏感深度森林对于正确分类情况具有较大提升。同时，代价敏感深度森林不仅可以利用在通过分类算法解决回归问题中，它还可以根据不同的代价矩阵，来解决不同的代价敏感问题。

第六章　总结与展望

§6.1 总结

本文主要解决了 P2P 汽车共享中租赁价格预测问题，一般方法是通过分类算法解决回归问题。然而，这个算法的两个缺陷限制了其在价格预测中的直接应用，即其忽视了离群值对于预测区间宽度均匀性的影响，以及越多的区间将会导致越差的分类性能。同时，由于价格预测问题是代价敏感的，而深度森林却缺乏合适的代价敏感模型。为解决上述问题，本文进行了以下研究和创新：

1) 针对 K 均值对于离群值较为敏感的问题，我们根据 K 均值的均匀作用和孤立森林的启发，提出了改进的 K 均值方法。对于 P2P 共享汽车的数据集，KMR 在少量区间的情况下，可以降低离群值的影响，此时的预测区间宽度较宽；KMD 在大量区间时，可以得到区间宽度较均匀的预测区间，此时的预测区间宽度较窄。使得所提出的改进 K 均值算法降低了区间宽度的均匀水平，这使得离散化所产生的区间宽度更加均匀。同时他还能应用在其他类似的价格预测模型中。
2) 然后我们将价格预测问题视为多类别分类问题，讨论了几个基本的传统算法，即支持向量机、神经网络、随机森林，以及最新的深度森林。结果发现，传统算法在价格预测问题中，区间数量越多，分类性能越差。这主要是因为，区间宽度影响预测的困难程度，区间宽度越大，越容易预测。还发现，随机森林的准确度越差，以其为基分类器的深度森林就会得到更高的改进。这主要是因为，深度森林的自适应深度使得集成学习获得了足够的多样性，从而改进了集成学习的性能。也就是说，深度森林的自适应深度特点有助于使更差的性能获得更高的提升，这降低了少量和大量区间之间的性能差异。这为深度森林开辟了新的应用领域，同时为集成学习的集成方法提供了借鉴。
3) 在通过分类算法解决回归问题中，被错误分类到与真实区间不同距离的区间会产生不同的错误，因此很难用分类准确度这个单一评价标准来衡量算法优劣，有必要将代价敏感的机器学习算法引入此模型。基于深度森林的优秀分类性能，我们提出了代价敏感深度森林。相较于传统深度森林，代价敏感深度森林可以在相同准确度下，获得更低的代价，也就是说，其错误分类的区间更加接近真实区间。相比于其他代价敏感方法，代价敏感深度森林对于正确分类情况具有较大提升。同时，代价敏感深度森林不仅可以利用在通过分类算法解决回归问题中，它还可以根据不同的代价矩阵，来解决不同的代价

敏感问题。

以上三个研究都是基于中国 P2P 共享汽车数据集研究的，另外，我们提出的 P2P 共享经济的价格预测模型，还能用于其他的价格预测中。

§6.2 展望

由于汽车共享服务中租赁价格的周期性，一般是以周为单位。然而，本文仅根据两周的数据关注了短期预测，未来的研究将会考虑长期预测。另外，还需要探索更有用的信息来提高模型表现，例如评价和特殊日期（即节假日）。

参考文献

[1] X Xu, L He, H Lu, A Shimada, et al. (2016) Non-linear Matrix Completion for Social Image Tagging. IEEE Access, PP (99): 1-1.

[2] J. Wright, A. Y. Yang, A. Ganesh, et al. (2008) Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) 210-227.

[3] Waqas J, Zhang Y, Lei Z. (2013) Collaborative neighbor representation based classification using l 2-minimization approach. Pattern Recognition Letters, 34(2): 201–208.

[4] Sadeghian A, Huang B. (2016) Robust probabilistic principal component analysis for process modeling subject to scaled mixture Gaussian noise. Computers & Chemical Engineering, 90: 62-78.

[5] Hampshire R C, Gaites C. Peer-to-peer carsharing: Market analysis and potential growth[J]. Transportation Research Record, 2011, 2217(1): 119-126.

[6] Le Vine S, Zolfaghari A, Polak J. Carsharing: evolution, challenges and opportunities[J]. Scientific advisory group report, 2014, 22: 218-229.

