

价格预测中代价敏感的机器学习算法及优化

导师：刘振丙 研究员

答辩人：马超

学号：1608202004

目录

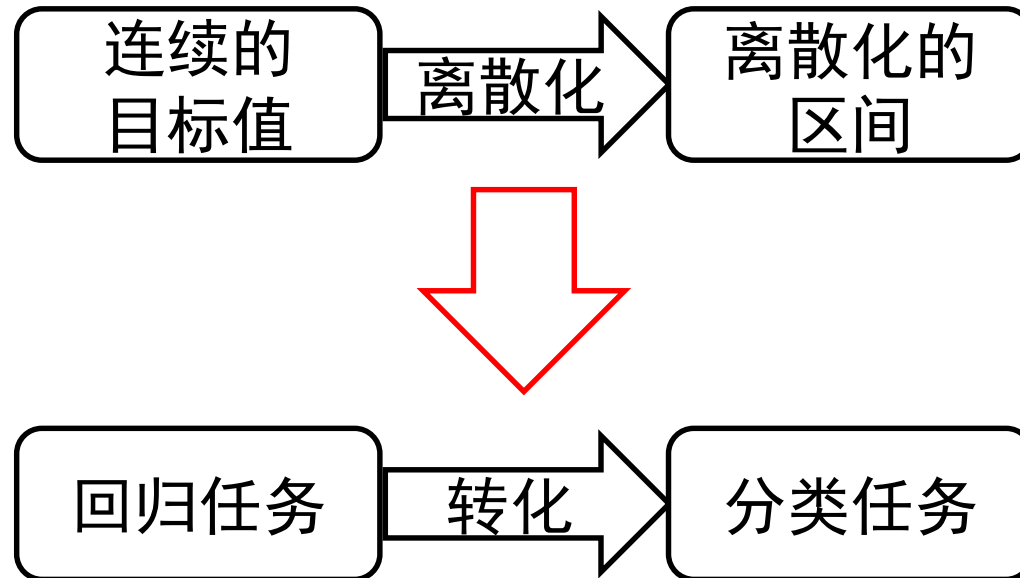
- 一. 研究背景与意义
- 二. 改进的K均值算法
- 三. 深度森林的自适应深度研究
- 四. 代价敏感深度森林
- 五. 总结

研究意义

- P2P共享
 - Person to Person
 - 更多选择
- 痛点
 - 新手价格判断
 - 价格预测系统

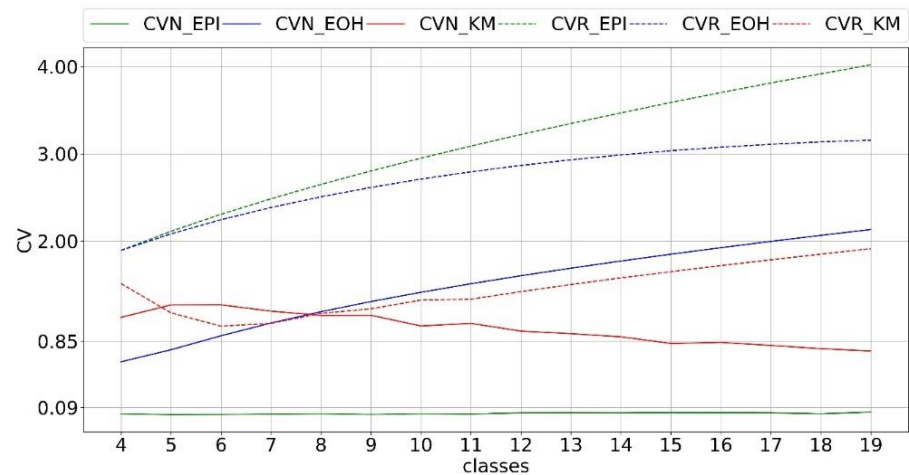
现有方法

- 通过分类算法解决回归问题
 - 不适合直接应用



问题

- 目标值离散化方法
 - 等概率区间、等宽度区间
 - K均值聚类 （对不能忽略的离群值敏感）
- 均匀的区域宽度
 - 稳定的预测
- 均匀的样本数量
 - 避免数据不平衡



