

文章编号: 1000-6788(2009)12-0134-07

基于 Bagging 与决策树算法的在线拍卖成交价格预测模型

刘 洋, 冯玉强, 邵 真

(哈尔滨工业大学 管理科学与工程系, 哈尔滨 150001)

摘 要 通过分析在线拍卖出价特点, 利用决策树和 Bagging 算法建立了一种全新的在线拍卖成交价格预测模型. 作者编写程序收集淘宝网在线拍卖交易数据 3310 条, 对有效出价记录 8275 条. 数据分析表明, 如不考虑未成交商品, 则有 40.4% 的交易可以利用出价次数精确计算最终成交价格. 如将未成交商品视为成交价格为 0, 该比例可提高为 79.55%. 据此发现, 作者通过预测出价次数间接对成交价格进行预测. 实验证明, 模型明显优于平均值预测, 并有 21.7% 的预测结果完全准确. 通过与 Heijst 发表于《Decision Support Systems》上的研究进行对比, 结果表明预测模型在样本需求量、运算时间, 及完全准确预测率上有明显优势. 由于模型训练时间仅为数秒, 为建立实时在线拍卖成交价格预测决策支持系统奠定了基础.

关键词 在线拍卖; 预测; Bagging 算法; 决策树

中图分类号 C931.9

文献标志码 A

Online auction final price forecasting model based on Bagging and decision tree

LIU Yang, FENG Yu-qiang, SHAO Zhen

(School of Management, Harbin Institute of Technology, Harbin 150001, China)

Abstract By analysing bidders' behaviors, the author proposed a new model which is based on the Bagging arithmetic and decision tree for predicting final prices in online auctions. The author collected 3310 transaction data and corresponding 8275 bids from Taobao. Data analysis shows that the final prices of 40.4% transactions can be calculated by using the times of bids. If the transactions which have no bid are included, the percentage can increase to 79.55%. Instead of predicting the final price directly, the author chose to predict times of bids first and then used it to calculate the final price. The experiment proves that the model substantially outperforms the naive method of predicting the category mean price, and 21.7% of predicted results are exactly equal to the real ones. The author also compared the model with Heijst's research which was published in Decision Support Systems. The result shows that the model is better in required training sample size, calculating time and the percentage of accurate prediction. For the training time is only a few seconds, our research can lay the foundation for developing real-time decision support systems.

Keywords online auction; prediction; Bagging arithmetic; decision tree

1 引言

近年来, 在线拍卖这一新兴的电子商务模式得到了飞速的发展. 与传统拍卖中拍卖商品多为收藏品等不

收稿日期: 2008-10-13

资助项目: 国家自然科学基金 (70572023); 黑龙江省自然科学基金 (zd200803-01)

作者简介: 刘洋 (1982-), 男, 哈尔滨, 博士研究生, 研究方向为电子商务, 数据挖掘, 企业信息化.

同, 在线拍卖交易商品种类繁多, 从电脑、服装到打折机票、优惠券等, 几乎无所不包含在内. 经常会发现多笔拍卖同时销售几乎相同的商品. 由于在线拍卖持续时间更长, 如果投标失败, 无疑会浪费消费者更多的时间. 那么如何在众多同时进行的拍卖中进行选择? 自己的估价是否会赢得拍卖? 这些无疑是在线拍卖投标人必须要面对的问题. 如果能在一定程度上对拍卖商品的成交价格进行预测, 这些问题将会得到很好的解决.

关于拍卖最终成交价格的研究很早就吸引了学者们的注意力. 早期对于成交价格的研究通常使用多元回归的方法. 虽然多元回归在一定程度上可以对最终成交价格进行预测, 但是由于多元回归分析基于线性的假设, 在很大程度上影响了预测的准确程度, 所以这些研究主要目的还是集中于识别在线拍卖最终成交价格影响因素上. 如 Wood^[1] 对 eBay 上发生的 7362 宗收藏品交易结果进行研究, 发现周末效应、图片效应、声望效应、拍卖时间效应是影响拍卖最终结果的四个显著因素; Bryan^[2] 等探讨了卖家信誉、投标人数等因素对成交价格的影响; Kauffman^[3] 研究了“托”对于成交价格的影响等等. 此外还有学者利用函数性数据分析 (FDA) 方法建立动态预测模型^[4-5].

