A Survey on Online Auction using Data Mining Techniques

Gomathi Sekar[1], Srinidhi Karthikeyan[2], Thenmozhi Baskaran[3], C. Jerin Mahibha[4]
[1,2&3]*Final Year UG Student* [4]*Sr.Assistent Professor*
Computer Science and Engineering
Meenakshi Sundararajan Engineering College
Chennai, India

Abstract

Thousands of people take part in Internet auctions every day, bidding on items from different places. Buyers and sellers alike benefit from the great opportunities that online auctions provide, but these auctions also provide criminals the opportunity to perpetrate fraud. The online auction environment is full of shill and fraud bidders and the losses incurred because of these offending activities are huge. This problem of Shill bidding and fraudulence can be solved by applying Data mining and Machine learning techniques. So, this paper will help us get an idea about different auction types and different data mining and ML algorithms that can be used to prevent the forgery that is prevalent in online auctions

Keywords - *Online auction; Data mining ;Fraud detection*

I. INTRODUCTION

Online Auction has significantly increased the variety of goods and services which can be bought and sold using the auction mechanisms. Online auction have broken all the barriers that were inhibiting the users for accessing the auction like geographical locations, time and a small target audience. Making auctions online, the numbers of users participating in the auctions have dramatically increased over time. It functions as, the bidder first starts quoting a smaller sum and then it gets increased over other bidders quoting higher amounts in order to win that particular good. The time limits for the auctions differ based on the domain where the auction takes place.

The main benefit of online auction over physical auction is that the user from different parts of the world can participate in the auction and then they are shipped globally. The objects sold can a single product of collection of many. Not all the users who are accessing the system are genuine and it is necessary to identify those who are not legitimate and try to increase theactual amount of the goods being sold.

Popular online auction sites like eBay protects the legitimate user from these swindlers by having a list called bidder block lists. If the user is from the list they are blocked from participating from the bidding. There are greater chances of selling pirated and stolen products.

The various Data Mining and Machine Learning techniques can be used to make the online auction safer and flexible for the user. Techniques like ID3, C4.5, C5.0, CART, Neural networks etc. are studied and analyzed here. This paper also present different other domains where these techniques were used successfully in other domains.

The rest of the paper is organized as follows. Section 2 presents an overview of different Data Mining techniques . Section 3 gives an idea about how CART can be used in different domains. Section 4 depicts how C4.5 can be used. Section 5 gives an outline of usage of other Machine Learning techniques used in acution systems. Section 6 concludes the work.

II. LITERATURE SURVEY

A. Overview Of Different Data Mining Techniques

There are different types of data mining techniques used namely Asssociation, Classification,Clustering, Prediction, Sequential patterns,Decision trees.

1. Association

It is a well known and simple data mining technique where, a pattern is identified considering a relationship between two transactions and hence called as Relation Technique. It is generally used in marketing and purchasing domains and generally used for market basket analysis. It is basically to tempt buyers with other products that are frequently bought items. For Example, when a user buys bread placing the jam and butter on the same rack might tempt the user to buy them together, thereby increasing the sales.

2. Classification

This is basically used to classify the items into predefined groups or classes using mathematical

techniques like decision trees, statistics etc. For example we will be able to use medical data to classify them as Diabetic and non-diabetic patients so that treatment can be made even more efficient by considering their medical conditions and groups they are placed in.

3.Clustering

It is an unsupervised ML technique to make data into groups based on similar characteristics to make then data meaningful. But in classification the objects are assigned to predefined classes. For example, a company might cluster customers based on their information like location, sales etc. to find the groups to target the right users for selling their products.

4. Prediction

It is a technique that identifies the relationship between dependent and independent variables. In simple , deriving the relationship between the thing we already know and the thing needed to predict the future.For example, when we have the previous records of volcanic data for the past few decades we will be able to predict the characteristics of the volcano when it is about to erupt. So when those charcteristcs are identified we will be able to do preventive deeds like clearing people in the vicinity etc.

5. Sequential Patterns

It is simply identify similar patterns or regular actions so that we will be able to understand what we should sell the user based on their previous transactions. For example, user's purchase details can be identified to recommend them other products similar to their previous purchase.

6. Decision trees

Decision tree is the most common data mining technique that is used these days. Here the decision tree is constructed based on the training data set given to the machine during the learning phase. Then the tree is used make the decision for the problems proposed based on the tree constructed.