改进的K均值算法

- 孤立森林
 - 将所有的样本彼此孤立
- 离群值
 - 数量较少
 - 特征值区别很大
- 结果
 - 离群值更容易划分出来

改进的K均值算法

- K均值的均匀作用
— 样本数量

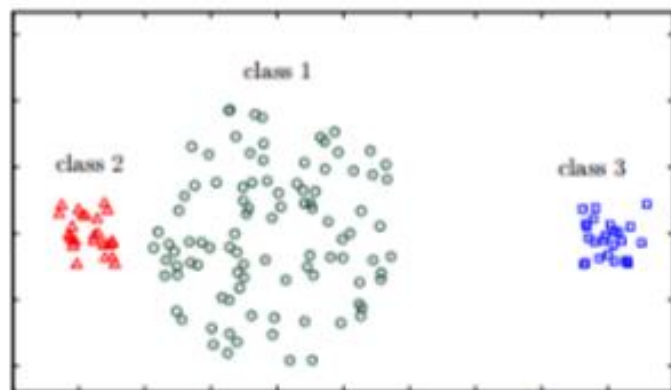
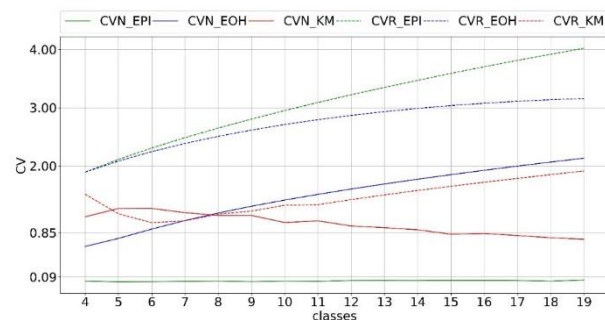


Figure 1: Clusters before K-means Clustering.

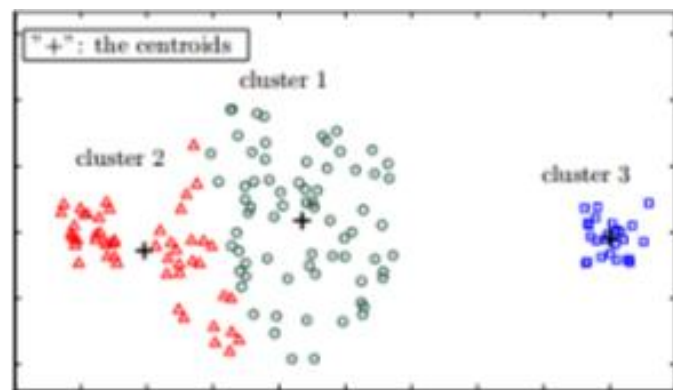


Figure 2: Clusters after K-means Clustering.

- 真实标签

K均值聚类标签

改进的K均值算法

- K均值的均匀作用
— 样本数量

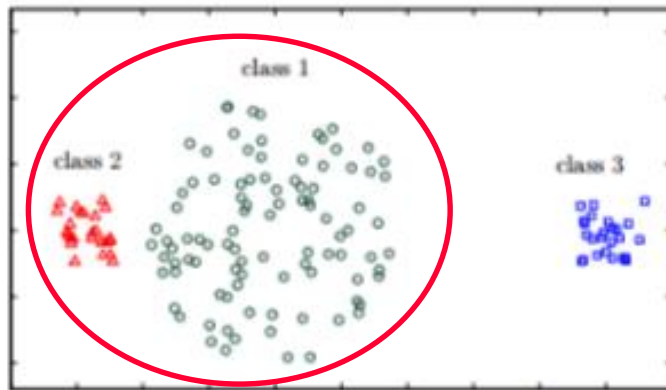
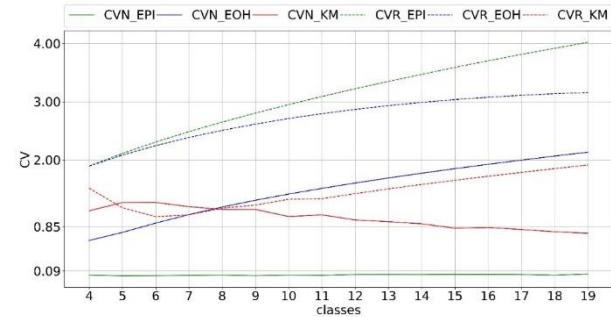


Figure 1: Clusters before K-means Clustering.

