Prediction of Click Frauds in Mobile Advertising

Mayank Taneja, Kavyanshi Garg, Archana Purwar, Samarth Sharma
Department of Computer Science Engineering and Information Technology
Jaypee Institute of Information Technology, Noida, India.
tanejamayank1@gmail.com

***Abstract*- Click fraud represents a serious drain on advertising budgets and can seriously harm the viability of the internet advertising market. This paper proposes a novel framework for prediction of click fraud in mobile advertising which consists of feature selection using Recursive Feature Elimination (RFE) and classification through Hellinger Distance Decision Tree (HDDT).RFE is chosen for the feature selection as it has given better results as compared to wrapper approach when evaluated using different classifiers. HDDT is also selected as classifier to deal with class imbalance issue present in the data set. The efficiency of proposed framework is investigated on the data set provided by Buzzcity and compared with J48, Rep Tree, logitboost, and random forest. Results show that accuracy achieved by proposed framework is 64.07 % which is best as compared to existing methods under study.**

Keywords- Recursive Feature Elimination (RFE); Hellinger Distance Decision Tree (HDDT); Click Fraud Detection

I. INTRODUCTION

Fraud against online advertisements is a problem in recent years that has gained a lot of attention because it is a serious threat to the advertising industry. The main coordinator in the advertising industry is the advertising commissioner, acting as a broker between advertisers and content publishers .An advertiser plans a budget, provides the commissioner with advertisements, and agrees on a commission for every customer action. A content publisher contracts with the commissioner to display advertisements on their websites, and gets commissions based on the traffic it drives to the advertisers. This model, however, may incentivise dishonest publishers to generate illegitimate clicks on their sites that is popularly known as **click fraud**. Click fraud degrades the reliability of online advertising systems and, if not kept under control, can lead to a contraction of the advertising market in the long term. Thus, a reliable click fraud detection system is needed to help the commissioners proactively prevent click fraud and assure their advertisers that their dollars have been well spent.

In 2007, Google Inc. [1] estimated that 10 percent of clicks on advertisements in their AdWords program were not legitimate user clicks, which translates into a one billion USD yearly revenue loss after filtering out these clicks. Due to the limitations of mobile phone networks, new methods are needed to evaluate the fraudulent behaviour of advertising publishers.

The proposed framework that uses combination of RFE and HDDT is able to predict the fraudulent behaviour as okay, fraudulent and observation.

Rest of the paper is organized into 5 sections. First section gives the Literature review of the work under study. Proposed System is described in section 2. Section 3 and section 4 propound experimental framework and results respectively. Last section shows the conclusion.

II. LITERATURE SURVEY

Click fraud detection is not an unaddressed problem but could largely be considered a secret art as little open research exists on the topic. The problem was structured around three key challenges namely high cost of both False Positive's and False Negative's with respect to legitimacy of the clicks , minority-class and multi-class issues , and training many models at scale. A simple MapReduce [1] framework for training data models at scale for feature selection was presented. The major features selected were natural language features, string based features, structural features, page-Type features, and crawl-based features. Hierarchical multi-class classification was proposed to deal with multiclass issue. Daniel Berrar has proposed a model using random forest to predict the status of a publisher based on its individual click profile [2]. The author had analysed the click patterns associated with 3081 publishers of online mobile advertisements (as per the given data set). The status of these publishers was known to be either fraudulent, under observation, or honest. He has obtained an accuracy of 49.99% on the blinded validation set and 42.00% on the blinded test set. It was realised that we will require as much information from the dataset to obtain higher precision therefore the attributes used to predict fraud clicks in mobile advertisements was elaborated to informative features. More importantly, as repeatedly said by experts feature engineering is the key for good performance. A lot of derived attributes had to be built which played a critical role in improving the performance. This particular challenge also has involved multiclass as well as class imbalance issues. Hence, Shivshankar and Manoj explained the concept of Hierarchical committee machines [8] that combined a set of diverse cost sensitive classifiers built using different set of attributes (datasets). Each committee machine was used to combine the responses of diverse

978-1-4673-7948-9/15/$31.00 ©2015 IEEE

classifiers on datasets that included different sets of derived attributes. The diverse classifiers included J48, K Star, LAD tree, AODE and REP tree. Finally, a committee machine was used to combine the responses from individual committee machines that were built on different datasets.

The methods proposed in the literature of click fraud had class imbalance issues [13] and also the features were not correctly selected .Hence, this paper proposes a novel framework comprising of Recursive Feature Elimination as feature selection method and Hellinger Distance based Decision tree as classifier to efficiently identify the fraudulent publishers from the data set.

III. PROPOSED SYSTEM

In this paper, we have developed a novel framework for prediction of click frauds in mobile advertisements. It consists of feature selection using Recursive Feature elimination and classification through Hellinger Distance Decision tree. The following subsections give the detailed description of RFE and HDDT classifier.

