Clustering and Labelling Auction Fraud Data

Preprint · August 2018

DOI: 10.6084/m9.figshare.6993308

Ahmad Alzahrani
University of Regina

Samira Sadaoui
University of Regina

Clustering and Labeling Auction Fraud Data

Conference Paper · January 2019

2 authors:

Ahmad Alzahrani
University of Regina
5 PUBLICATIONS **6** CITATIONS

Samira Sadaoui
University of Regina
76 PUBLICATIONS **220** CITATIONS

Clustering and Labeling Auction Fraud Data

Ahmad Alzahrani and **Samira Sadaoui**

Computer Science Department, University of Regina, Regina, SK, Canada
{alzah234,sadaouis}@uregina.ca

Abstract. Although shill bidding is a common fraud in online auctions, it is however very tough to detect because there is no obvious evidence of it happening. There are limited studies on SB classification because training data are difficult to produce. In this study, we build a high quality labeled shill bidding dataset based on recently scraped auctions from eBay. Labeling shill biding instances with multi-dimensional features is a tedious task but critical for developing efficient classification models. For this purpose, we introduce a new approach to effectively label shill bidding data with the help of the robust hierarchical clustering technique CURE. As illustrated in the experiments, our approach returns remarkable results.

Keywords: Auction Fraud · Shill Bidding · Hierarchical Clustering · CURE · Silhouette · Data Labeling

1 Introduction

In the last three decades, we witnessed a significant increase in exchanging goods and services over the Web. According to the World Trade Organization, the worldwide merchandise during the period 1995-2015 was over 18 billion [1]. Online auctions are a very profitable e-commerce application. For instance, in 2017, eBay claimed that the net revenue attained 9.7 billion US dollars, and the number of active users hit 170 million [2]. Regardless of their popularity, e-auctions remain very vulnerable to cyber-crimes. The high anonymity of users, low fees of auction services and flexibility of bidding make auctions a great incubator for fraudulent activities. The Internet Crime Complain Center announced that auction fraud is one of the top cyber-crimes [2]. As an example, the complaints about auction fraud in only three states, California, Florida and New York, reached 7,448 in 2016 [2]. Malicious moneymakers can commit three types of fraud, which are pre-auction fraud, such as auctioning of black market merchandise, in-auction fraud that occurs during the bidding time, such as Shill Bidding (SB), and post-auction fraud, such as fees stacking. Our primary focus is on the SB fraud whose goal is to increase the profits of sellers by placing many bids through fake accounts and colluding with other users. SB does not leave any obvious evidence unlike

[1] https://www.wto.org/english/res_e/statis_e/its2015_e/its2015_e.pdf

[2] https://www.statista.com

the two other auction fraud. Indeed, buyers are not even aware that they have been overcharged.

Identifying relevant SB strategies, determining robust SB metrics, crawling and preprocessing commercial auction data, and evaluating the SB metrics against the extracted data make the study of SB fraud very challenging as demonstrated in our previous technical paper [1]. In addition, labeling SB instances with multi-dimensional features is a critical phase for the classification models. In the literature, labeling training data is usually done manually by the domain experts, which is quite a laborious task and prone to errors. Due to the lack of labeled SB training datasets, the prime contribution of this paper is to produce high-quality labeled SB data based on commercial auction transactions that we extracted from eBay and preprocessed [1]. As illustrated in Figure 1, we introduce a new approach to effectively label SB data with the help of data clustering. Firstly, we split the SB dataset into several subsets according to the different bidding durations of the extracted auctions. Secondly, we efficiently partition each SB subset into clusters of users with similar bidding behaviour. Last, we apply a systematic labeling method to each cluster to classify bidders into normal or suspicious.

Hierarchical clustering is significantly preferable over partitioning clustering because it provides clusters with a higher quality [10]. This type of data clustering has been utilized successfully in numerous fraud studies [8, 10]. In fact, we employ the Clustering Using REpresentatives (CURE) technique to produce the best differentiation between normal and suspicious activities. CURE [11] has proved over the years to be a highly efficient clustering method in terms of eliminating outliers and producing high quality clusters, especially for large-scale training datasets. The labeled SB dataset that we produced can be utilized by the state-of-the-art supervised classification methods. Furthermore, the accuracy of new predictive models can also be tested using our SB training dataset.

