Pricing Analysis in Online Auctions Using Clustering and Regression Tree Approach

Preetinder Kaur, Madhu Goyal, and Jie Lu

School of Software, University of Technology, Australia
preetinder.kaur@student.uts.edu.au, {madhu,jielu}@it.uts.edu.au

Abstract. Auctions can be characterized by distinct nature of their feature space. This feature space may include opening price, closing price, average bid rate, bid history, seller and buyer reputation, number of bids and many more. In this paper, a price forecasting agent (PFA) is proposed using data mining techniques to forecast the end-price of an online auction for autonomous agent based system. In the proposed model, the input auction space is partitioned into groups of similar auctions by k-means clustering algorithm. The recurrent problem of finding the value of k in k-means algorithm is solved by employing elbow method using one way analysis of variance (ANOVA). Based on the transformed data after clustering, bid selector nominates the cluster for the current auction whose price is to be forecasted. Regression trees are employed to predict the end-price and designing the optimal bidding strategies for the current auction. Our results show the improvements in the end price prediction using clustering and regression tree approach.

Keywords: Online auctions, Price forecasting, Bidding strategies, Data mining, Clustering, Regression trees.

1 Introduction

Predicting the end price of an auction has become an increasingly important area of research because buyers and sellers can be offered a great benefit by using the above predicted prices [1][2][5]. The online auctions are exchange mechanisms which produce a large amount of transaction data. This data can be exploited to forecast the final prices of the auction items, if utilized properly. Researchers' efforts can be noticed in the area of forecasting end-price of an auction using machine learning techniques, functional data analysis, time series analysis [1][2][3][4][5][6][7][8]

Software agent technology is one of the most popular mechanisms used in on-line auctions for buying and selling the goods. Software agent is a software component that can execute autonomously, communicates with other agents or the user and monitors the state of its execution environment effectively [9][10][11] [13]. Agents can use different auction mechanisms (e.g. English, Dutch, Vickery etc.) for procurement of goods or reaching agreement between agents. Agents make decisions on behalf of consumer and endeavor to guarantee the delivery

L. Cao et al. (Eds.): ADMI 2011, LNAI 7103, pp. 248–257, 2012.

of item according to the buyers preferences. These are better negotiators than human being in terms of monitoring, remembering and emotional influence. In these auctions buyers are faced with difficult task of deciding amount to bid in order to get the desired item matching their preferences. This bid amount can be forecasted effectively by analyzing the data produced as an auction progresses (historical data). This forecasted bid can be exploited by the bidding agents to improve their behaviors. Also the analysis of plethora of data produced in the online auction environment can be done by using DM techniques [1,12][4][7][8][14].

Predicting the end price depends on many factors, such as item type, type of auction, quantity available, opening price, number of bidders, average bid amount and many more. Price dynamics of the online auctions can be different even when dealing with auctions for similar items. Functional data analytical tools have been used to characterize different type of auctions that exhibit different price dynamics in [8]. In this paper, a price forecasting agent (PFA) is proposed using data mining techniques to forecast the end-price of an online auction for autonomous agent based system. A clustering based approach is used to characterize different type of auctions. In the proposed model, the input auctions are clustered into groups of similar auctions based on their characteristics using k-means algorithm. To decide the value of k in k-means algorithm is a recurrent problem in clustering and is a distinct issue from the process of actually solving the clustering problem. The optimal choice of k is often ambiguous, increasing the value of k always reduce the error and increases the computation speed. In this paper, we are exploring Elbow approach using one way analysis of variance (ANOVA) to estimate the value of k. Based on the transformed data after clustering and the characteristics of the current auction whose price is to be forecasted, bid selector nominates the cluster. Regression trees are employed to the corresponding cluster for forecasting the end-price and to design the bidding strategies for the current auction.

The rest of the paper is organized as follows. In section 2 we discuss the related work to the topic. Section 3 illustrates the design of PFA- Price Forecasting Agent describing the data mining mechanism developed for forecasting the end-price and optimizing the bidding strategies for online auctions. Section 4 depicts the experimental results. Section 5 discusses the conclusions of the paper and presents directions for the future work.