[7] Shaheen S, Martin E. Assessing early market potential for carsharing in China: a case study of Beijing[J]. 2006.

[8] Degirmenci K, Breitner M H. Carsharing: A literature review and a perspective for information systems research[J]. 2014.

[9] Shaheen S A, Chan N D, Micheaux H. One-way carsharing’s evolution and operator perspectives from the Americas[J]. Transportation, 2015, 42(3): 519-536.

[10] Martin E, Shaheen S. Impacts of Car2Go on vehicle ownership, modal shift, vehicle miles traveled, and greenhouse gas emissions: an analysis of five North American Cities[J]. Transportation Sustainability Research Center, UC Berkeley, 2016, 3.

[11] Schöneburg E. Stock price prediction using neural networks: A project report[J]. Neurocomputing, 1990, 2(1): 17-27.

[12] Choi J H, Lee M K, Rhee M W. Trading S&P 500 stock index futures using a neural network[C]//Proceedings of the third annual international conference on artificial intelligence applications on wall street. 1995: 63-72.

[13] Trippi R R, DeSieno D. Trading equity index futures with a neural network[J]. Journal of Portfolio Management, 1992, 19: 27-27.

[14] Limsombunchai V. House price prediction: hedonic price model vs. artificial neural network[C]//New Zealand Agricultural and Resource Economics Society Conference. 2004: 25-26.

[15] Goodman A C, Thibodeau T G. Housing market segmentation and hedonic prediction accuracy[J].

Journal of Housing Economics, 2003, 12(3): 181-201.

[16] Kim K, Han I. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index[J]. Expert systems with Applications, 2000, 19(2): 125-132.

[17] Kim K. Financial time series forecasting using support vector machines[J]. Neurocomputing, 2003, 55(1-2): 307-319.

[18] Madge S. Predicting stock price direction using support vector machines[J]. Independent work report spring, 2015.

[19] Kara Y, Boyacioglu M A, Baykan Ö K. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange[J]. Expert systems with Applications, 2011, 38(5): 5311-5319.

[20] Hagen R. How is the international price of a particular crude determined?[J]. OPEC Review, 1994, 18(1): 127-135.

[21] Xie W, Yu L, Xu S, et al. A new method for crude oil price forecasting based on support vector machines[C]//International Conference on Computational Science. Springer, Berlin, Heidelberg, 2006: 444-451.

[22] Khashman A, Nwulu N I. Intelligent prediction of crude oil price using Support Vector Machines[C]//2011 IEEE 9th International Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2011: 165-169.

[23] Kulkarni S, Haidar I. Forecasting model for crude oil price using artificial neural networks and commodity futures prices[J]. arXiv preprint arXiv:0906.4838, 2009.

[24] Jammazi R, Aloui C. Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling[J]. Energy Economics, 2012, 34(3): 828-841.

[25] Lee C Y, Huh S Y. Forecasting long-term crude oil prices using a Bayesian model with informative priors[J]. Sustainability, 2017, 9(2): 190.

[26] Frew J, Jud G. Estimating the value of apartment buildings[J]. Journal of Real Estate Research, 2003, 25(1): 77-86.

[27] Meese R, Wallace N. House price dynamics and market fundamentals: the Parisian housing market[J]. Urban Studies, 2003, 40(5-6): 1027-1045.

[28] Selim H. Determinants of house prices in Turkey: Hedonic regression versus artificial neural network[J]. Expert systems with Applications, 2009, 36(2): 2843-2852.

[29] Liu J G, Zhang X L, Wu W P. Application of fuzzy neural network for real estate prediction[C]//International Symposium on Neural Networks. Springer, Berlin, Heidelberg, 2006: 1187-1191.

[30] Wang X, Wen J, Zhang Y, et al. Real estate price forecasting based on SVM optimized by PSO[J]. Optik-International Journal for Light and Electron Optics, 2014, 125(3): 1439-1443.

[31] Gu J, Zhu M, Jiang L. Housing price forecasting based on genetic algorithm and support vector machine[J]. Expert Systems with Applications, 2011, 38(4): 3383-3386.

[32] Lus C A, Torgo L, Gama J. Regression using classification algorithms[M]. IOS Press, 1997.