A. Recursive feature elimination (RFE)

RFE algorithm is a greedy method that only hopes to find the best possible combination for classification. Although, RFE does not necessarily return an optimal solution due to its greedy nature, it is distinguished as one of the most effective methods of feature selection as the weakest feature is removed after each step rather than selecting the strongest feature.

RFE algorithm can be broken into two steps. Firstly, an initial solution is obtained by finding weakest of the features by ranking them according to the correlation with the features and classes present. This solution is the set of features remaining after removing the weakest one .In the second step; we do the process recursively till negative correlation value is found. Thus the number of features are minimised to improve their quality.

B. Hellinger Distance Decision Tree (HDDT)

This is a decision tree classifier based on Hellinger distance [12], which is a measure to find the similarity between two probabilistic distributions. Hellinger distance decision tree is robust and insensitive to skewness. As it is to find out the overlap between two statistical distributions, overestimation of overlap will result in the low accuracy of the classification. To obtain accurate result, we have to assume optimum numbers of partitions describing the overlap

For two discrete probability distributions P and Q, their Hellinger distance [10] is calculated using (1).

$$\boldsymbol{H(P,Q)} = \frac{\mathbf{1}}{\sqrt{\mathbf{2}}}\sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{k}}\left(\sqrt{\boldsymbol{p_i}} - \sqrt{\boldsymbol{q_i}}\right)^{\mathbf{2}}} \qquad (1)$$

Where P and Q are the selected features; p_i and $q_i$are the discrete values of distribution P and Q. And k is the total number of tuples.

This Hellinger distance has to follow the property that sum of the discrete values should be non-negative .We have used Hellinger distance between two distributions as criteria to convert the attributes to a decision tree based on the dataset obtained after applying feature selection technique (RFE) .

The data set under experimental work is skewed because it has “OK” class as the majority in the data set. Hence, we have chosen HDDT as classifier to address the class imbalance issue. HDDT uses Hellinger distance as the decision tree splitting criteria rather than using the traditional criteria such as entropy and log likelihood.

The performance of proposed system is evaluated using accuracy [11] as popular metric that refers to the ability of system to correctly predict the class label of new or unseen data. Accuracy is computed using (2).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

where,

True positives (TP) = no. of correct classifications predicted as yes (or positive).

True negatives (TN) = no. of correct classifications predicted as no (or negative)

False positive (FP) = no. of examples that are incorrectly predicted as yes (or positive) when it is actually no (negative).

False negative (FN) = no. of examples that are incorrectly predicted as no when it is actually yes

IV. EXPERIMENTAL SETUP

A. Data set description

In this section, proposed model is validated on data set provided by Buzzcity [9] to analyse the behaviour of publishers and classify fraudulent publishers from legitimate publishers. This data set was divided into two data sets. First dataset contains the information of 3081 publishers with 4 attributes as publisher ID, account no, address and status. The second data set contains 3173834 click details on advertisements by all of the above publishers. These details include 9 attributes namely click id, publisher id, ip address, agent, category, TimeAt, country, campaign id and referrer. Dataset is described by three distinct classes as status in publisher information i, e. "OK", "Fraud" and "Observation".

B. Data preprocessing and feature selection

Since the attributes of provided two data sets cannot be used directly for classification with sufficient accuracy, data sets were processed to extract features that add value to the data set used under study. In order to obtain the features that can give us the maximum information, we have extracted features from the existing attributes .To create features, we select each attribute separately and model the partner's click pattern with respect to that attribute .Many parameters such as AvgClickPerMin, and AvgClickPer5Min for TimeAt attribute shown in Table 1 are created . These Parameters are identified using statistical measures like maximum, average, skewness, variance and ratio. The data set created after this consists of these 21 extracted features as well 7 original features. Therefore the total numbers of features were 28.The names of features extracted are listed in the Table1.

C. Experimental Framework

We have applied RFE and wrapper feature selection approaches on data set obtained from previous section. The experiments were made with and without sampling to deal with skewness of classes present in the data set as shown in Table 2. SMOTE is used as SMOTE sampling technique in the experiments. The performances of both approaches were measured using classifiers as J48, Logitboost, REP and Random Forest. RFE has performed better as compared to wrapper with all the classifiers except J48. Therefore, RFE is chosen as feature selection for our proposed system. Data set obtained after RFE is used for classification using HDDT. Before applying the HDDT classifier, multiclass data set is converted to binary class dataset which was achieved by converting all the "Observation" to "OK". We have repeated the experiment by converting "Observation" to "Fraud". Then, we have calculated the accuracy by taking an average of both experiments. RFE and HDDT was implemented using python 2.7 where as wrapper feature selection technique, SMOTE, J48, Rep Tree, Random Forest, Logitboost are done using Weka 6.4

TABLE I. DETAILS OF FEATURES EXPLORED FROM THE GIVEN ATTRIBUTES OF THE DATASET.