The rest of the paper is organized as follows. Section 2 discusses related work on data clustering in the context of online auctions. Section 3 describes the SB patterns used in this research as well as the SB data produced from the commercial website eBay. Section 4 explains how to apply the hierarchical clustering CURE according to different bidding durations. Section 5 exposes a new approach to label the clusters of bidders. Finally, Section 6 summarizes the results of this study and highlights important research directions as well.

2 Related Work

Many researchers utilized data clustering to examine auctions from different angles, such as studying the dynamics of auction prices and bidder behaviour. For instance, [13] introduced an approach to model and analyze the price formation as well as its dynamics to characterize the heterogeneity of the price formation process. The proposed functional objects represent the price process by accommodating the structure format of bidding data on eBay. Then, the

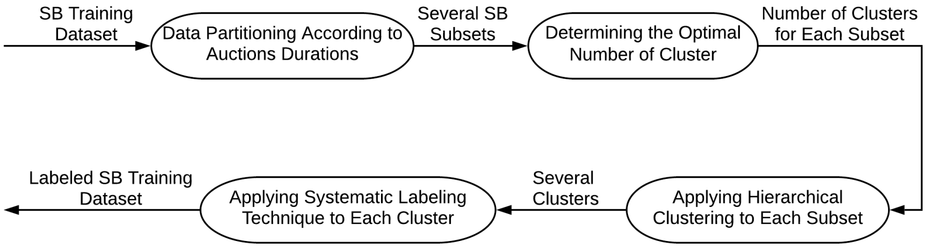
SB Training Dataset
Data Partitioning According to Auctions Durations
Several SB Subsets
Determining the Optimal Number of Cluster
Number of Clusters for Each Subset
Applying Hierarchical Clustering to Each Subset
Several Clusters
Applying Systematic Labeling Technique to Each Cluster
Labeled SB Training Dataset

Fig. 1. The Labeling Process of Shill Bidding Training Data

curve clustering is used to partition auctions by grouping similar price profiles. Finally, differential equations are used to specify the price of each group.

Another work [7] measured the similarity of bidder behaviour using specific attributes, such as bidder feedback rating, average increment difference and number of bids. Then, a centroid-based hierarchical clustering approach is presented to group similar bidders. Each produced cluster is then labeled manually according to the overall bidder behaviour in that cluster.

The study [14] suggested a SB detection model utilizing k-mean clustering technique. The latter groups similar buyers in one class to differentiate between general buyers and shill bidders. There are four features that represent a buyer: “how long the buyer has been in the auction”, “the times of buyer bids”, “the average response time of the buyer” and ‘the absolute average discrepancy of buyer bids”. Based on these features, k-mean classifies bidders into one of the two clusters: general buyers and shill bidders. However, we believe the second and third features do not really reflect SB since a buyer might be very interested in winning the auction. Besides, there are stronger patterns that highly identify SB. Since SB behaviour is somehow similar to real bidding [5], there is a possibility that some SB samples fell in the normal class.

In [4], the authors proposed a two-step clustering model to recognize bidding strategies. The hierarchical clustering is the first step to produce a dendrogram and an agglomeration schedule table to find the best number of clusters. The next step employs k-mean clustering to provide more details about the bidders’ strategies in each cluster. The experiment was operated on an outdated data (2003) collected from Taobao.com. According to the agglomeration coefficients, the optimal number of clusters is three: “early bidding strategy”, “snipe bidding strategy”, and the third one groups bidders that enter the auction early and remain for a long time. The first cluster has shown low values for the given features, which indicates that the bidders’ strategy is to enter and exit auctions early and participate infrequently. The second cluster has displayed high values for some features and low values for others. This illustrates that bidders enter and exit auctions late and rarely participate. The final cluster has administered the bidders' strategy where bidders enter early and stay for a long time and highly participate in the auctions.

[20] proposed a model based on hedonic regression and fuzzy logic expert system (FLES) to analyze bidder behaviour. The hedonic regression is used to select key variables that are passed to FLES to produce a knowledge base about the relationships between variables, like auction characteristics. Since the examined data have no relational information, k-mean is employed to obtain the minimum squared-error clusters. So, each training sample is classified to low, medium or high membership degree. The issue here is that the study is based on an outdated dataset (2004). Also, there is a potential for fabricating feedback ratings conducted between shill bidders and fraudulent bidder rings [7].

