Weighted Functional K-Nearest Neighbor Method and Application in Close Price Prediction of Online Auction

Zhoushanyue He
Dept. Mathematics, Beijing Normal University, Beijing, P. R. China
Email: hezhoushanyue@gmail.com

Abstract: A weighted functional K-nearest neighbor (weighted fKNN) method has been proposed to predict the close price of online auction based on the traditional fKNN. Weighted fKNN takes into account the difference between the distance measures, and generally produces better fit. A case study of the Ebay record of Palm M515 7-day auctions from March, 2003 to May, 2003 has been done, noting that the mean absolute prediction error is less than 5%, a better prediction than the traditional fKNN.

Keywords: Internet e-commerce; online auction; K-nearest neighbor; functional data analysis

基于加权距离的函数型 K 近邻法及其在网上拍卖最终成交价格预测中的应用

何州杉月
数学科学学院，北京师范大学，北京，中国，100875
Email: hezhoushanyue@gmail.com

摘　要：为研究网络拍卖行为得价格趋势，本文在函数型 K 近邻法的基础上，提出了基于加权距离的函数型 K 近邻法，并对 2003 年 3 月到 5 月之间，Ebay 上对 Palm M515 的 7 天英式拍卖的数据进行了实例研究。结果发现该模型的平均绝对预测误差率在 5%左右，相对于原有模型有很大提高，说明了 fKNN 算法的合理性，为该算法在网络拍卖中的进一步应用提供了实证分析。

关键词：网络电子商务；网上拍卖；K 近邻法；函数型数据分析

1 引言

拍卖是通过一系列明确的规则和买者竞价所决定的价格来决定资源配置的一种市场机制。McAfee 和 Mcmillan 这样描述拍卖：“拍卖是市场参与者根据报价按照一系列规则决定资源的分配和价格的一种市场机制”[1]。随着经济的发展，拍卖已逐渐成为产品市场和交易市场必不可少的交易制度，作为一种有效的市场定价方式已经受到人们的广泛认可。互联网的出现和普及，为拍卖的网络化创造了条件，演变出新的拍卖形式——网上拍卖。与传统拍卖相比，网上拍卖由于其对互联网的依托，形成了以下一些特点：

1．从买方角度上来说，网上拍卖不再是只有少数人才能参与，而成为了每个网民都可以参与的交易方式。

2．从卖方角度上来说，网上拍卖的壁垒远远低于传统拍卖，进入市场变得容易且成本低廉，于是网上拍卖的物品也不仅限于昂贵的文物和艺术品，而是囊括了各式各样的产品。

3．从拍卖规则上来说，网上拍卖中竞买人不需要同时竞价，这使得竞买人可以在时间上有更大的灵活性，而且可以产生更大的拍卖市场。

4．网络的距离性使得拍卖中的信息不对称更加突出。

上述特点使得传统的拍卖模型并不能直接应用到网上拍卖上。与传统拍卖的研究文献相比，对网上拍卖的研究比较少。究其原因，一方面是因为网上拍卖出现得较晚，另一方面由于一部分研究者忽略了网上拍卖本质上的特殊性。目前对网上拍卖的研究主要是采用理论和实验的办法。理论研究的核心就是拍卖理论，即采用博弈论的方法来研究网上拍卖的特定问题，如 Sulin, Andrew and Zhang (2003) [2]。实验的方法主要通过对网上拍卖丰富的数据进行统计研究，来对特定问题进行解释或者预测，如 David Lucking-Reiley et al. (2007) [3]。

978-1-935068-40-2 © 2010 SciRes.

网上拍卖的一个重要研究对象是最终成交价格。对于竞买人来说，对最终成交价格的预测可以增加其信息量，从而更好地做出决策；对于卖方来说，预测最终成交价格可以帮助其决定最佳出售时间和评估库存价值。目前国内尚没有对网上拍卖最终成交价格预测的研究，而国外对这个问题的研究正逐渐引起注意。Ghani 和 Simmons (2004) [4]提出了一个只利用了拍卖开始时的信息来进行最终成交价格预测的模型。Wang et al. (2008) [5]指出，通过将拍卖价格动态信息（价格增加速度和加速度）纳入考虑可以提高预测的精确性。Dass et al. (2009) [6]以当代印度艺术品的网上拍卖为例提出了一个相似的最终成交价预测模型。Jank 和 Zhang (2008) [7]设计出一套可以自动从成百上千的拍卖中挑选出一场合适的网上拍卖并且决定何时出价的系统，对最终成交价格的预测是其中重要的一环。