近年来, 开始有少数学者通过数据挖掘技术对成交价格进行预测, 并对各种算法进行比较, 试图找到最有效的算法^[6]. 目前, 进行预测多使用如神经网络^[7]、决策树^[8]等算法, 并取得了很好的效果. 但在线拍卖最终成交价格属于连续变量, 训练过程所需运算时间较长, 往往需要几小时甚至几天, 无法达到实时性要求, 限制了这些方法在现实中推广使用. 因此, 如何更进一步缩短模型运行时间成为了当前在线拍卖价格预测研究领域的热点.

不难看出, 造成算法运行时间过长的根本原因在于预测变量为连续变量. 因此, 对预测变量进行离散化是解决该问题的关键. 显然, 出价次数与成交价格之间存在着高度的相关性, 同时又满足属于离散变量这一条件. 因此, 如果可以找出成交价格与出价次数之间的关系, 就可以将连续问题转换为离散问题, 进而减少运算时间.

按照以上思路, 在第二节中, 作者抓取淘宝网在线拍卖真实交易数据, 并对其进行分析, 找出成交价格与出价次数之间的关系; 在第三节中, 利用两变量之间的关系对连续问题离散化, 并利用 Bagging 算法与决策树算法建立模型; 第四节中, 利用淘宝网真实交易数据进行实验, 并对实验结果进行评价, 同时与相关研究进行详细比较; 文章最后给出研究结论.

2 数据分析

2.1 数据获取

本文研究数据来自淘宝网. 淘宝网采用的拍卖机制为在拍卖过程中, 买家在拍卖结束之前分别对价格和购买数量投标, 商品以全部领先者中出价最低者的价格成交, 优先满足价格最高者. 当出现出价相同时, 先出价者获得商品. 例如, 卖家出售某商品 2 件, 共 4 人投标, 各购买 1 件. 其中 A 出价 15 元; B 出 14 元; C 出 14 元, 但出价时间晚于 B; D 出价 13 元. 则最终成交形式为 A、B 各拍得商品 1 件, 成交价格为 14 元, C、D 投标失败.

本文作者利用 C# 开发了一个针对淘宝网在线拍卖交易的数据收集 Agent 程序. 该程序的基本原理是通过获取淘宝网上已经结束的拍卖商品所在的 HTML 页面, 利用程序抽取出 HTML 页面中包含的详细信息, 将其存储入 SQL Server 2000 数据库中. 关于数据抓取更多信息可参阅文献 [9].

抓取数据包含拍卖商品详细信息及出价历史记录. 其中, 拍卖商品详细信息记录了拍卖本身的信息, 其中包括卖家 ID、卖家名称、卖家信誉、卖家好评率、产品 ID、产品名称、开始时间、结束时间、起始价格、加价幅度、成交价格、产品类型 (全新、二手、闲置)、是否提供发票、是否提供保修、是否提供先行赔付、是否提供 7 天无理由退货换货这 16 个属性. 出价记录则详细记录了拍卖过程中投标人的参与过程.

2.2 数据分析

利用该程序, 作者收集了淘宝网女鞋类商品中结束时间在 2008 年 5 月 26 日至 6 月 20 日之间的拍卖数据. 除去少数在拍卖开始不久即被卖家关闭的交易, 共获得拍卖商品详细信息数据 3310 条, 及与之相对应的出价记录 8275 条. 获取数据分布特点如表 1 所示.