[33] Bibi S, Tsoumakas G, Stamelos I, et al. Regression via Classification applied on software defect estimation[J]. Expert Systems with Applications, 2008, 34(3):2091-2101.

[34] Janssen F, Johannes Fürnkranz. Heuristic Rule-Based Regression via Dynamic Reduction to Classification[C]// IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011. DBLP, 2011.

[35] Ahmad A, Halawani S M, Albidewi I A. Novel ensemble methods for regression via classification problems[J]. Expert Systems with Applications, 2012, 39(7):6396-6401.

[36] Wu J, Xiong H, Chen J, et al. A Generalization of Proximity Functions for K-Means[C]// icdm. IEEE Computer Society, 2007.

[37] Liu F T, Kai M T, Zhou Z H. Isolation Forest[C]// 2008 Eighth IEEE International Conference on Data Mining. 2009.

[38] Hsu C W, Lin C J. A Comparison of Methods for Multiclass Support Vector Machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2):415-425.

[39] Weston J, Watkins C. Multi-class support vector machines[R]. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May, 1998.

[40] Schwenker F. Hierarchical support vector machines for multi-class pattern recognition[J]. Knowledge-based intelligent engineering systems and applied technologies KES 2000, 2000.

[41] Bottou, Léon, Cortes C, Denker J S, et al. Comparison of classifier methods: a case study in handwritten digit recognition[C]// International Conference on Pattern Recognition. IEEE Computer Society, 1994.

[42] Krebel U. Pairwise classification and support vector machines[M]// Advances in kernel methods. MIT Press, 1999.

[43] Mhaskar H, Liao Q, Poggio T. When and why are deep networks better than shallow ones?[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[44] Cutler A, Cutler D R, Stevens J R. Random Forests[J]. Machine Learning, 2004, 45(1):157-176.

[45] Liu F T, Ting K M, Yu Y, et al. Spectrum of Variable-Random Trees[J]. Journal of Artificial Intelligence Research, 2008, 32(1):355-384.

[46] Schwenker F. Ensemble Methods: Foundations and Algorithms [Book Review][J]. IEEE Computational Intelligence Magazine, 2013, 8(1):77-79.

[47] Krogh A, Vedelsby J. Neural Network Ensembles, Cross Validation, and Active Learning[C]// International Conference on Neural Information Processing Systems. MIT Press, 1995.

[48] Zhou Z H. Ensemble Methods - Foundations and Algorithms[M]. Taylor & Francis, 2012.

[49] Breiman L. Bagging Predictors[J]. Machine Learning, 1996, 24(2):123-140.

[50] Johnson R W. An Introduction to the Bootstrap[J]. Teaching Statistics, 2001.

[51] Freund Y, Schapire R E. A desicion-theoretic generalization of on-line learning and an application to boosting[J]. 1995.

[52] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8):832-844.

[53] Kolen J F, Pollack J B. Back propagation is sensitive to initial conditions[C]// Conference on Advances in Neural Information Processing Systems. Morgan Kaufmann Publishers Inc. 1990.

[54] Dietterich T G, Bakiri G. Solving Multiclass Learning Problems via Error-Correcting Output Codes[J]. Journal of Artificial Intelligence Research, 1995, 2(1):263--286.

[55] Breiman L. Randomizing Outputs to Increase Prediction Accuracy[J]. Machine Learning, 2000, 40(3):229-242.

[56] Zhou Z H, Feng J. Deep Forest: Towards An Alternative to Deep Neural Networks[J]. 2017.

[57] Zhang Y, ZHOU, ZhiHua. Cost-Sensitive Face Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 32(10):1758-1769.

[58] Correa Bahnsen A, Aouada D, Ottersten, Björn. Example-Dependent Cost-Sensitive Decision Trees[J]. Expert Systems with Applications, 2015, 42(19):6609-6619.

[59] Liu Z, Ma C, Gao C, et al. Cost-sensitive collaborative representation based classification via probability estimation with addressing the class imbalance[J]. Multimedia Tools and Applications, 2017.

[60] Egon Kocjan, Igor Kononenko. Regression as cost-sensitive classification. PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUZBA 2009, 2009.

[61] Method N. Tukey's range test[J]. 2015.

[62] Gama J, Pinto C. Discretization from data streams: applications to histograms and data mining[C]//Proceedings of the 2006 ACM symposium on Applied computing. ACM, 2006: 662-667.