本文在 Zhang, Jank 和 Shmueli (2010) [8]提出的函数型 K 近邻法（functional K-nearest neighbor）的基础上予以改进，提出加权函数型 K 近邻法（weighted functional K-nearest neighbor）,并以实例说明此方法能够更加精确地对网上拍卖的最终成交价格进行预测。

由于拍卖是一个动态的过程，要尽可能精确地预测最终成交价格，我们不仅要考虑拍卖商品的属性、拍卖开始时间等静态变量，还需要考虑拍卖中的时变变量（比如出价次数）和动态变量（比如拍卖价格的增长速率）。本文中使用的变量主要有：

1. 拍卖的结束时间。由于文中所用的拍卖数据都来源于 7 天拍卖，所以拍卖的结束时间跟拍卖的开始时间具有同等的意义。

2. 卖方等级，也就是网上拍卖系统对卖方售卖信誉、服务质量等的量化指标

3. 开拍价格

4. 拍卖过程中的出价次数。对由一个竞买人多次出价的情况，按其出价次数计。

5. 当前价格，即预测起始点 t 时刻的价格，在本文的实例中我们取预测起始时间为拍卖开场后 6.95 天，即距离结束大概 72 分钟。

6. 价格轨迹，使用 Beta 模型拟合拍卖的价格轨迹，轨迹中包含了价格增加速度和价格加速度（分别为轨迹函数的一次导数和二次导数）的信息

2 基于加权距离的函数型 K 近邻法

2.1 Beta 模型拟合价格轨迹

注意到拍卖价格曲线的特点：首先，由于部分拍卖的出价点较少，为了保证参数估计的精确性，应该采用含尽可能少的参数的模型；其次，由于拍卖中价格是单调变化的，故拟合时应选用保持单调性的方法。我们采用 Jank, Shmueli and Zhang(2010) [9]提出的 Beta 模型来进行预测。与传统的 p 样条法和单调样条法相比，Beta 模型使用的参数更少，保持单调递增，同时能够描述多种拍卖价格曲线的特点，并且可以给出曲线间的 KL 距离的解析表达式，为后续计算提供了方便。

Beta 模型，实质就是用 Beta 累积分布函数来拟合价格走势。Beta 累积分布函数如下：

$$F(x,\alpha,\beta)=\frac{\int_0^x u^{\alpha-1}(1-u)^{\beta-1}du}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} \quad (1)$$

由于 Beta 累积分布函数必过（0，0）和（1，1）两点，因此在使用 Beta 模型进行拟合时，应先将出价时间和价格按比例变到 0 和 1 之间：

$$\begin{aligned}\overline{t_0}&=\vec{t}/7\\ \overline{p_0}&=\vec{p}/\max(\vec{p})\end{aligned} \quad (2)$$

由于 Beta 累积分布函数的形状完全由参数 α，β 确定，故只需要求出使得

$$p_0=\frac{\int_0^{t_0} u^{\alpha-1}(1-u)^{\beta-1}du}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} \quad (3)$$

即可。

2.2 定义合適的距离度量

要采用 K 近邻估计法来预测最终成交价格，关键在于合适的距离度量。

事实上，距离度量的选择要符合意义明确，容易估算的特点。对于本文所使用的 6 种变量各自的距离度量，我们采用 Zhang, Jank 和 Shmueli (2010) [8]的方法来定义。具体来说::

1．对拍卖的结束时间，距离等于其相差的天数；

2．对卖方等级，距离等于其相差的级数；

3．对开拍价格，距离等于其价格差；

4．对出价次数，距离等于次数之差；

5．对当前价格，距离等于当前价格差；

6. 对价格轨迹，距离等于轨迹曲线间的 KL 距离。对于 Beta 累计分布函数来说，KL 距离有着较为简单的表示形式 [10]：

$$\begin{aligned}&\mathrm{D}_{KL}(\alpha,\beta,\alpha',\beta')=\\&\ln\frac{B(\alpha',\beta')}{B(\alpha,\beta)}-(\alpha'-\alpha)\psi(\alpha)\\&-(\beta'-\beta)\psi(\beta)\\&+(\alpha'+\beta'-\alpha-\beta)\psi(\alpha+\beta)\end{aligned}\quad(4)$$