表 1 数据各属性分布特点

成交情况		商品类别		发票		保修	
成交	1135	全新	2469	提供	51	提供	199
		闲置	303				
未成交	2175	二手	538	无	3259	无	3111
合计	3310		3310		3310		3310
运费		先行赔付		7 天无条件退换货		出售数量	
买家承担	0	提供	547	提供	196	1	2670
卖家承担	3310	无	2763	无	3114	>1	640
合计	3310		3310		3310		3310
起始价格		最小加价幅度		卖家信誉		卖家好评率	
均值	86.97		5.37		415.09		93.57
方差	279.86		14.16		1151.92		22.98

在全部的 3310 条商品信息中, 成交商品记录 1135 条, 即只有这 1135 件商品有出价记录. 通过对数据进行分析我们发现, 很多的消费者按以下策略进行投标. 设卖家规定起始价格为 p , 最小加价幅度为 ε . 当消费者为第一个出价者, 其出价为 p ; 当消费者不为第一个出价者, 设前一个人出价为 q , 则该消费者出价 $q + \varepsilon$, 即以卖家规定的最小加价幅度进行加价. 为了分析其所占比例, 作者对数据进行统计, 结果如表 2 所示:

表 2 出价记录数据分析

按出价记录统计		按成交商品统计	
以最小加价幅度加价	5526	全部以最小加价幅度加价	458
其他	2479	其他	667
合计	8275	合计	1135

从数据中可以看出, 在全部的 8275 条出价记录中, 有 66.8% 的出价记录是以最小加价幅度进行加价的. 在全部成交的 1135 个商品中, 某一商品所有的出价全按照最小加价幅度进行加价的为 458 个, 占全部商品的 40.4%. 如果将未成交的商品视为出价 0 次, 成交价格为 0, 则意味着有 2633 条交易记录 (未成交的 2175 加 458) 的拍卖成交价格可以通过出价次数进行准确计算, 占全部交易记录的比例高达 79.55%.

3 模型建立

3.1 预测方法

从上一部分的分析中我们可以看出, 绝大部分的拍卖成交价格可以通过预测拍卖出价次数以达到对成交价格的间接预测. 计算公式如下:

$$\begin{cases} \text{价格} = 0, & \text{出价次数} = 0 \\ \text{价格} = \text{起始价格} + (\text{出价次数} - 1) \times \text{最小加价}, & \text{出价次数} \neq 0 \end{cases} \quad (1)$$

由于出价次数为离散属性, 对其进行分类, 可以显著减少运算时间、提高精度. 正是基于此, 本文采用分类的方法建立预测模型, 以出价次数作为类标号, 之后利用分类得出的出价次数计算价格, 而不再直接对最终成交价格进行预测.

在众多的分类算法中, 决策树因其具有优良的性能, 一直作为分类算法的首选. 在本文研究中, 也将使用决策树方法. 同时, 由于所使用的数据在一定程度上属于异质商品数据, 存在较大的变化性, 加之分类属性数量较少, 会在一定程度上影响分类的效果. 因此, 在本文研究中进一步通过 Bagging 算法对决策树算法加以集成, 对其进行改进.

3.2 决策树

决策树是一个类似于流程图的树结构, 由决策节点和叶节点组成, 树最顶层的决策节点称为根节点. 每一个决策节点含有若干个判定条件 C , 对应着若干个分枝. 每个分枝代表一个测试输出, 不包含分枝的节点是叶节点, 代表类或类分布. 一棵典型的决策树如图 1 所示:

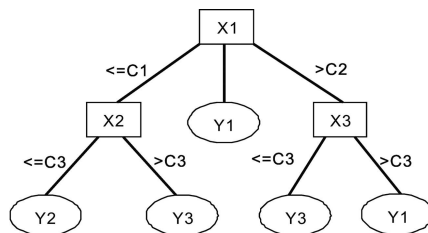


图 1 决策树示意图

从决策树的根节点开始, 沿着决策树从上到下的遍历过程中, 每一个节点都可能会遇到一个测试, 对应着每个节点上判定条件的不同, 测试输出导致不同的分枝, 最后会达到一个叶节点. 这个过程就是利用决策树进行分类的过程, 即利用若干变量来判断所属的类别.