2 Related Work

In the recent literature, different approaches have been presented for end price prediction in the online auctions environment. A data mining based multi-agent system has been designed in favor of a multiple on-line auctions environment for selecting the auction, in which the traded item will be sold at the lowest price [4]. The K-means clustering technique has been used to classify auctions into discrete clusters. Clustering operates dynamically on multiple auctions as bid price changes in running auctions. The results of the dynamic clustering

are fed into the agents, and by employing probability-based decision making processes, agents deduce the auction that is most likely to close at the lowest price. Experimental results have demonstrated the robustness of the designed system for multiple on-line auctions with little or no available information.

Forecasting [3] has been proposed as a time series problem and has been solved using moving averages and exponential smoothing models. Authors in this paper also emphasized that there is no one best forecasting technique for a particular set of data. Authors used sequence mining to improve the decision mechanism of an agent for predicting bidding strategy of a set of trading agents [7]. Prediction are mostly based on the continuously growing high dimensional bids history, so authors identified that sequence mining technique can be employed to classify the high dimensional frequent patterns in the bids history. Also the sliding window concept has been used for feeding the continuous classifier. Classification and meta-classification are applied to predict the final bid.

Predicting end price of an online auction has been stated as a machine learning problem and has been solved using regression trees, multi-class classification and multiple binary classification [2]. Among these machine learning techniques, posing the price prediction as a series of binary classification has been proved to be the best suited method for this task. In the literature, along with the machine learning techniques, traditional statistical methods have also been used to forecast the final prices of the auction item, but the experimental results demonstrated that the machine-learning algorithms such as BP networks outperform traditional statistical models [5]. Seeking the effect of clustering of data was also the concern of the authors [5]. Clustering has increased the accuracy of the results in case of BP networks but decreased the accuracy in logistic regression.

A support system for predicting the end-price of an online auction based on the item description using text mining and boosting algorithm has been proposed [6]. Emphasis is given by the authors on capturing the relevant information for incorporating into the price forecasting models for ongoing online auction [1]. A novel functional K-nearest neighbor (fKNN) forecaster based on functional and non-functional distance has been proposed and has been proved to be effective particularly for heterogeneous auction populations.

LS-SVM (Least Square Support Vector Machine) algorithm has been introduced by Zhou and Wang for forecasting in online electronic commerce [14]. Authors first improved the SVM to solve the sparsity and time lag problems existing in the traditional method and then they established the LS-SVM online forecast model based on the time factor elimination. Experimental results demonstrated almost same values for the forecasted and the actual price.

3 PFA-Price Forecasting Agent

A clustering based method is used to forecast the end-price of an online auction for autonomous agent based system. In the proposed methodology the input auctions are partitioned into groups of similar auctions depending on their different characteristics. This partitioning has been done by using k-means clustering

algorithm. The value of k in k-means algorithm is determined by employing elbow method using one way analysis of variance (ANOVA). Based on the transformed data after clustering and the characteristics of the current auction, bid selector nominates the cluster for price forecasting. End-price is predicted and bidding strategies are designed by using regression trees for the nominated cluster.

The price forecasting and bidding agent is represented in Fig. 1. Formally our approach consists of four steps. First, data is extracted from the bid server as per the requirements to form the agents knowledge base for online auctions. Let A be the set of the attributes collected for each auction then A=$\{a_1, a_2, ..., a_j\}$, where j is the total number of attributes. Then based on the auctions characteristics, similar auctions are clustered together. Secondly, k-estimator agent determines the best number of partitions for the overall auction data and then the set of similar auctions are clustered together in k groups. Let C be the set of clusters then C = $\{c_1, c_2, ..., c_k\}$, where k is the total number of clusters. Thirdly, based on the transformed data after clustering and the characteristics of the current auction, bid selector nominates the cluster for price forecasting. Finally, regression tree is employed to forecast the end price and to design the bidding strategies for the selected cluster.