S. No.	Attribute	Features
1.	TimeAt	AvgClickPerMin AvgClickPer5Mins AvgClickPer10Mins AvgClickPer1Hr AvgClickPer2Hrs AvgClickPer6Hrs Var1min Var5mins Var10mins Var1hr Var2hrs Var6hrs
2.	IP Address	VarIP Ip/click
3.	Agent	VarAgent Agent/click
4.	CampaignID	VarCid Cid/Click
5.	Country	VarCntry Cntry/click
6.	Referrer	Refer/click

TABLE II. ACCURACY (IN %) OF DATASET AGAINST COMBINATION OF FEATURE SELECTION AND CLASSIFICATION TECHNIQUES

Classifier	**Wrapper**		**RFE**	
Sampling	**No**	**Yes**	**No**	**Yes**
J48	94.96	96.4	94	97
LogitBoost	85.3	97.1	87.7	97.13
REP Tree	93.8	97.1	94.29	97.23
Random Forest	96.7	96.3	96.06	96.1

V. RESULTS AND DISCUSSIONS

The result obtained from proposed system is mentioned in Table 3. Results show that accuracy achieved from RFE+ HDDT is 64.07%. We also have compared our framework with other classifiers namely Random forest, Logitboost, REP tree and J48.Table 3 shows that none of them have achieved accuracy higher than 50.37 on the data set under experimental study. Results are best as compared to existing classifier under experimental study.

TABLE III. CLASSIFICATION RESULTS OF PROPOSED METHOD WITH OTHER METHODS ON THE BLINDED TEST SET.

Method	**Accuracy (%)**
RFE+HDDT (Proposed)	**64.07**
RFE+Randomforest	49.99
RFE+LogitBoost	44.82
RFE+RepTree	46.82
RFE+J48	50.37

VI. CONCLUSION

Click fraud detection is a serious problem and has to be addressed intelligently. Hence ,this paper has developed a novel framework to detect fraudulent partners based on click data associated with mobile phone internet surfing .New features based on the attributes were generated, and these features were used to model the behaviour of each partner's click. To achieve a stable framework to predict click frauds, new feature evaluation technique namely RFE and HDDT as classification algorithm were applied on the data set under study .Results obtained from the experiments are quite promising in terms of accuracy as compared to existing models such as J48, Logitboost, REP tree and Random forest .

REFERENCES

[1] D. Scully, M. E. Otey, and M. Pohl, " Detecting Adversarial Advertisements in the wild" , KDD'11 Conference on knowledge discovery and data mining, 2011.

[2] D. Berrar, "Random forests for the detection of click fraud in mobile advertising", Journal of Machine Learning Research 1, 2000.

[3] C. Phua, E. Cheu, G. Yap, and K. Sim." Feature Engineering for Click Fraud Detection". FDMA 2012: International Workshop on Fraud Detection in Mobile Advertising, 2012.

[4] A. Metwally, D. Agrawal, A. E. Abbadi and Q. Zheng."Hide and Seek: Detecting Hit Inflation Fraud in Streams of Web Advertising Networks",.2006,www.cs.ucsb.edu/research/tech-reports/reports/2006-06.pdf.

[5] C. Phua, E.Cheu, G. Yap, K. Sim, M. Nguyen,"feature engineering for click fraud detection," FDMA 2012: International Workshop on Fraud Detection in Mobile Advertising, 2012.

[6] M. Häger, and T. Landergren, "Implementing best practices for fraud detection on an online advertising platform", .Department of computer science and engineering, Chalmers University of technology Goteborg, Sweden, 2010.

[7] V. Anupam, A. Mayer, K. Nissim, B. Pinkas, and M. Reiter. "On the Security of Pay-Per-Click and Other Web Advertising Schemes". in *Proc. 8th International Conference on World Wide Web*, pp. 1091–1100, 1999.

[8] S. Shivashankar, and P. Manoj, " Hierarchical committee Machines for fraud Detection in mobile Advertising", FDMA 2012: International Workshop on Fraud Detection in Mobile Advertising, 2012.

[9] K. S. Perera, B. Neupane, M. A. Faisal , Z. Aung, W.L. Woon. "A Novel Ensemble Learning-Based Approach for Click Fraud Detection in Mobile Advertising", International Workshop on Fraud Detection in Mobile Advertising, 2012.

[10] M.S. Nikulin,"Hellinger distance", in Hazewinkel Michiel, Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4, 2001

[11] A. Purwar and S. Singh, “Hybrid Prediction Model with missing value Imputation for medical data “ , Expert Systems with Applications , Elsevier , vol 42, no 13, pp. 5621-5631,2015.

[12] Hoens, T. Ryan and Qian, Qi and Chawla, Nitesh V. and Zhou, Zhi-Hua, “Building Decision Trees for the Multi-class Imbalance Problem”, in *Proc.* 16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining vol part 1, pp. 122- 134, 2012.

[13] A. Purwar and S. Singh, "Issues in Data mining: A comprehensive survey", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Dec 2014.