决策树的建立分两部分: 第一部分是决策树的生成过程, 第二部分是决策树的剪枝过程. 在生成阶段是利用样本数据对决策树进行训练, 直到精度无法提高或者已经达到用户满意时停止. 当决策树创建时, 由于数据中的噪声和孤立点, 许多分枝反映的是训练数据中的异常. 剪枝方法用于处理这种过分适应数据问题. 通常, 这种方法使用统计度量, 剪去最不可靠的分枝, 这将导致较快分类, 提高树独立测试数据正确分类的能力.

使用决策树具有以下好处. 首先, 利用决策树方法可以自行判断哪些属性会在类别判断中起作用; 其次, 决策树不再像回归那样受到线性问题的影响.

3.3 Bagging 算法

Bagging 算法是由 Breiman^[10] 于 1996 年提出. 其理论依据源于 Valiant 提出的计算学习理论——PAC(Probably approximately correct) 学习模型, 是一种提升任意给定学习算法精度的普遍方法^[11]. 其基本思想是将弱学习算法“提升”为一个具有任意精度的强学习算法. 该算法是近年来机器学习领域的主流算法之一, 在很多应用领域都表现出了优良的特性.

Bagging 不需要弱学习算法的先验知识, 只要给定足够多的数据, 通过寻找比随机猜测稍好的弱分类器, 就能够得到一个比较好的最终强分类器, 而不是一开始就试图设计一个分类比较精确的算法; 同时它还有理论支持, 只要有足够多的数据以及弱分类器就能够达到任意的预测精度.

与 Bagging 算法类似的还有 Boosting 算法, 也是近年来的研究热点. 与 Boosting 相比, 稳定性是 Bagging 能否提高预测准确率的关键因素^[10-12], Bagging 算法对如决策树等不稳定的学习算法能提高预测的准确度, 而对稳定的学习算法效果不明显. 此外, Bagging 的各预测函数可以并行生成, 而 Boosting 的只能顺序生成. 因此, Bagging 算法可以节省大量的训练时间. 这也是本文选择使用 Bagging 算法而不使用 Boosting 算法的原因.

Bagging 算法的核心思想是给定训练集 $S = ((x_1, y_1), (x_2, y_2), \dots, (x_N, y_N))$ 和弱学习算法, 对该算法进行 T 次调用, 每次调用时只用训练集 S 中的某个子集作为当前训练集, 每一个训练例在某轮训练集中可以多次或根本不出现. 经 T 次调用后, 可得到 T 个不同的分类器 h_1, h_2, \dots, h_T , 当对一个实例进行分类时, 分别调用这 T 个分类器, 得到 T 个结果. 最后对分类问题把这 T 个分类结果中出现次数最多的类赋予测试实例. 算法逻辑如下所示:

输入: 训练集 S , 训练循环次数 T

循环 $t = 1, 2, \dots, T$

从训练集 S 中进行第 t 次抽样 (有放回抽样)

用弱学习算法进行训练, 得到弱的假设模型 h_t ;

输出: 对未知样本 X 分类时, 每个模型 h_t 都得出一个分类, 得票最高的为未知样本 X 的类别.

在本文中所使用的弱学习算法即为上一部分所述的决策树.

4 实验及结果分析

为对我们的模型进行测试, 遵照上节所述模型, 本文的实验分两阶段进行. 第一阶段预测商品出价次数

是否为零, 即预测商品最终是否会成交; 第二阶段预测该商品出价次数.

4.1 商品是否成交预测实验

本部分研究首先对前文所述的 3310 条“女鞋”类数据进行随机分组. 其中, 训练样本 2649 条, 占全部数据的 80%; 测试样本 661 条, 占全部数据的 20%. 之后, 将商品对应的出价次数加入到原始数据中, 并对数据进行处理, 将出价次数划分为“出价次数为 0”与“出价次数非 0”两大类, 作为类标号. 最后, 利用本文建立的 Bagging 算法集成决策树模型进行训练及预测, 结果如表 3 所示:

表 3 模型测试结果表

	成交	未成交	合计
样本量	221	440	661
正确预测	181	408	589
错误预测	40	32	72
测准率	81.9%	92.7%	89%

这一结果可以说明, 本文所建立的模型在预测商品是否会最终成交时会取得非常好的效果. 尤其是在预测商品未成交时, 效果尤为明显.