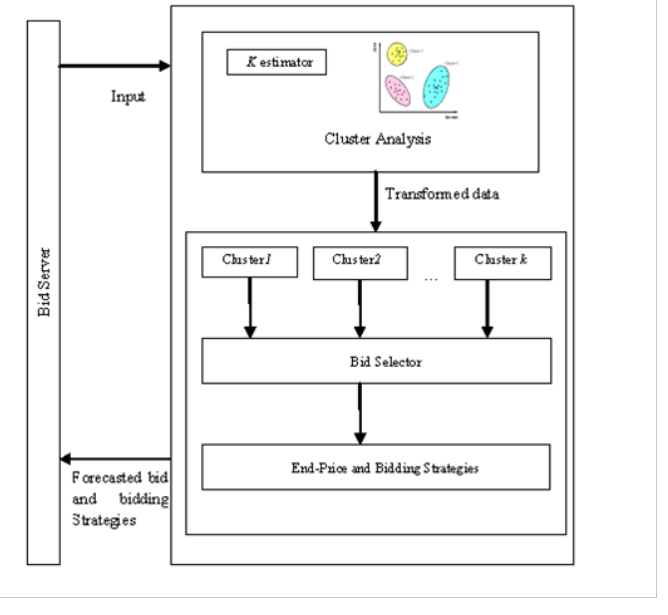

Fig. 1. PFA-Price Forecasting Agent

3.1 K-means Cluster Analysis

The main idea of k-means clustering is to define k centroids, one for each cluster. Initially k data are randomly chosen to represent the centroids. Each data point is assigned to the group that has the closest centroid. After assignment of each data point, positions of the k centroids are re-calculated as the average value of every cluster. These steps repeat until the centroids no longer move or minimizing the objective function J.

$$j = \Sigma_{j=1}^{k} \Sigma_{i=1}^{n} \|x_i^{(j)} - c_j\|^2 \quad (1)$$

Where $\|x^{(j)} - c_j\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre. It indicates the distance of the n data points from their respective cluster centers.

K-means clustering technique is used here to categorize different types of auctions based on some predefined attributes from the vast feature space of online auctions. The feature space may include average bid amount, average bid rate, no. of bids, item type, seller reputation, opening bid, closing bid, quantity available, type of auction, duration of the auction, buyer reputation and many more. In this paper, to classify different types of auctions, we focus on only a set of attributes; opening bid, closing price, number of bids, average bid amount and average bid rate. Now A=$\{OpenBi, ClosePi, NUMi, AvgBi, AvgBRi\}$

Where A is the set of attributes for an auction.

$OpenB_i$ is the starting price of i^{th} auction

$CloseP_i$ is the end price of i^{th} auction

NUM_i is the total number of bids placed in i^{th} auction

$AvgB_i$ is the average bid amount of i^{th} auction and can be calculated as $Avg(B_1, B_2, ...B_l)$ where B_1 is the 1^{st} bid amount, B_2 is the second bid amount and B_l is the last bid amount for i^{th} auction.

$AvgBR_i$ is the average bid rate of i^{th} auction and can be calculated as $\frac{1}{n}\frac{B_{i+1}-B_i}{t_{i+1}-t_i}$

where B_{i+1} is the amount of $(i+1)^{th}$ bid, B_i is the amount of i^{th} bid, t_{i+1} is the time at which $(i+1)^{th}$ bid is placed and t_i is the time at which i^{th} bid is placed.

3.2 K-estimator

To decide the value of k in k-means algorithm is a recurrent problem in clustering and is a distinct issue from the process of actually solving the clustering problem. The optimal choice of k is often ambiguous, increasing the value of k always reduce the error and increases the computation speed. The most favorable method to find k adopts a strategy which balances between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster. There are several approaches to decide the value of k: rule of thumb, the elbow method, information criterion approach, an information theoretic approach, choosing k using the Silhouette and cross-validation.