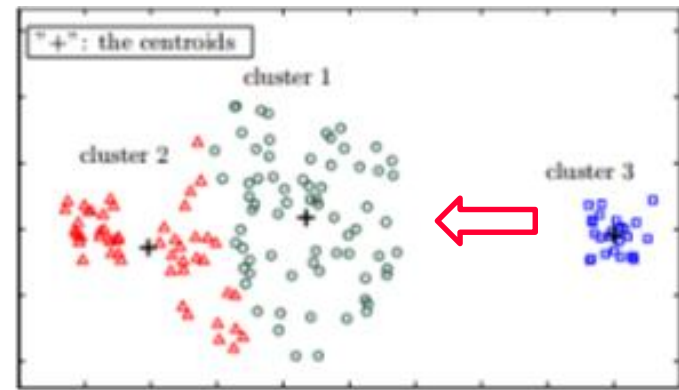


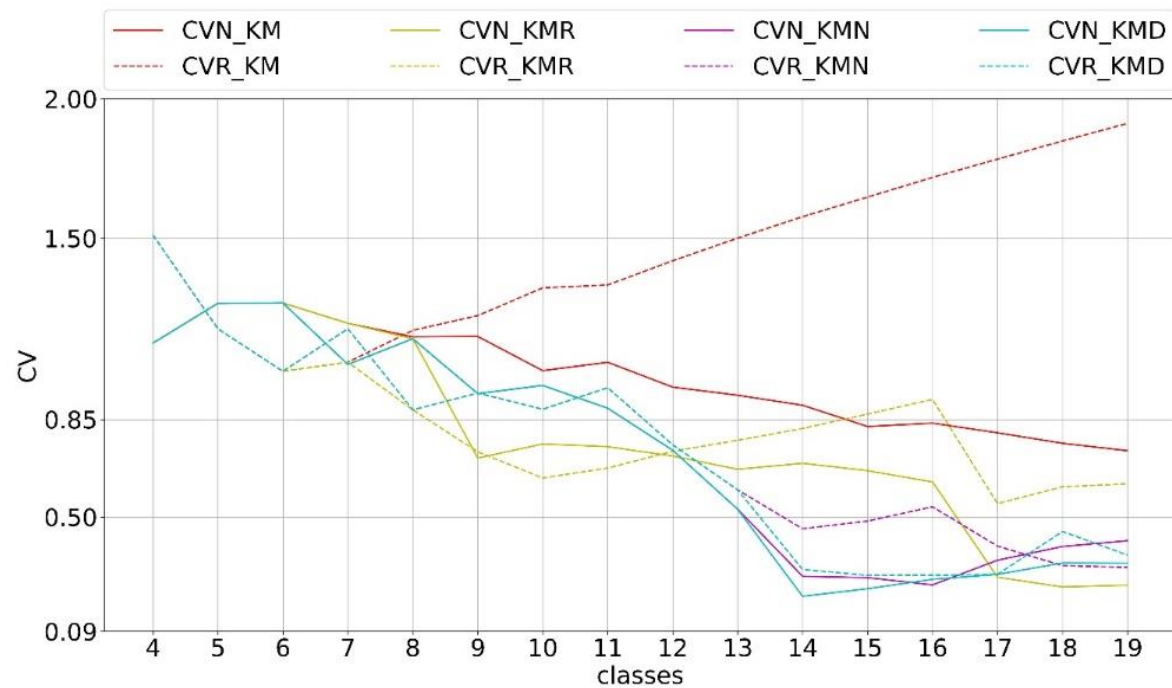
Figure 2: Clusters after K-means Clustering.