4.2 商品成交价格预测实验

这一部分研究所用数据为前文所述的 1135 条成交数据. 按照随机抽样原则对其进行划分, 选择 80% 作为训练样本, 20% 作为测试样本, 得到训练样本 918 条, 测试样本 217 条. 将商品对应的出价次数加入到原始数据中, 作为类标号使用. 利用本文提出的模型进行训练及预测, 结果中正确预测出价次数 72 条, 占全部样本的 33.2%; 利用公式 (1) 进行价格计算, 共有 47 条完全正确, 即有 21.7% 的成交价格可以进行准确预测, 如图 2 所示.

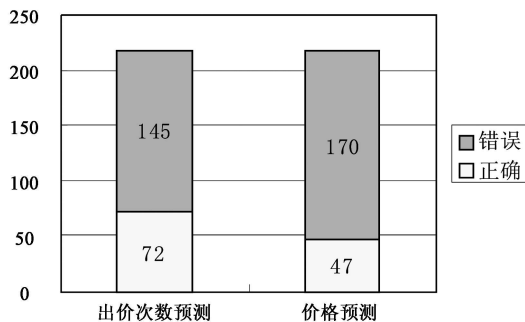


图 2 价格预测结果

为对模型性能进行更好的评价, 本文选择平均绝对误差 (MAE) 与平均相对误差 (MER) 对模型效果进行更进一步的分析. 其中:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\text{预测值}_i - \text{实际值}_i| \quad (2)$$

$$MRE = \frac{1}{N} \sum_{i=1}^N \frac{|\text{预测值}_i - \text{实际值}_i|}{\text{实际值}_i} \quad (3)$$

同时, 本文还将预测结果与常用来评价统计方法性能的平均值预测进行比较. 结果表明, 无论在绝对精度和相对精度上都显著优于平均值预测效果, 如表 4 所示.

表 4 预测模型性能比较表

	Bagging	平均值
MAE	25.91	MAE 49.60
MRE	0.62	MRE 4.23

显然, MAE 指标与 MRE 指标仅仅反映出模型预测误差的平均水平. 为获得更多信息, 本文对两种预测方法相对误差的累积分布进行统计, 累积分布如图 3 所示.

从图中可以看出, 本文所使用的 Bagging 集成决策树的方法预测精度要远远优于平均值预测. 值得注意的是, 从图中所示的相对误差与累积分布交点可以看出, Bagging 集成决策树方法中, 有 40% 的预测结果误差波动在 20% 以内; 而利用平均值方法时, 预测误差在 20% 以内的仅仅约占 10% 左右. 可以说明本文在预测精度上远优于平均值方法.

此外, 我们进一步将本文模型预测结果与 Heijst 发表于《Decision Support Systems》上的模型进行了对比. 由于 Heijst 在其文章中使用了 4 组不同的数据, 其中两组属于非同质产品, 为 Nike 运动鞋与 Canon 数码相机; 两组同质产品, 为 H700 摩托罗拉蓝牙耳机, 30G 苹果 MP3. 因本文数据属于非同质产品数据, 所以

我们选择与其模型中的非同质数据进行对比. 结果如表 5 所示:

表 5 模型性能指标对比表

本文模型	MAE		MRE	
		25.91		0.62
Heijst 所用模型	Nike 运动鞋		Canon 相机	
	MAE	MRE	MAE	MRE
	14.3	0.35	72.02	0.59

在这里可以看到, 本文模型预测 MRE 值在一定程度上略弱于 Heijst 模型预测结果.

我们认为主要是由以下原因造成:

1) 本文所用数据具有更高的异质性. Heijst 所用的 Nike 数据与 Canon 数据虽然在一定程度上属于异质数据, 但终究属于一个品牌产品. 而本文所用的女鞋数据属于不同品牌, 差异化程度更高. 这一点也可以从 Heijst 论文中的累积分布图中看出. 以 Canon 数据为例, 如图 4.