[63] Yang Y, Webb G I. Discretization for naive-Bayes learning: managing discretization bias and variance[J]. Machine learning, 2009, 74(1): 39-74.

[64] Torgo L, Gama J. Search-based class discretization[C]//European Conference on Machine Learning. Springer, Berlin, Heidelberg, 1997: 266-273.

[65] Habbema J D F. Cases of Doubt in Allocation Problems[J]. Biometrika, 1974, 61(2):313-324.

[66] Elkan C. The foundations of cost-sensitive learning[C]//International joint conference on artificial intelligence. LAWRENCE ERLBAUM ASSOCIATES LTD, 2001, 17(1): 973-978.

[67] Ma Y, Ding X. Face Detection Based on Cost-Sensitive Support Vector Machines[J]. Lecture Notes in Computer Science, 2002, 2388:260-267.

[68] Lee W, Fan W, Stolfo S J, et al. Cost-Sensitive Modeling for Intrusion Detection[J]. Journal of

Computer Security, 2002, 10(1-2):5-22.

[69] Viaene S, Van Gheel D, Ayuso M, et al. Cost-Sensitive design of claim fraud screens[C]// International Conference on Advances in Data Mining: Applications in Image Mining, Medicine and Biotechnology, Management and Environmental Control, and Telecommunications. Springer-Verlag, 2004:78-87.

[70] Friis J, Williams N, Zadrozny B. Cost-Sensitive Knowledge Discovery: A Case Study[J]. 2000.

[71] Lee Y, Lin Y, Wahba G. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data[J]. Journal of the American Statistical Association, 2004, 99(465): 67-81.

[72] Mez Hidalgo J, Pez M M, Sanz E P. Combining text and heuristics for cost-sensitive spam filtering[C]// The Workshop on Learning Language in Logic and the, Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2000:99-102.

[73] Núñez M. Economic Induction: A Case Study[C]//EWSL. 1988, 88: 139-145.

[74] Bahnsen A C, Aouada D, Ottersten B. Example-dependent cost-sensitive logistic regression for credit scoring[C]//2014 13th International Conference on Machine Learning and Applications. IEEE, 2014: 263-269.

[75] Sahin Y, Bulkan S, Duman E. A cost-sensitive decision tree approach for fraud detection[J] Expert Systems with Applications, 2013, 40(15): 5916-5923.

[76] Wu G, Chang E Y. KBA: Kernel boundary alignment considering unbalanced data distribution[J]. IEEE Transactions on knowledge and data engineering, 2005, 17(6): 786-795.

[77] Masnadi-Shirazi H, Vasconcelos N, Iranmehr A. Cost-sensitive support vector machines[J]. arXiv preprint arXiv:12120975, 2012,

[78] Li Y F, Kwok J T, Zhou Z H. Cost-Sensitive Semi-Supervised Support Vector Machine[C]// Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, Usa, July. 2010.

[79] Chawla N V. Data mining for unbalanced datasets: An overview[M]. Data mining and knowledge discovery handbook. Springer. 2009: 875-886.

[80] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[81] Xia Y, Liu C, Liu N. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending[J]. Electronic Commerce Research and Applications, 2017, 24: 30-49.

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

[83] Han J, Pei J, Kamber M. Data mining: concepts and techniques[M]. Elsevier, 2011.

[84] Liu H, Hussain F, Tan C L, et al. Discretization: An enabling technique[J]. Data mining and knowledge discovery, 2002, 6(4): 393-423.

[85] Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features[M]//Machine Learning Proceedings 1995. Morgan Kaufmann, 1995: 194-202.

[86] Chmielewski M R, Grzymala-Busse J W. Global discretization of continuous attributes as preprocessing for machine learning[J]. International journal of approximate reasoning, 1996, 15(4): 319-331.

[87] Tay F E H, Shen L. A modified chi2 algorithm for discretization[J]. IEEE Transactions on knowledge and data engineering, 2002, 14(3): 666-670.

[88] Kohavi R, Sahami M. Error-based and entropy-based discretization of continuous features[C]//KDD. 1996: 114-119.