More recently, [12] applied k-mean clustering to categorize bidders' habits. The observations are obtained by the k-mean and then passed to the Baum-Welch algorithm and Hidden Markov Model. Three main clusters were suggested according to the values of the given features, which are low, medium and high cluster values. A bidder habit with values beyond these clusters values is considered as a fraudster. The experiment showed that only two simple features were given to identify the clusters: the number of auctions that a bidder participated in and the number of submitted bids by that bidder. Thus, if more features were considered to define the clusters, then the samples distribution on each cluster might be changed. As a result, this may influence the outcomes of the detection model. Also, the clustering is based on a dataset that is not adequately described, and only ten samples were used for explaining the results.

Lastly, [10] applied a hierarchical clustering to group users with similar bidding behaviour. The centroid linkage is used as the similarity measure. The described SB patterns were computed for all the bidders of each of the generated clusters. Then, the authors introduced a semi-automatic approach to label each cluster according to the general behaviour of users in that cluster and the weights of the fraud patterns.

3 Shill Bidding Overview

SB is a well-known auction fraud, and yet it is the most difficult to detect since it behaves similarly to normal bidding [5, 8]. The aim of this fraud is to increase the price or desirability of the auctioned product through imitation accounts and collusion with other users. In other words, shill bidders do not tend to win the auction but to increase the revenue of the seller. SB leads buyers to overpay for the items, and for high priced items, buyers will lose a substantial amount of money. As mentioned in [5], excessive SB could lead to a market failure. Thus, e-auctions may lose their credibility [5]. In fact, several sellers and their accomplices have been prosecuted due to SB activities, including:

- In 2007, a jewellery seller was accused of conducting SB fraud on eBay, and had to pay $400,000 for a settlement. Also, he and his employees were prevented from engaging in any online auctioning activities for four years [3].

[3] https://www.nytimes.com/2007/06/09/business/09auction.html

Table 1. SB Patterns and their Characteristics

Name	**Definition**	**Category**	**Source**	**Weight**
Bidder Tendency (BT)	Engages exclusively with few sellers instead of a diversified lot	Bidder	User history	0.5
Bidding Ratio (BR)	Participates more frequently to raise the auction price	Bid	Bidding period	0.7
Successive Outbidding (SO)	Successively outbids himself even though he is the current winner	Bid	Bidding period	0.7
Last Bidding (LB)	Becomes inactive at the last stage to avoid winning	Bid	Last bidding stage	0.5
Early Bidding (EB)	Tends to bid pretty early in the auction to get users attention	Bid	Early bidding stage	0.3
Winning Ratio (WR)	Participates a lot in many auctions but rarely wins any auctions	Bidder	Bidder history	0.7
Auction Bids (AB)	Tends to have a much higher number of bids than the average of bids in auctions selling the same product	Auction	Auction history	0.3
Auction Starting Price (ASP)	Offers a small starting price to attract genuine bidders	Auction	Auction history	0.3

- In 2010, a seller faced a £50,000 fine after being found outbidding himself on eBay. He claimed that: "*eBay let me open up the second account and I gave all my personal details and home address to do so.*" [4].
- In 2012, the online auction Trade Me had to pay $70,000 for each victim after the investigation discovered SB fraud conducted by a motor vehicle trader in Auckland. The fraud was carried out for one year, and caused a significant loss for the victims. Trade Me blocked this trader from using their site, and referred the case to the Commerce Commission for a further investigation [5].
- In 2014, a lawsuit was filled against Auction.com by VRG in California claiming that the website allowed SB. The bid of $5.4 million should have secured the property as the plaintiff declared, and yet the winning price was 2 million more. Auction.com was accused of helping the property loan holder, which is not fair for genuine bidders. The California state passed a law on July 1, 2015, which requests the property auctioneers to reveal bids they submit on a seller's behalf [6]. The spokeswoman for the California Association of Realtors said: "*To the best of our knowledge, we are the only state to pass this sort of legislation, even though we believe shill bidding to be prevalent all over the country.*"

By examining throughly the literature on the SB strategies [6, 8, 17], we compiled in Table 1, the most relevant SB patterns. Each pattern, which is a training

[4] http://www.dailymail.co.uk/news/article-1267410/Ebay-seller-faces-fine-bidding-items-raise-prices.html

[5] https://www.trademe.co.nz/trust-safety/2012/9/29/shill-bidding

[6] https://nypost.com/2014/12/25/lawsuit-targets-googles-auction-com

feature, represents a unique aspect of the bidding behaviour in auctions. The feature uniqueness will lead to an improved predictive peformance.