其中，B 和 ψ 分别表示 Beta 函数和 Digamma 函数。

计算出距离之后，将各变量之间的距离分别除以各变量距离的最大值。这样，就将所有的距离都化为 0 和 1 之间，使得各变量的距离度量具有一定的可比性。

得到已知拍卖 i 与待预测拍卖各变量的距离度量后（表示为 $d^i{}_1 \sim d^i{}_6$），我们采用加权平均的方法来得到拍卖间总的距离度量（设权重为 $w_1 \sim w_6$）。对各变量的距离度量进行加权平均，得到已知拍卖 i 与待预测拍卖的总体距离的度量 $D_i=\sum_{n=1}^{6} d^i{}_n w_n$，。将已知拍卖按照与待预测拍卖 P 的距离从小到大排列后，记为 $\{D_{(i)}\}, i=1,2,...,N$，其中 N 为已知拍卖数。我们通过与待预测拍卖距离最小的前 K 个已知拍卖的数据来预测拍卖 P 的最终成交价格。

$$\hat{\mathrm{P}}_{end}=P_{current}+\frac{\sum_{i=1}^{K}[(P_{(i)end}-P_{(i)current})/D_{(i)}]}{\sum_{i=1}^{K}D_{(i)}}\quad(5)$$

其中，$P_{(i)end}$ 表示排序后的已知拍卖的最终成交价格，$P_{(i)current}$ 表示排序后的已知拍卖的当前价格，$P_{current}$ 表示待预测拍卖的当前价格，$\hat{P}_{end}$ 表示待预测拍卖最终成交价格的估计值。

2.3 权重和参数 K 的选择

本文将遗传算法和交叉验证法结合起来搜索参数 K 和最优权重。关于遗传算法详细内容可参见 Ji Genlin (2004) [11]。将数据集分为两组，一组作为训练集，一组作为测试集，用模型在测试集上的平均绝对预测误差率（mean absolute prediction error, MAPE）

$$\mathrm{MAPE}=\frac{1}{M}\sum_{i=1}^{M}\left|\frac{P_{end,i}-\hat{P}_{end,i}}{P_{end,i}}\right|\quad(6)$$

作为预测精确度的度量（其中 M 为预测拍卖的数量），其相反数作为个体适应度。让不同组合的权重和参数 K 通过多次选择、重组和变异后，从所得到的组合中选择使模型在测试集中的 MAPE 达到最小的权重和参数 K 的组合，将其作为最优的权重与参数 K。

2.4 模型算法

设预测起始点为 t，

（一）将已知的拍卖数据分为 2 组，一组为训练集，一组为测试集。

（二）用 Beta 模型拟合训练集、测试集和预测集的直到 t 的价格轨迹。

（三）计算训练集中样本到测试集中样本以及训练集中样本到预测集中样本的每个变量的距离，并进行标准化。

（四）利用遗传算法，将权重与 K 的不同组合视为个体，以模型在测试集上的平均绝对预测误差（MAPE）的相反数作为个体适应度。通过多次选择、重组和变异后，得到使 MAPE 达到极小的权重与 K 的组合，将其作为模型中的权重和 K 值。

（五）以第（五）步中得到的权重对训练集中样本到预测集中样本的每个变量的距离进行加权平均，得到总的距离度量。

（六）利用（四）中计算的 K 值和（五）中总距离度量，根据公式（5）计算预测集中拍卖最终成交价格的预测值。

3 实例

本文采用了在 2003 年 3 月到 5 月之间，Ebay 上对 Palm M515 的 7 天英式拍卖的共 136 组拍卖的数据。数据集中除包括上述 1～4 项变量之外，还包括在每次拍卖中竞买人出价的时间以及价格。拍卖商品的同质性使得我们可以不用考虑商品属性对拍卖的影响，从而大大地简化了预测过程。

我们将这 136 场拍卖按照结束时间的先后顺序排列后分为 3 组，其中第一组包含 69 场拍卖数据，作为训练集，用于模型参数估计，第二组包含 33 场拍卖数据，作为测试集，用于参数选择，第三组包含 34 场拍卖数据，作为预测集，用于评价模型预测能力的好坏。

遗传算法通过选择使模型在测试集上 MAPE 达到最小的权重和 K 值，我们得到 K=2，W=[0.1109, 0.2126, 0.0903, 0.1977, 0.2314, 0.1571]，相应的平均绝对预测误差率为 4.63%，低于 Zhang, Jank 和 Shmueli (2010) [8]中采用普通求和的 MAPE 值。注意到在 Zhang,

978-1-935068-40-2 © 2010 SciRes.