[89] Bakar A A, Othman Z A, Shuib N L M. Building a new taxonomy for data discretization techniques[C]//2009 2nd Conference on Data Mining and Optimization. IEEE, 2009: 132-140.

[90] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Cognitive modeling, 1988, 5(3): 1.

[91] Pineda F J. Generalization of back-propagation to recurrent neural networks[J]. Physical Review Letters, 1987, 59(19):2229-2232.

[92] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

致谢

本论文在导师刘振丙老师的精心指导下已顺利完成。在攻读硕士学位的三年里，从选题方向、资料收集、深入研究、论文撰写等各个环节，无不渗透刘老师的悉心教导和无微关怀。他严谨的治学，渊博的学识，以及对科学研究执着与热忱，使得我本人在今后的人身发展有了更充实的沉淀和更明确的目标，以及了解了自身的不足和改进的方法。面对较深的理论及前沿技术，刘老师多次提供参加学术会议和相关培训的机会，以及在相关教材和器材给予的资助，还支持我获得学校公派留学名额，并积极帮助我联系联合培养导师，这些都为我完成本论文的研究和写作起到决定性作用，致此我表示由衷的感谢！

感谢桂林电子科技大学的杨辉华教授、蓝如师老师、王子民老师、潘细朋老师，在学术氛围、理论研究以及论文写作方面给予的指导和帮助；感谢师兄师姐蒋淑洁、何其佳、高春洋、徐涛给予的帮助与指导；感谢同门卢勇全、方旭升对我的帮助和支持；感谢师弟师妹姬欢欢、李泽亚、王文颢、李鑫龙、李蔚蔚给予的支持；感谢实验室这个大家庭的关心和帮助，让我两年的生活如此温馨、如此多姿多彩。

感谢桂林电子科技大学研究生院的段雪峰副院长、于立娟老师对我出国留学事宜的帮助和支持。感谢新加坡南洋理工大学张杰教授的邀请；感谢新加坡国立大学曹志广老师，不仅在学术上给予悉心指导，还在为人和生活上为我竖立了标杆；感谢师兄师姐张露、武垚欣、韩星烁、Chan Jyh Huah、Jim Cherian 的帮助与指导；感谢在新加坡遇到的所有人，这一年的经历让我不仅提升了科研能力，还增长了见识、体验到了不同的文化和生活。

感谢论文评审专家和论文答辩委员会专家。因为诸位的认真评价、由衷建议以及严谨治学为论文的完善付出了辛劳的汗水，也使我了解到论文中存在的不足以及需要改进的地方；感谢国家自然科学基金、广西省自然科学基金、桂林电子科技大学研究生出国（境）研究奖学金项目、桂林电子科技大学研究生优秀学位论文培育项目、桂林电子科技大学研究生创新项目等对本论文的资助。

最后，再一次感谢我的导师刘振丙老师！感谢我的母校桂林电子科技大学！祝各位老师身体健康，工作顺利！愿我的母校桂林电子科技大学教育事业蓬勃发展！

作者在攻读硕士期间的主要研究成果

论文

[1] **Chao Ma**, Zhenbing Liu, Zhiguang Cao, Lu Zhang, Jie Zhang. Using Cost-sensitive Deep Forest to Predict the Price for a P2P Car-sharing Service in China[J]. Pattern Recognition. (Submitted)

[2] Zhenbing Liu, **Chao Ma**, Chunyang Gao, Huihua Yang, Tao Xu, Rushi Lan, Xiaonan Luo. Cost-sensitive collaborative representation based classification via probability estimation with addressing the class imbalance[J]. Multimedia Tools and Applications, 2018, 77(9):10835-10851.

[3] Zhenbing Liu, Tao Xu, **Chao Ma**, Chunyang Gao, and Huihua Yang. T-test based Alzheimer's disease diagnosis with multi-feature in MRIs[J]. Multimedia Tools and Applications, 2018, 77(22): 29687-29703.

[4] 刘振丙, 姬欢欢, **马超**, 蒋淑洁, 杨辉华. 基于 Gabor 优化协同表示的近红外药品鉴别[C]. 2018 药物质量分析与过程控制分会首届学术报告会.

科研项目

[1] 类别不平衡问题中代价敏感的协同表示算法研究, 桂林电子科技大学研究生科研创新项目, 2018 年 4 月-2020 年 4 月, 在研, 负责人.