4 Production of Shill Bidding Data from Online Auctions

4.1 Auction Data Extraction and Preprocessing

To obtain a reliable SB training dataset, it must be built from actual auction data. Nevertheless, producing high quality auction data is itself a burdensome operation due to the difficulty of collecting data from auction sites on one hand, and the challenging task of preprocessing the raw data on the other hand. The latter consumes a significant time and effort, around 60% to 80% of the entire workload [15]. In our technical study [1] , we employed the professional scraper Octopars [7] to collect a large number of auctions for one of the most popular products on eBay. The extracted dataset contains all the information related to auctions, bids and bidders. We crawled completed auctions of the iPhone 7 for three months (March to June 2017). We chose iPhone 7 because it may have attracted malicious moneymakers due to the following facts:

- Its auctions attracted a large number of bids and bidders.
- It has a good price range with the average of $578.64 (US currency). Indeed, there is a direct relationship between SB fraud and the auction price [5].
- The bidding duration varies between 1 (20.57%), 3 (23.2%), 5 (16.23%), 7 (38.3%) and 10 (1.7%) days. In long duration, a dishonest bidder may easily mimic usual bidding behaviour [5]. However, as claimed in [3], fraudulent sellers may receive positive rating in short duration. Thus, we considered both durations.

Table 2 presents the statistics after preprocessing the scraped auction data. This operation was very time consuming as it required several manual operations [1]: 1) removing redundant and inconsistent records, and also records with missing bidder IDs; 2) merging several attributes into a single one; 3) converting the format of several attributes into a proper one; 4) assigning an ID to the auctions. For example, the two attributes Date and Time in each auction are converted into seconds; as an example 1 day and 10 day durations are converted into 86,400 and 864,000 seconds respectively.

4.2 SB Data Production

The algorithms to measure the SB patterns are presented in [17, 1]. Each metric is scaled to the range of [0, 1]; a high value indicates a suspicious bidding behaviour. We evaluated each metric against each bidder in each of the 807 auctions [1]. Therefore, we obtained a SB training dataset with a total of 6321 samples. A sample is described as a vector of 10 elements: the eight SB features along with Auction, and Bidder identification numbers. Once labeled, an instance will denote the conduct (normal or suspicious) of a bidder in a certain auction.

[7] https://www.octoparse.com

Table 2. Preprocessed Auctions of iPhone 7

No. of Auctions	807
No. of Records	15145
No. of Bidder IDs	1054
No. of Seller IDs	647
Avg. Winning Price	\$578.64
Avg. Bidding Duration	7
No. of Attributes	12

5 Hierarchical Clustering of Shill Bidding Data

Since the produced SB data are not labeled, data clustering, an unsupervised learning method, can be utilized to facilitate the labeling operation. Clustering is the process of isolating instances into K groups w. r. t. their similarities. The clustering techniques fall into one of the following categories: 1) Partitioning-based, such as K-medoids and K-means; 2) Hierarchical-based, such as BIRCH, GRIDCLUST and CURE; 3) Density-based, such as DBSCAN and DBCLASD; 4) Grid-based, such as STING and CLIQUE. In our work, we select agglomerative (bottom-top) hierarchical clustering where instances are arranged in the form of a tree structure using a proximity matrix.

5.1 CURE Overview

Among the hierarchical clustering methods, we choose CURE because it is highly performant in handling large-scale multi-dimensional datasets, determines non-spherical shapes of the clusters, and efficiently eliminates outliers [11]. Random sampling and partitioning techniques are utilized to handle the large-scale problem and to speed up the clustering operation. Each instance is first considered as an individual cluster, and then the cluster with the closest distance/similarity is merged into it in order to form a new cluster [11]. Two novel strategies have been introduced in CURE:

- **Representative Points (RPs)**, which are selected data points that define the cluster boundary. Instead of using a centroid, clusters are identified by a fixed number of RPs that are well dispersed. Clusters with the closest RPs are merged into one cluster. The multiplicity of RPs allow CURE to obtain arbitrary clustering shapes.
- **Constant shrinking factor (α)**, which is utilized to shrink the distance of RPs towards the centroid of the cluster. This factor reduces noise and outliers.