In this paper, we are exploring Elbow approach using one way analysis of variance (ANOVA) to estimate the value of k. Number of clusters are chosen so that adding another cluster doesn't give much better modeling of the data. A graph has been plotted for percent of variance against the number of clusters. At some point the marginal gain will drop, giving an angle in the graph Fig. 2. The number of clusters is chosen at this point.

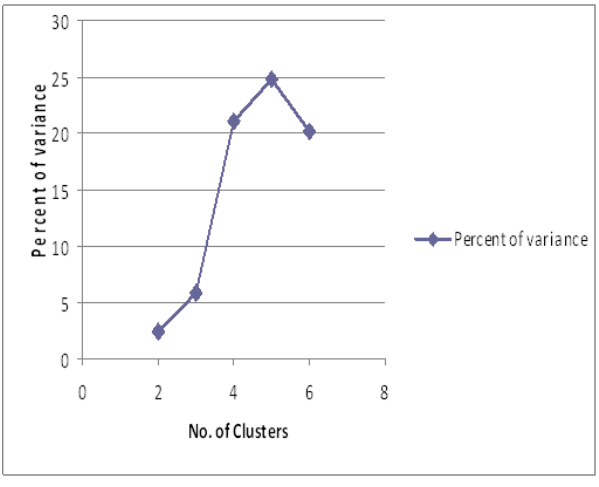

Fig. 2. Choosing the value of k

Below is the mathematical model explained for the value of k.

$$percentofvariance = \frac{\Sigma_{n=1}^{k}\Sigma_{i=1}^{N_n}\left(X_{in}-\bar{X}\right)^2 - \Sigma_{n=1}^{k}\Sigma_{i=1}^{N_n}\left(X_{in}-\bar{X_n}\right)^2}{\Sigma_{n=1}^{k}\Sigma_{i=1}^{N_n}\left(X_{ij}-\bar{X}\right)^2} \quad (2)$$

Where k is the total no. of clusters N_n is total no. of elements in the n^{th} cluster $\bar{X_n}$ is the mean of distances of auctions from the cluster center in n^{th} cluster X_{in} is the distance of i^{th} auction in n^{th} cluster from its cluster center.

3.3 Bid Selector

In order to decide that the current auction belongs to which cluster, the bid selector is activated. Based on the transformed data after clustering and the characteristics of the current auction, bid selector nominates the cluster for the current auction whose price is to be forecasted.

3.4 Extracting Bidding Strategies

Regression Tree is an analytic procedure for predicting the values of a continuous response variable from continuous predictors. We can derive simple if-then rules

from these predictions. In this paper, we are employing regression tree to predict the end price and to design the optimal bidding strategies for the targeted cluster. This targeted cluster is the cluster which bid selector nominates for the current auction whose end price is to be forecasted. Regression trees are built in two phases. The first phase is growing phase and the second phase is pruning phase. In the growing phase, the tree is grown until no error reduction on the training possible or a pre-defined threshold has been reached. The resulting model usually over-fits the data. This is overcome by the pruning the tree. The best tree is chosen when both the output variable variance and the cost-complexity factor are minimized. The trade-off between these two criterions should be considered. There are two ways to accomplish this. One is test-sample cross-validation and the other is v-fold cross-validation. In this paper, test-sample cross-validation method is used to determine the best pruned tree for designing the optimal bidding strategies of the online auction.

4 Experimentation

In the proposed approach end-price is predicted and the bidding strategies are designed for an online auction by exploiting regressions trees on each cluster which are generated by applying k-means algorithm on the input auctions dataset. The outcome of the proposed model with clustering is compared with the classic (without clustering) model for price prediction. The improvement in the error measure for each cluster for a set of attributes gives support in favor of the proposed model using clustering. Optimal bidding strategies are designed by employing regression trees on each cluster and the model is evaluated by test-sample cross validation approach. Our dataset includes the complete bidding records for 149 auctions for new Palm M515 PDAs. All are 7-day auctions that were transacted on eBay between March and June, 2003 (a sample is available online at `http://www.rhsmith.umd.edu/digits/statistics/pdfs_docs/Palm_7day_149auctions.xls`) Table 1 shows the statistical information of the data set.