III. USING CART TECHNIQUE

In[5], Different sampling techniques are used to analyze the performance of CART and c5.0. Considering confusion matrix as a parameter the different sampling techniques like simple random, Systematic random and Stratified random sample is used.

In Simple sampling technique, the samples are randomly selected from the dataset of some size, so that all the instances will have the same probability, thereby reducing biases in the data being used. The next type is the systematic sampling technique, the instances are selected at random intervals from the dataset. But when all the data selected are homogenous then, the result may not be accurate. The last type of sampling is Stratified random sampling in which the entire dataset is divided into groups called strata by using the simple random sampling method and the instances are got from different stratum. This can guarantee that there is only negligible bias in the selected data and can provide greater accuracy in the results.

The conclusion is that the type of sampling technique plays a major role in determining the accuracy of the algorithm. the confusion matrix shows a greater difference in the accuracy of the results when the stratified random sampling technique is used.

In[11], Merbouha compares various classification techniques where he concludes that CART has drawbacks like the tests are always binary, it depends only on the Gini impurity values to decide and the technique of pruning the trees are complex. So, we can understand other techniques in the next section.

IV. USAGE OF C4.5 TECHNIQUE

In[6], several attributes for providing loans are considered and a decision tree is built using the C4.5 algorithm. Performance parameters like accuracy, recall, accuracy are calculated with different partitioning of the datasets like 90%:10%, 80%:20%, 70%: 30 %, 60 %: 40 are analysed.

It was found that the 80% training data and 20% of test data gives the highest performance. It is also important to note that not only the partition is important but also the quality of the dataset is important to improve the performance of the algorithm. It was concluded that Pruning method used in C4.5 is not that efficient and missing values cannot be addressed in C4.5.

In[7],Soil quality of a province is predicted using the C4.5 algorithm.A decision tree is constructed using C4.5 with several attributes of soil like organic matter and nitrogen, sodium, potassium content. The information gain and entropy for the different attributes are calculated. The maximum property of the information gain rate to divide dataset for the first time, to generate the decision tree.

Later the tree is pruned. It was found that the accuracy of the results were 92.71% when compared with the original results.

V.USING OTHER DATA MINING AND MACHINE LEARNING TECHNIQUE

In[1], Benjamin et al. proposed the data mining technique incremental neural network approach to

identify the suspicious bidder in online auction. Here, the neural network will make use of training dataset which contains the historical auction data and it will be revised with new data every time .Using that dataset it will differentiate the behavior of bidder.This will help to decide which person is the suspicious bidder and produces more accurate results. However, in this approach we need to reinitialise a network every time when new data enters so, it is very expensive process and also time consuming method.

In[2], Fei Dong et al. proposed two main techniques are model checking technique and Dempster-Shafer theory. The model checking technique used to detect suspicious shill bidders in auction based on bidding pattern-based temporal formulas. The Dempster-Shafer theory will justifies the shill suspects by collecting the large amount of evidences about bidders and it will combine with new evidences .Based on the evidences it will give shill score to each bidder which will classifies the trusted and suspect bidder. Using Dempster-Shafer theory gives more accurate result but using Bayesian inference method it improves the system performance.

In[3],Wen-Hsi Chang et al. proposed Clustering technique for discovering the fake transaction records in auction. The X-means(extended version of K-means)clustering technique analyze the behavioral changes of the fraudsters and identify them in the earlier phase. This paper also used C4.5 algorithm which was applied for explaining the rules of the labeled fraudulent behaviors. These techniques identifies fraudulent behavior in earlier phase and along with that apply different instance-based learning algorithms to refine the behavioral changes of fraudsters.

In[8], Govardhani analyzed et al. used various machine algorithms such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). He conducted experiments with the datasets and noticed the accuracies between them. Then used pandas and various packages to load datasets and performed the numerical calculations. Both the ANN and SVM gave better results when compared to the other machine learning algorithms. He proposed the Parameter tuning (Grid Search technique) which is used to find the best quality parameters. The Grid Search will test several combinations of quality parameters and and will return the best selection that gives best accuracy. The proposed system is applied to both the ANN and SVM algorithms but the ANN has more accuracy than the SVM.

In[9], Nadeem Akhtar et al proposed the system which uses dynamic packet filtering and buffering to enable the effective bulk recording of large traffic streams. Then he the congestion is encountered by the system. So then he discussed about the concept of Very Fast Decision Tree Learner(VFDT) used for Data mining based on Hoeffding Algorithm which could handle data to tune million of packets.implemented the Bitmap indexing in data mining with Fastbit ,which is the first robust end-to-end system. But It takes lesser time to mine data in lesser time than it takes to input them from disk. And also it does not store data in the main memory and hence storage space is required only for that tree and associated statistics. He compared VFDT with C4.5 and it shows that C4.5 is more accurate than VFDT. Finally it can be incorporated with DSMS tool like TelegraphyCQ.