The worst case computational complexity of CURE is estimated to $O(N^2 \, log \, N)$, which is high when N is large (N is the number of instances) [18]. Since the SB data clustering is an offline operation, so the running time is not an issue. The only disadvantage of CURE is that the two parameters RP and α have to be set up by users.

To run the experiments, we utilize the Anaconda-Navigator environment for running Python 3, and incorporate the CURE program developed by Freddy Stein and Zach Levonian. CURE code (in Python) is available at GitHub.com[8].

5.2 SB Data Preparation

Since the bidding duration is used as a denominator for the two patterns Early Bidding and Last Bidding, the large gap between different durations greatly affects the computation results. The computed value of the fraud pattern for 10 days is far smaller than for 1 day. So, before applying CURE, we first partition the SB dataset into five subsets according to the five durations (1, 3, 5, 7 and 10 days) as presented in Table 3.

Table 3. SB Dataset Partitioning According to Bidding Durations

Subset	1 Day	3 Days	5 Days	7 Days	10 Days
No. of Auctions	166	187	131	309	14
No. of Instances	1289	1408	1060	2427	137

5.3 Optimal Number of Clusters

It is always difficult to decide about the optimal number of clusters of a training dataset. Besides, normal and shill bidder behaviour is somehow similar. Thus, determining the best number of clusters is an essential step to achieve a better interpretation for classifying similar SB instances [19]. There are several methods to address this problem, such as Elbow, Dendrogram, the Rule of Thumb and Silhouette. In our study, we employ the Silhouette method where each group is represented as a silhouette based on the separation between instances and the cluster's tightness [16]. The construction of a silhouette requires the clustering technique to generate the partitions and also to collect all approximates between instances [16]. K-mean clustering algorithm has been successfully utilized for this task due to its simplicity and effectiveness [19]. Consequently, we apply K-mean to estimate the number of clusters for each of the five SB subsets.

Next, for each subset, we examine the silhouette scores of 19 clusters (2 to 20 clusters), and choose the best number based on the best silhouette score. We have noticed that once the peek (denoting the optimal number) of the silhouette score is reached on a certain number of clusters, the silhouette score gradually decreases with the increasing of the number of clusters. In Figure 2, we give an example for the 7-day bidding duration for which the optimal number of clusters is eight since the highest silhouette score (0.4669) is obtained on that number. In Table 4, we expose the best number of clusters for each of the five SB subsets. The total number of produced clusters is 29.

[8] https://github.com/levoniaz/python-cure-implementation/blob/master/cure.py

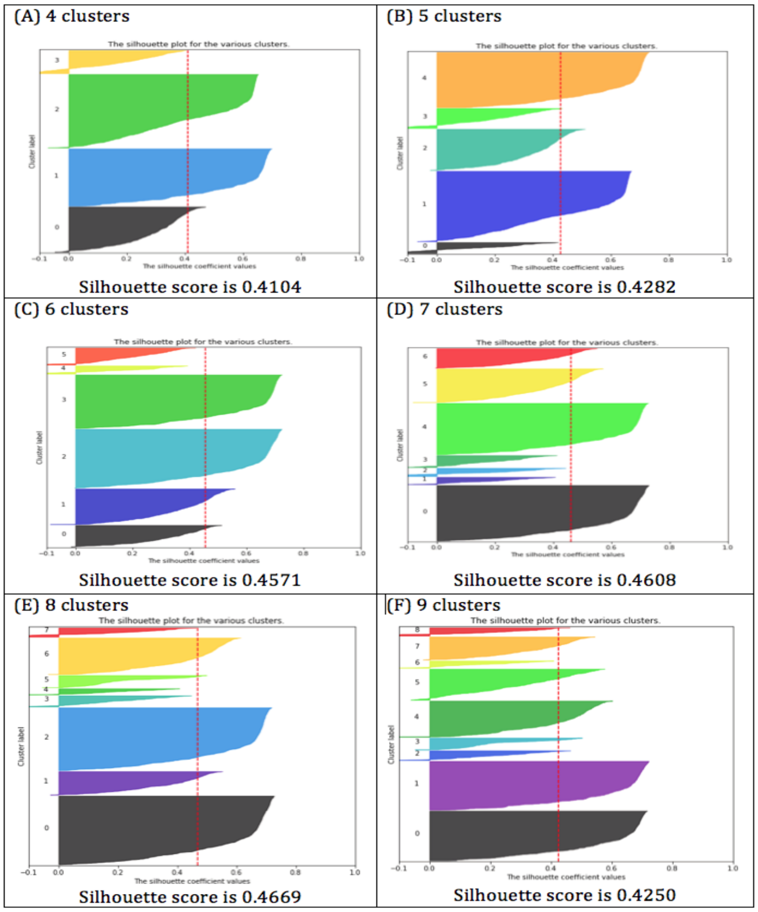

Fig. 2. Optimal number of clusters for 7 day bidding duration. Silhouette score is examined 19 times. We show the top Silhouette scores.