- 真实标签

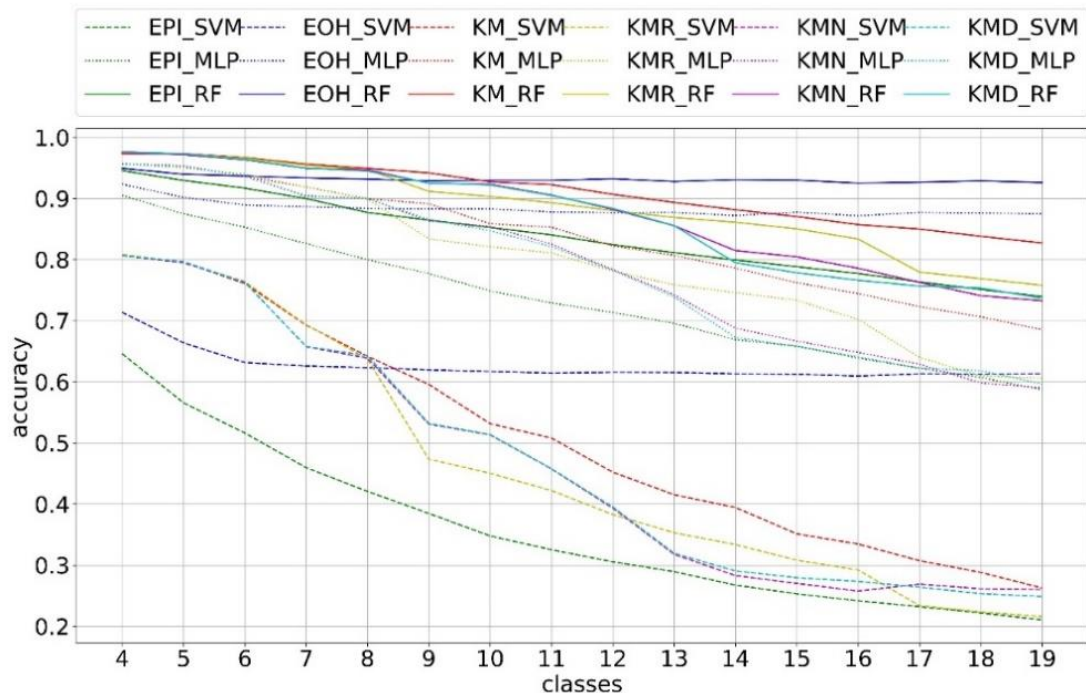
K均值聚类标签

实验结果及分析

- 不同的最离群区间
 - 三种不同的改进K均值算法




问题

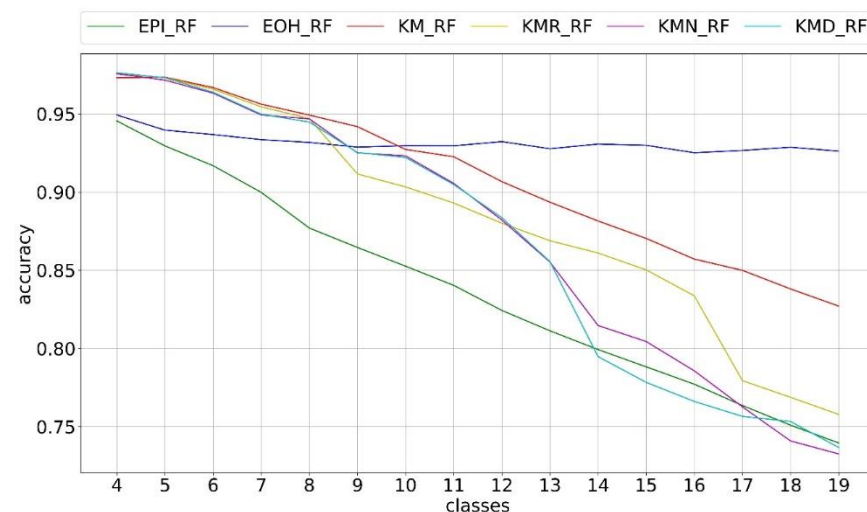
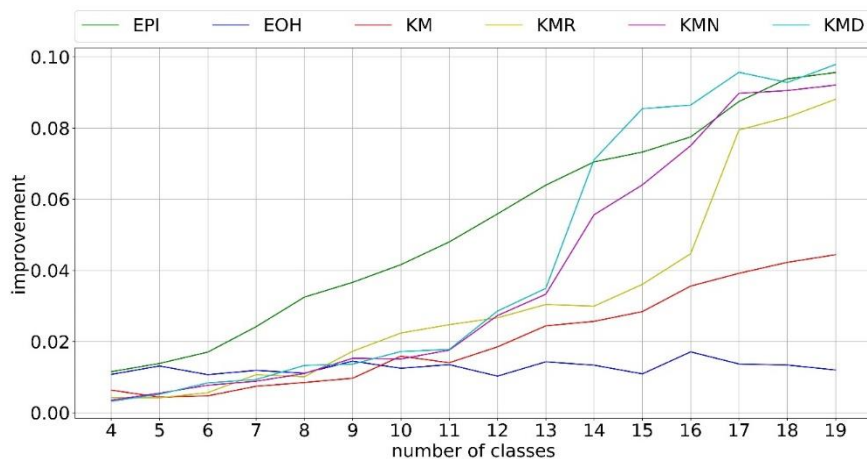