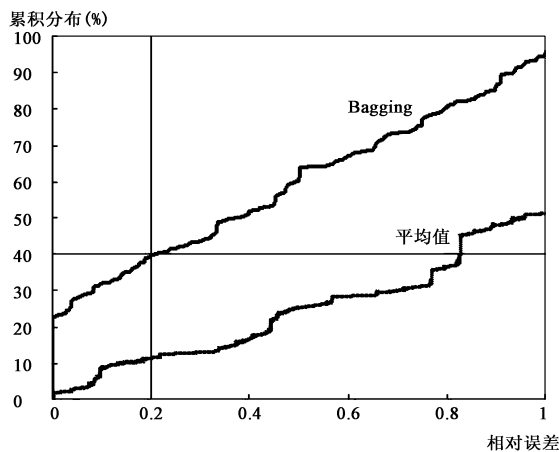


图 3 预测误差累积分布图

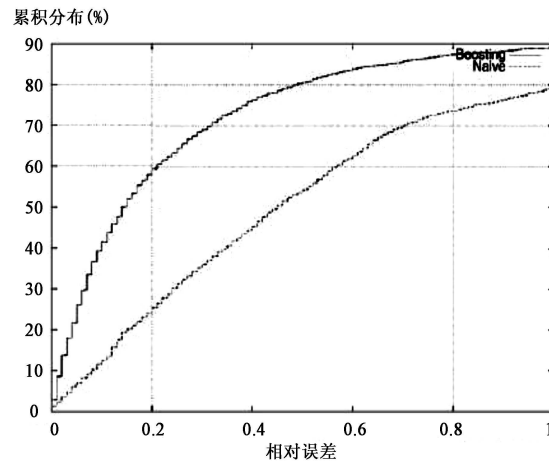


图 4 Heijst 模型中 Canon 数据累积分布图

从中可以看出, 利用平均值进行预测, 相对误差在 100% 之内的 Canon 数据达到 80%. 而本文数据仅仅为 50% 左右. 体现出本文数据具有更大的差异性, 因此, 在一定程度上造成了预测精确程度的下降.

2) 本文所用的训练样本数量更少. Heijst 所用模型中 Nike 训练样本为 4437 条, Canon 训练样本为 3683 条. 而本文的成交数据训练样本仅仅为 918 条, 远小于 Nike 与 Canon 训练样本数量. 这也是造成本文模型预测精度略弱于 Heijst 模型的原因. 可以推测, 当本文数据更具有同质性, 或扩大训练样本数量后, 本文模型预测精度将更进一步提高.

本文模型虽然由于数据异质性和训练样本数量制约, 在精度上略弱于 Heijst 模型. 但是, 在以下方面体现出了一定的优势:

1) 首先, 本文模型可以在一定比例上实现成交价格的精确预测. 从前面的分析中可以看出, 在成交记录中有 21.7% 的商品成交价格预测结果与实际相等. 而从 Heijst 模型的累积分布图中可以看出, 其模型精确预测的比例非常低, 几乎为 0.

2) 由于预测模型预测属性向量维度仅仅为 10, 维度的降低显著提高了运行效率. 本文模型在训练及预测过程中仅需要数秒时间, 与 Heijst 模型要运行数小时相比, 时间上有了非常显著的提高. 而以秒为单位的训练时间也为建立实时成交价格预测系统奠定了基础.

3) 由于降低了维度, 本文模型对训练样本的需求量显著降低. 在本文中的训练样本仅仅为 918 条数据. 与类似研究所用 4000 条以上训练数据相比, 仅仅为其 25% 左右.

综上所述, 本文所用预测模型取得了良好的效果.

5 结论

本文通过分析在线拍卖出价记录特点发现, 在全部包含出价记录的商品中, 有 40% 以上的商品完全以最

小加价幅度进行加价, 即其最终成交价格可以通过预测出价次数进行精确计算. 如将未成交商品视为出价 0 次, 则这一比例会提高到高达 79.55%. 基于此, 本文提出了一种全新的在线拍卖成交价格预测模型. 与前人研究中直接对在线拍卖成交价格进行预测不同, 本文通过分类算法估计出价次数, 进而间接对商品成交价格进行预测. 通过实际数据对模型进行验证, 结果表明模型运行效果良好. 对于全部商品是否成交预测准确率达 89%; 对于含出价的交易记录, 约 21.7% 的商品成交价格可以达到完全准确的预测; 40% 的商品预测结果误差在 20% 以内. 这一结果显著优于利用平均价格进行预测.