Table 1. Description of the data

	Min	Max	Mean	Std. Deviation
Opening Bid(OpenB)	0.01	240	40.552	69.7142
Closing Price(CloseP)	177	280.65	228.245	16.098
# Bids(NUM)	2	51	21.358	10.155
Avg bid amt(AvgB)	77.473	243.75	148.912	33.133
Avg bid rate(AvgBR)	-2196.268	42679.939	11624.091	8143.416

OpenB: Starting price of an auction.
CloseP: End price of an auction.
NUM: Total number of bids placed in an auction.

AvgB: Average bid amount of each auction.
AvgBR: Average bid rate of each auction.

The value of k for k-means algorithm is estimated by Elbow approach using one way analysis of variance (ANOVA). Percent of variance is calculated after performing clustering for subsequent values of k using k-means algorithm for estimating the point where marginal gain drops. In our experiment, this point occurs after five numbers of clusters as shown in Table. 2. and Fig. 2. So we divide the input space in five clusters considering a set of attributes i.e. opening bid, closing price, no. of bids, average bid amount and average bid rate for a particular auction. These five clusters contain 20.27%, 34.46%, 18.92%, 20.95% and 5.41% of auctions data.

The prediction performance is evaluated using the root mean square error (RMSE) measure. This is used as a measure of the deviation between the predicted and the actual values. The smaller the value of RMSE, the closer is the predicted values to the actual values. Typically the RMSE is defined as

$$RMSE = \sqrt{\frac{1}{m} \Sigma_{i=1}^{m} \left(y_i - \bar{y_i}\right)^2} \tag{3}$$

Where $\bar{y_i}$ is the forecasted bid for the i^{th} auction based on the estimated model, y_i is the actual bid amount for the i^{th} auction and m is the total no. of auctions in the cluster.

Table 2. Percent of variance after subsequent clustering

No. of Clusters	Percent of Variance
2	2.46
3	5.8
4	21.06
5	24.75
6	20.22

The improvement in root mean square error after clustering for each of the five clusters are 16.43%, 29.47%, 20.76%, 31.17% and 64.91% respectively. This indicated that the proposed price forecasting agent can improve the accuracy of the forecasted price for online auctions.

Optimal bidding strategies are designed by employing regression trees on each cluster. To find the best tree, we have used the test-sample cross validation approach. We randomly divide the data into 80% training and 20% validation set and applied the tree-growing algorithm to the training data and grows the largest tree which over-fits the data. Then the tree is pruned by calculating the cost complexity factor at each step during the growing of the tree and deciding the number of decision nodes for the pruned tree. The tree is pruned to minimize the sum of (1) the output variable variance in the validation data, taken a terminal node at a time, and (2) the product of the cost complexity factor and the number

of terminal nodes. In our experiments, we applied regression trees on the second cluster only and designed the bidding rules for the same. Regression tree for cluster 2 is shown in Fig. 3. The non-terminal nodes present test/decisions on one or more attributes and the terminal nodes reflect the decision outcomes. Fig. 4 gives the rule set that sum up this regression tree.

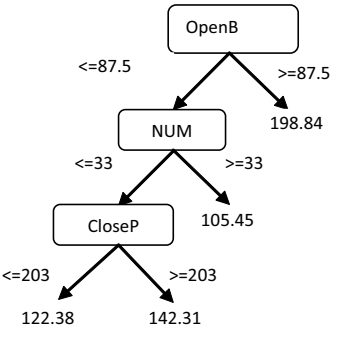

Fig. 3. Regression tree for cluster 2

1 If OpenB <=87.5	**And NUM** <=33	**And CloseP** <=203	**Then AvgB:** 122.38
2 If OpenB <=87.5	**And NUM** >=33		**Then AvgB:** 105.45
3 If OpenB >=87.5			**Then AvgB:** 198.84
4 If OpenB <=87.5	**And NUM** >=33	**And CloseP** >=203	**Then AvgB:** 142.31