Jank 和 Shmueli (2010）[8]中，作者通过变量选择，提出只用时变因素来度量距离精度最高，而在本文中，对于时变因素的权重亦较高（0.1977 和 0.2314），从而从侧面说明本文模型的合理性。

进一步的分析可以发现，使用遗传算法进行最优参数搜索，容易发生过度拟合的现象。因此，我们用拟合后的模型预测第 3 组拍卖数据即预测集的最终成交价格，得到的平均绝对误差率为 5.13%（模型预测结果与实际最终成交价格如图 1 所示）。这与模型在第 2 组数据上的表现相差不大，故可以认为在本例中不存在明显的过度拟合。

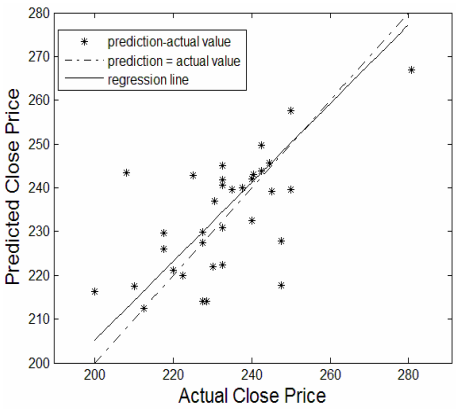

Figure 1 the comparison of predicted close prices and actual close prices of prediction set

图 1 模型预测结果与实际最终成交价格对比

4 结论

本文在 Zhang, Jank 和 Shmueli (2010) [8]的基础上，提出了基于加权平均构造距离度量的函数型 K 近邻法（weighted fKNN）。利用遗传算法得到最佳权重和 K 后，求出所有已知拍卖到待预测拍卖的距离度量，并选出与待预测拍卖距离最短的 K 组已知拍卖，以距离的倒数为权重对这 K 组拍卖自预测时刻至收市的价格增量做加权平均，得到待预测拍卖的价格增量预测值，再加上待预测拍卖在预测时刻的价格，得到待预测拍卖的最终成交价格预测。与原始的函数型 K 近邻法在同一数据集上的预测结果相比，预测精度有了明显地提高

致 谢

感谢马里兰大学的 Galit Shmueli 教授提供数据及对本文结果发表评论。

References（参考文献）

[1] R. P. McAfee, J. M. Auctions and Bidding [M]. Journal of Economic Perspectives. 1989, 70:2-33

[2] Sulin Ba, Andrew B. Whinston, and Han Zhang. Building trust in online auction markets through an economic incentive mechanism [J]. *Decision Support Systems*, 2003, V35,I3,pp.273-286

[3] David Lucking-Reiley, Doug Bryan, Naghi Prasad, Daniel Reeves. Pennies from eBay: the Determinants of Price in Online Auctions [J]. *The Journal of Industrial Economics*. 2007, V55,I2,p.223-235

[4] Ghani, R. and Simmons, H. Predicting the end-price of online auctions. In International Workshop on Data Mining and Adaptive Modeling Methods for Economics and Management, 2004, Pisa, Italy.

[5] Wang.S., Jank.W., and Shmueli.G. Explaining and forecasting online auction prices and their dynamics using functional data analysis [J]. *Journal of Business and Economic Statistics*, 2008, 26(2):144-160.

[6] Dass. M., Jank. W., and Shmueli, G. Dynamic price forecasting in simultaneous online art auctions. In Casillas and Martnez-Lpez (Eds.): Marketing Intelligent Systems using Soft Computing, [M], 2009, Springer.

[7] Jank. W. and Zhang.S. An automated and data-driven bidding strategy for online auctions. Technical report, University of Maryland, available at http://ssrn.com/abstract=1427212

[8] Zhang.S., Jank.W., Shmueli.G. Real-Time Forecasting of Online Auctions via Functional K-Nearest Neighbors, *International Journal of Forecasting*, 2010, vol. 26, pp. 666-683

[9] Jank. W, Shmueli.G, Zhang.S. A Flexible Model for Estimating Price Dynamics in Online Auctions, *JRSS C*, In Press

[10] Raubera, T., Braun, T., & Berns, K. Probabilistic distance measures of the dirichlet and beta distributions. *Pattern Recognition*, 2008, 41(2), 637-645.

[11] Ji Genlin, Survey on Genetic Algorithm. *Computer Applications and Software*, 2004,v21-2,pp.69-73
吉根林，遗传算法研究综述。计算机应用与软件，2004，21-2，pp.69-73