5.4 Cluster Generation

CURE has three parameters that need to be setup: representative points (RPs), shrinking factor (α) and optimal number of clusters. Based on the results of silhouette, we have obtained the optimal number of clusters for each of the five SB subsets. The two parameters RPs and α are defined by selecting the configuration that provides the best instance distribution among the specified clusters. Thus, CURE is applied with different values of RPs and α starting from the default values (5 for RPs and 0.1 for α). We throughly conducted trial-and-error experiments for all the clusters (29 clusters in total) of the five SB subsets to determine the best values of the parameters (Table 4). The best parameters' values are selected based on the best distribution of a subset population between the defined clusters. As an example, in Table 5, we present the results for the eight clusters that we have generated previously for the 7-day bidding duration subset. The best value configuration is shown in bold. Each cluster consists of users with similar bidding behaviour. As we can see, the clusters 4, 7 and 8 have very few bidders, which are most probably outliers i.e. suspiciuous.

Table 4. Optimal Number of Clusters and Optimal CURE Parameters

SB Subset	**No. of Samples**	**No. of Clusters**	**Silhouette Score**	**RPs**	α
1 Day	1289	7	0.4597	5	0.05
3 Days	1408	7	0.4672	5	0.01
5 Days	1060	5	0.4758	5	0.05
7 Days	2427	8	0.4669	10	0.001
10 Days	137	2	0.5549	5	0.1

Table 5. CURE Clustering for 7-day Bidding Duration (8 Clusters)

RP	α	**Cl.#1**	**Cl.#2**	**Cl.#3**	**Cl.#4**	**Cl.#5**	**Cl.#6**	**Cl.#7**	**Cl.#8**
5	0.1	136	1438	1	2	657	190	2	1
5	0.05	657	328	2	1408	1	28	2	1
5	0.01	1438	640	1	1	17	1	1	328
5	0.001	21	166	2	1	2	25	2209	1
10	0.1	2	1	657	1410	22	8	137	190
10	0.05	2	133	2	1	2	31	2066	190
10	0.01	1	1	1	1	135	31	2067	190
10	**0.001**	**189**	**654**	**1410**	**3**	**137**	**31**	**1**	**2**

6 Labeling Shill Bidding Data

In algorithm 1, we show the steps to label the bidders of a given cluster. A cluster belongs to a certain SB subset. As shown in Table 6, for each subset, we first compute the mean and STandard Deviation (STD) of each fraud pattern for all the instances in that subset. Then, we compute the average of the means (Avg. Means) and average of the STDs (Avg. STDs) of all the patterns for that subset. We consider the value of $(Avg.Means + \frac{1}{2}Avg.STDs)$ since it produces the best decision line that separates between normal and suspicious instances as depicted in Figure 3. Then, we calculate the average of the means of all the patterns for the cluster. So, if the average mean of the cluster is greater than the decision line of the subset, then instances are labeled as suspicious (1) in that cluster, otherwise, they are labeled normal (0).

Table 6. Characteristics of SB Subsets

Subset	1 Day	3 Days	5 Days	7 Days	10 Days
Mean of each pattern per subset					
BT	0.1434	0.1394	0.1419	0.1455	0.1162
BR	0.1287	0.1328	0.1235	0.1273	0.1021
SO	0.0996	0.1047	0.0872	0.1149	0.0620
LB	0.4624	0.4511	0.4676	0.4678	0.4746
EB	0.4314	0.4192	0.4318	0.4348	0.4575
WR	0.3812	0.3718	0.3810	0.3533	0.3496
AB	0.2120	0.1936	0.2403	0.2567	0.2926
ASP	0.5007	0.4301	0.4478	0.4801	0.7123
Avg. Means	0.2949	0.2802	0.2901	0.2975	0.3208
STD of each pattern per subset					
BT	0.1973	0.1884	0.1984	0.2019	0.1811
BR	0.1246	0.1330	0.1243	0.1377	0.1165
SO	0.2764	0.2811	0.2583	0.2917	0.2215
LB	0.3773	0.3753	0.3917	0.3783	0.3931
EB	0.3775	0.3742	0.3921	0.3802	0.3968
WR	0.4356	0.4373	0.4402	0.4345	0.4398
AB	0.2323	0.2426	0.2646	0.2658	0.2575
ASP	0.4931	0.4831	0.4863	0.4908	0.4510
Avg. STDs	0.3142	0.3143	0.3194	0.3226	0.3071