Fig. 4. Rule set for cluster 2

5 Conclusions

In this paper we presented a clustering and regression trees based approach to forecast the end-price and to find the optimal bidding strategies of an online auction for autonomous agent based system. In the proposed model, the input auctions are partitioned into groups of similar auctions by k-means clustering algorithm. The recurrent problem of finding the value of k in k-means algorithm is solved by employing elbow method using one way analysis of variance (ANOVA). Then k numbers of regression models are employed to estimate the forecasted price of the online auction. Based on the transformed data after clustering, bid selector nominates the cluster for the current auction whose price is to be forecasted. Regression trees are employed to the corresponding cluster for forecasting the end-price and to design the bidding strategies for the current

auction. The outcome of the proposed model with clustering is compared with the classic model for price prediction. The improvement in the error measure for each cluster for a set of attributes gives support in favor of the proposed model using clustering. Optimal bidding strategies are designed by employing regression trees on each cluster and evaluating the model by test-sample cross validation approach. Further work will be focused in two directions; first, to improve the prediction model by exploiting decision trees and classification methods and secondly, study the importance of each attribute on the end price prediction for online auctions.

References

1. Zhang, S., Jank, W., Shmueli, G.: Real-time forecasting of online auctions via functional K-nearest neighbors. International Journal of Forecasting (2010)
2. Ghani, R., Simmons, H.: Predicting the end-price of online auctions. In: Proceedings of the International Workshop on Data Mining and Adaptive Modeling Methods for Economics and Management, Held in Conjunction with the 15th European Conference on Machine Learning (2004)
3. Guo, W., Chen, D., Shih, T.: Automatic forecasting agent for e-commerce applications. In: 20th International Conference on Advanced Information Networking and Applications (AINA 2006), pp. 1–4 (2006)
4. Kehagias, D.D., Mitkas, P.A.: Efficient E-Commerce Agent Design Based on Clustering eBay Data. In: IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops, pp. 495–498 (2007)
5. Xuefeng, L., Lu, L., Lihua, W., Zhao, Z.: Predicting the final prices of online auction items. Expert Systems with Applications 31, 542–550 (2006)
6. Heijst, D., Potharst, R., Wezel, M.: A support system for predicting ebay end prices. Econometric Institute Report (2006)
7. Nikolaidou, V., Mitkas, P.A.: A Sequence Mining Method to Predict the Bidding Strategy of Trading Agents. In: Cao, L., Gorodetsky, V., Liu, J., Weiss, G., Yu, P.S. (eds.) ADMI 2009. LNCS, vol. 5680, pp. 139–151. Springer, Heidelberg (2009)
8. Jank, W., Shmueli, G.: Profiling price dynamics in online auctions using curve clustering. Technical report, Smith School of Business, University of Maryland, pp. 1–28 (2005)
9. Greenwald, A., Stone, P.: Autonomous bidding agents in the trading agent competition. IEEE Internet Computing, 52–60 (2001)
10. Byde, A., Preist, C., Jennings, N.: Decision procedures for multiple auctions. In: ACM First International Joint Conf. on Autonomous Agents and Multi-Agent Systems, pp. 613–620 (2002)
11. Anthony, P., Jennings, N.: Evolving bidding strategies for multiple auctions. In: Proc. of 15th European Conf. Artificial Intelligence, pp. 178–182 (2002)
12. Cao, L., Gorodetsky, V., Mitkas, P.A.: Agent Mining: The Synergy of Agents and Data Mining. IEEE Intelligent Systems 24(3), 64–72 (2009)
13. Goyal, M., Kaushik, S., Kaur, P.: Automated Fuzzy Bidding Strategy for Continuous Double Auctions Using Trading Agents Attitude and Market Competition. International Journal of Agent Technologies and Systems 2(4) (2010)
14. Min, Z., Qiwan, W.: The On-line Electronic Commerce Forecast Based on Least Square Support Vector Machine. In: Second International Conference on Information and Computing Science, ICIC 2009, vol. 2, pp. 75–78 (2009)