- 支持向量机、多层感知器、**随机森林**
- 区间数量越多，分类性能越差

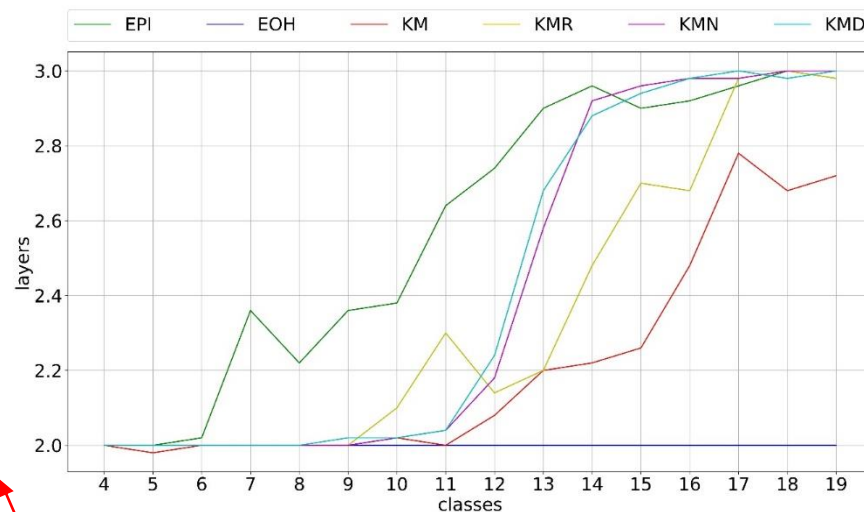
深度森林

- 深度森林
 - 深度学习+集成学习
 - 深度自适应
 - 深度学习
 - 模型复杂度
 - 深度比宽度更重要
 - 集成学习
 - 合适的基分类器性能
 - 足够的多样性
- 