In [10],Kiran Drogar used several machine learning algorithms for Stock market predictions. The different data mining models have been developed and their performances are compared. He proposed the simple architecture and it involves feature extraction from the given dataset, supervised classification of training dataset and test dataset, and result evaluation. For the comparative study of Supervised learning algorithms, he calculated the accuracy and F-measure. By the comparison of accuracies of different algorithms he concludes that Naive Bayes performs best for small datasets and Random Forest performs best for large datasets. He concludes that reduction in the technical indicators reduces the accuracy of each algorithm.

VI. COMPARISION OF PERFORMANCE OF DIFFERENT CLASSIFICATION ALGORTIHMS

In[11],Merbouha compares various classification algorithms. Firstly the comparison between ID3 and C4.5 , where he concludes C4.5 is better than ID3, because the performance parameters like accuracy and execution time is better for c4.5. Secondly, the comparison between C4.5 and C5.0 implies that C5.0 has more advantages like Scalability and use of unordered rule sets, applicable values were improved. It also improves predictive accuracy.Thirdly, CART is compared with C5.0 , which infers that there are both pros and cons in the usage of them but the speed and accuracy of C5.0 is better than CART.

VI.CONCLUSION

We come to a conclusion after going through several data mininig algoritms especially classiifcation algorithms that C4.5 algortihm can be used to prevent frauds and shill bidding in online auction systems to produce more accurate results.Even though the algorithms have some drawbacks like lacking accuarcy when used with smaller datasets and smaller variation

in that data having greater impact on the tree, it has produced nearly successful results in various domains and simple to implement without any complications.Moreover the Accuracy of results are not only based on the algorithm used but several other factors like partition of test and trainig set, quality of the data in dataset etc. So, by taking all these factors into considertion we will be able to get better results using C4.5 algorthm .

REFERENCES

[1] Benjamin J. Ford, Haiping Xu∗ and Iren Valova. "A Real-Time Self-Adaptive Classifier for Identifying Suspicious Bidders in Online Auctions", The Computer Journal, Vol. 56 No. 5, 2013.

[2] Fei Dong1 , Sol M. Shatz1 , and Haiping Xu2. "Inference of Online Auction Shills Using Dempster-Shafer Theory", Sixth International Conference on Information Technology: New Generations,2009.

[3] Wen-HsiChang,Jau-ShienChang."Using Clustering Techniques to Analyze Fraudulent Behavior Changes in Online Auctions",International Conference on Networking and Information Technology,2010.

[4] Prashant K. Khobragade and Latesh G. Malik."Data Generation and Analysis for Digital Forensic Application using Data mining", Fourth International Conference on Communication Systems and Network Technologies, 2014.

[5] M.Balamurugan and S. Kannan, "Performance analysis of cart and C5.0 using sampling techniques," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, 2016, pp. 72-75.

[6] R.K.Amin, Indwiarti and Y. Sibaroni, "Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region)," *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, Nusa Dua, 2015, pp. 75-80. doi: 10.1109/ICoICT.2015.7231400.

[7] L.Dongming, L. Yan, Y. Chao, L. Chaoran, L. Huan and Z. Lijuan, "The application of decision tree C4.5 algorithm to soil quality grade forecasting model," 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI), Wuhan, 2016, pp. 552-555. doi: 10.1109/CCI.2016.7778985

[8] L.N.Pondhu and G. Kummari, "Performance Analysis of Machine Learning Algorithms for Gender Classification," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, 2018, pp. 1626-1628.

[9] M.A.Qadeer, N. Akhtar and F. Khan, "Comparison of Tools for Data Mining and Retrieval in High Volume Data Stream," *2009 Second International Workshop on Knowledge Discovery and Data Mining*, Moscow, 2009, pp. 252-255.

[10] I.Kumar, K. Dogra, C. Utreja and P. Yadav, "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, 2018,pp.1003-1007. doi: 10.1109/ICICCT.2018.8473214

[11] Badr HSSINA, Abdelkarim MERBOUHA, Hanane EZZIKOURI and Mohammed ERRITALI, "A comparative study of decision tree ID3 and C4.5" International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Advances in Vehicular Ad Hoc Networking and Applications 2014.