Algorithm 1 : Labeling Bidders in a Cluster

```
Require: AvgMeans and AvgSTDs of the corresponding SB subset
Compute $MeanCluster$
if ($MeanCluster \geq (AvgMeans + \frac{AvgSTDs}{2})$) then
    for x=1 to $NumberBiddersCluster$ do
        $LabelBidder_x = 1$ (Suspicious)
    end for
else
    for x=1 to $NumberBiddersCluster$ do
        $LabelBidder_x = 0$ (Normal)
    end for
end if
```

To validate our approach, we choose randomly 5 auctions among the 7-day duration auctions (in total 309) and select randomly one bidder in each auction (Table 7). As we can observe from this table, the shill bidding instances were successfully labeled by our approach. For example, bidder "g***r" has 4 fraud patterns with very high values; among them 2 have a high weight and 1 a medium weight. Therefore, the activity of this bidder in auction ID # 2370 is suspicious. On the other hand, the bidder "k***a" has all his fraud patterns with very low values; this indicates that this bidder behaved normally in the auction ID# 900. All these results are consistent with the labels produced by our approach.

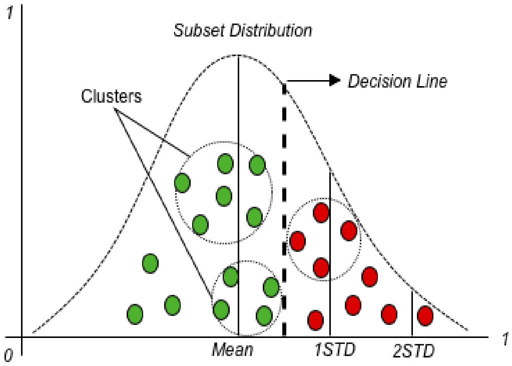

Fig. 3. Decision line of a subset and its labeled clusters

Table 7. SB Instances and their Labels

AuctionID	1009	900	2370	432	1370
BidderID	z***z	k***a	g***r	0***0	o***-
BT	0.75	0.4705	0.8333	0.5	0.04615
BR	0.3461	0.3076	0.2	0.3333	0.0857
SO	1	0	1	0	0.5
LB	0.5667	0.1909	0.0350	0.2199	0.2966
EB	0.5409	0.1909	0.0239	0.0043	0.2060
WR	0.75	0.4	1	0.5	0
AB	0	0	0.3333	0	0.0526
ASP	0	0	0.9935	0	0
Generated Label	**1**	**0**	**1**	**0**	**0**

Table 8 provides all the final results of the labeling task of our SB training dataset. There are 5646 instances categorized as normal and 675 instances as suspicious.

Table 8. Final Results of Instance Labeling

SB Subset	1 Day	3 Days	5 Days	7 Days	10 Days	Total
No. of Normal Instances	1135	1303	975	2098	135	5646
No. of Suspicious Instances	154	105	85	329	2	675

7 Conclusion and Future Work

There are limited classification studies on the SB fraud due to the difficulty of producing training data on one hand and labeling multi-dimensional instances on the other hand. Our aim in this paper is to effectively label SB instances based on the hierarchical clustering CURE that showed a remarkable capability for partitioning the online behaviour of bidders. First, we divide the SB dataset into several subsets according to the different bidding durations of the auctions that we scraped from eBay. Then, we efficiently partition each SB subset into clusters of users with similar bidding behaviour. At last, we apply a systematic labeling approach to each cluster to classify bidders into normal or suspicious.

In the following, we highlight two important research directions:

- The generated SB dataset is highly imbalanced, which will negatively impact the performance of classifiers as demonstrated in numerous studies such as [9]. The decision boundary of the fraud classifiers will be biased towards the normal class, which means suspicious bidders will be poorly detected. Handling the class imbalance problem is a continuous area of study [21]. In our research, we will investigate this problem by testing different types of techniques, such as data sampling and cost-sensitive learning, to determine the most suitable technique for our SB dataset.