实验分析



随机森林



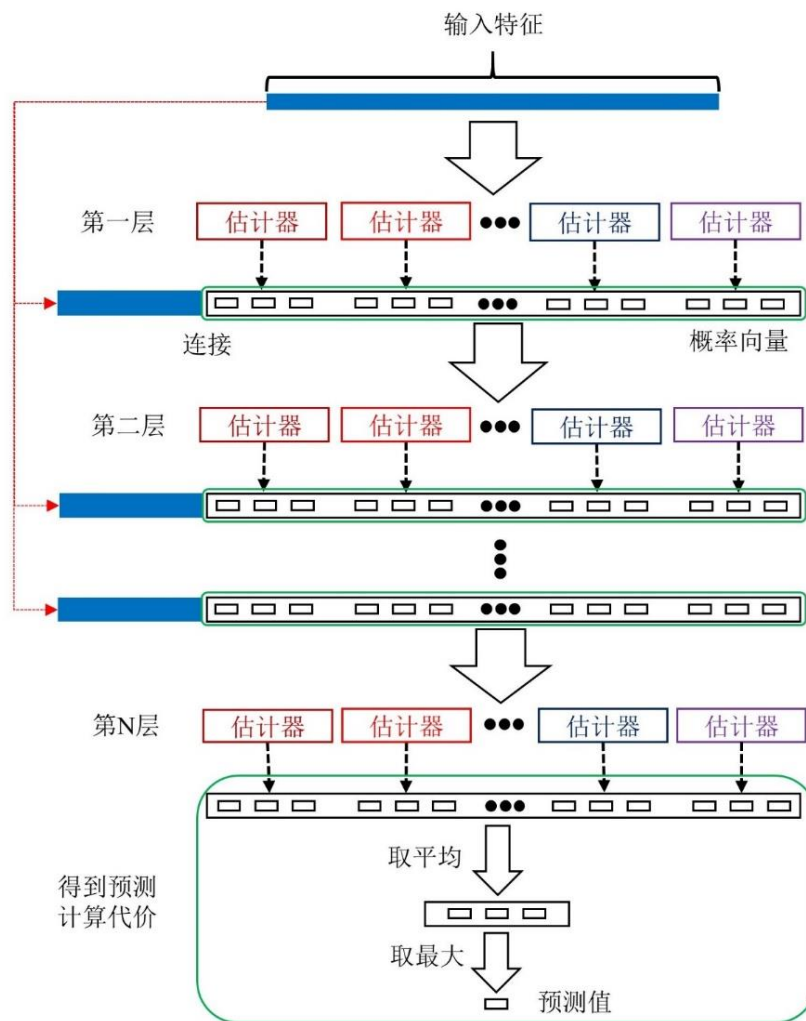
深度森林的层数

深度森林的改进

问题

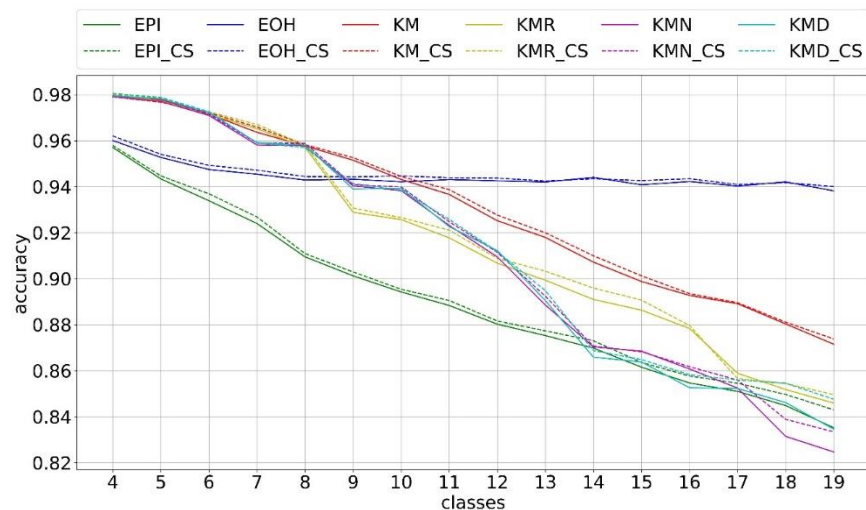
- 深度森林缺乏代价敏感框架
- 通过分类算法解决回归问题
 - $[100, 199]$ $[200, 299]$, $[500, 599]$
 - 代价敏感问题

代价敏感深度森林

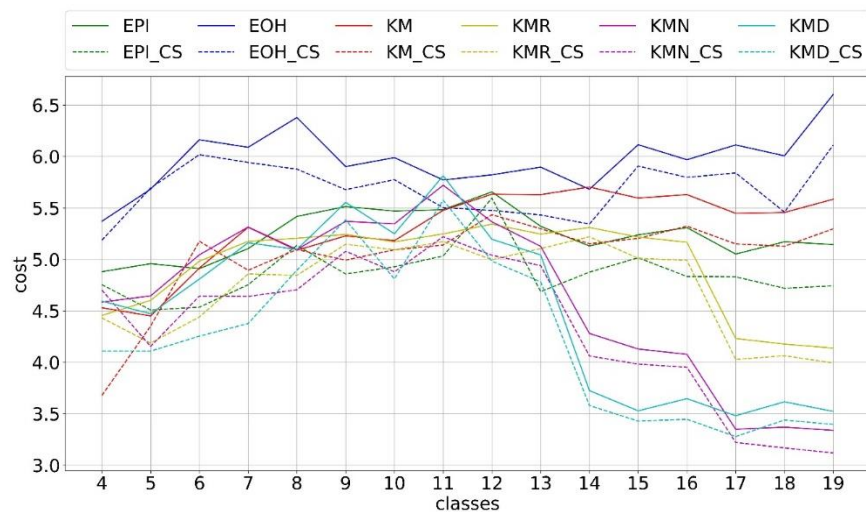


实验结果及分析

- 准确度



- 代价
- 5.6%



主要贡献

- 改进了K均值算法，降低了离群值对K均值的影响
- 发现了深度森林的自适应深度有助于通过分类算法解决回归问题
- 将代价敏感框架引入到深度森林算法

主要研究成果

- 论文

- Pattern Recognition

- 二区 已投稿

- Multimedia Tools and Applications

- 四区 已发表

- 科研项目

- 研究生科研创新项目

- 负责人

谢 谢