本文模型实验结果与 Heijst 发表于《Decision Support Systems》的研究结果相比, 在预测精度上略显不足. 作者认为这主要是由于本文所用数据具有更大的差异性, 同时训练样本仅为 Heijst 模型训练样本的 25% 左右造成的. 但本文所用的模型也体现出了一定的优势性. 首先, 本文模型可以实现 21.7% 左右的精确预测, 而 Heijst 模型精确预测比例较低, 几乎为 0; 其次, 本文方法类标号数量较少, 因而可以减少运算时间、提高预测准确率. 同时, 通过减少分类属性向量维度, 从而减少了对训练样本的需求数量. 本文中所用训练样本仅仅为类似研究的 1/4 左右, 但仍然取得了较好的效果. 更重要的是, 本文模型训练和预测仅仅需要数秒的时间, 与 Heijst 模型需要数小时的训练时间相比, 对时间的需求显著降低. 这为将来建立实时的在线拍卖成交价格预测决策支持系统打下了坚实的基础.

致谢 感谢《Decision Support Systems》杂志副主编、加洲大学长滩分校 Melody Kiang 教授在研究阶段给予的指导与研究工具上的支持. 在这里表示对 Melody Kiang 教授由衷的感谢.

参考文献

- [1] Wood C A. What factors drive final price in Internet auctions? An empirical assessment of coin transactions on eBay[C]// Annual Convention of the Institute for Operations Research and the Management Sciences. Miami Beach, Fla, 2001.
- [2] Bryan D, Lucking-Reiley D, Prasad N, et al. Pennies from eBay: The determinants of price in online auctions[J]. Journal of Industrial Economics, 2007, 55(2): 223–233.
- [3] Kauffman J, Wood C A. The effects of shilling on final bid prices in online auctions[J]. Electronic Commerce Research and Applications, 2005(4): 21–34.
- [4] Srinivas K R, Mayukh D. Modeling on-line art auction dynamics using functional data analysis[J]. Statistical Science, 2006, 21(2): 179–193.
- [5] Wang S S, Jank W, Shmueli G. Explaining and forecasting online auction prices and their dynamics using functional data analysis[J]. Journal of Business & Economic Statistics, 2008, 26(2): 144–160.
- [6] Ghani R, Simmons H. Predicting the end-price of online auctions[C]//Proceedings of International Workshop on Data Mining and Adaptive Modelling Methods for Economics and Management Held in Conjunction with the 15th European Conference on Machine Learning. Italy: Pisa, 2004.
- [7] 李雪峰, 刘鲁, 吴丽花. 在线拍卖商品最终成交价格预测 [J]. 计算机工程, 2006, 32(18): 189–191.
Li X F, Liu L, Wu L H. Prediction of the end-price of online auction item[J]. Computer Engineering, 2006, 32(18): 189–191.
- [8] Heijst D V, Potharst R, Wezel M V. A support system for predicting eBay end prices[J]. Decision Support Systems, 2008, 44(4): 970–982.
- [9] Kauffman R J, March S T, Wood C A. Design principles for long-lived internet agent[J]. International Journal of Intelligent Systems in Accounting, Finance & Management, 2000(9): 217–236.
- [10] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123–140.
- [11] Valiant L G. A theory of the learnable[J]. Communications of the ACM, 1984, 27(22): 1134–1142.
- [12] 李毓, 徐成贤. 修剪 Bagging 集成的方法及其应用 [J]. 系统工程理论与实践, 2008, 28(7): 105–110.
Li Y, Xu C X. A method for pruning Bagging ensembles and its applications[J]. Systems Engineering — Theory & Practice, 2008, 28(7): 105–110.