- Ensemble learning has produced reliable performance for many practical applications. The goals defined by ensemble learning are lowering the model's error ratio, avoiding the overfitting problem, and reducing the bias and variance errors. The most common ensemble methods are Boosting and Bootstrap Aggregation (Bagging). Thus, we will employ ensemble learning to develop a robust SB detection model, and examine the most fitting ensemble strategy for our SB dataset.

Acknowledgments

The first author would like to thank the Saudi Arabian Cultural Bureau in Canada and the Umm Al-Qura University for their financial support.

References

1. Alzahrani, A., Sadaoui, S.: Scraping and preprocessing commercial auction data for fraud classification. arXiv preprint arXiv:1806.00656 (2018)
2. Center, I.C.C.: 2015 internet crime report. In: 2015 IC3 Report. IC3 (2016)
3. Chang, J.S., Chang, W.H.: Analysis of fraudulent behavior strategies in online auctions for detecting latent fraudsters. Electronic Commerce Research and Applications **13**(2), 79–97 (2014)
4. Cui, X., Lai, V.S.: Bidding strategies in online single-unit auctions: Their impact and satisfaction. Information & Management **50**(6), 314–321 (2013)

5. Dong, F., Shatz, S.M., Xu, H.: Combating online in-auction fraud: Clues, techniques and challenges. Computer Science Review **3**(4), 245–258 (2009)
6. Dong, F., Shatz, S.M., Xu, H.: Reasoning under uncertainty for shill detection in online auctions using dempster–shafer theory. International Journal of Software Engineering and Knowledge Engineering **20**(07), 943–973 (2010)
7. Ford, B.J., Xu, H., Valova, I.: Identifying suspicious bidders utilizing hierarchical clustering and decision trees. In: IC-AI. pp. 195–201 (2010)
8. Ford, B.J., Xu, H., Valova, I.: A real-time self-adaptive classifier for identifying suspicious bidders in online auctions. The Computer Journal **56**(5), 646–663 (2012)
9. Ganguly, S., Sadaoui, S.: Classification of imbalanced auction fraud data. In: Canadian Conference on Artificial Intelligence. pp. 84–89. Springer (2017)
10. Ganguly, S., Sadaoui, S.: Online detection of shill bidding fraud based on machine learning techniques. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. pp. 303–314. Springer (2018)
11. Guha, S., Rastogi, R., Shim, K.: Cure: an efficient clustering algorithm for large databases. In: ACM Sigmod Record. vol. 27, pp. 73–84. ACM (1998)
12. Gupta, P., Mundra, A.: Online in-auction fraud detection using online hybrid model. In: Computing, Communication & Automation (ICCCA), 2015 International Conference on. pp. 901–907. IEEE (2015)
13. Jank, W., Shmueli, G.: Studying heterogeneity of price evolution in ebay auctions via functional clustering. Handbook of information systems series: Business computing pp. 237–261 (2009)
14. Lei, B., Zhang, H., Chen, H., Liu, L., Wang, D.: A k-means clustering based algorithm for shill bidding recognition in online auction. In: Control and Decision Conference (CCDC), 2012 24th Chinese. pp. 939–943. IEEE (2012)
15. Ravisankar, P., Ravi, V., Rao, G.R., Bose, I.: Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems **50**(2), 491–500 (2011)
16. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
17. Sadaoui, S., Wang, X.: A dynamic stage-based fraud monitoring framework of multiple live auctions. Applied Intelligence **46**(1), 197–213 (2017)
18. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. Annals of Data Science **2**(2), 165–193 (2015)
19. Yu, H., Liu, Z., Wang, G.: An automatic method to determine the number of clusters using decision-theoretic rough set. International Journal of Approximate Reasoning **55**(1), 101–115 (2014)
20. Zhang, J., Prater, E.L., Lipkin, I.: Feedback reviews and bidding in online auctions: An integrated hedonic regression and fuzzy logic expert system approach. Decision Support Systems **55**(4), 894–902 (2013)
21. Zhang, S., Sadaoui, S., Mouhoub, M.: An empirical analysis of imbalanced data classification. Computer and Information Science **8**(1), 151–162 (2015)