Network and Complex Systems
ISSN 2224-610X (Paper) ISSN 2225-0603 (Online)
Vol.11, 2020

The Effect of Feature Reduction in Click Fraud Detection: Review

Hazar Wanous* Nitul Dutta
School of Computer Engineering, Marwadi University, Rajkot, Gujarat, India

Abstract
It is almost impossible for online activities being without fraud. Online ads face a major threat represents by fake clicks which happen because of bots or some mischievous people. Several studies have solved the problem using machine learning algorithms. Some of them have solved only the problem of automatic click fraud (which carried out using bot), to classify physical or bot click. While many recent researches have detected click fraud problem in spite of clicks type. This paper presents a survey of methods used to detect fraud clicks on ads. It presents advantages, as well as disadvantages of each method, in general, Most recent studies in this field, have focused on features preprocessing before classification, because of the problems' type which imposed existence many related features and this may lead to overfitting. So the solution is applying dimensional reduction algorithms, to get better results and avoid overfitting.
Keywords: Click Fraud, dimensional reduction, features, Online advertising, pay_per_click.
DOI: 10.7176/NCS/11-01
Publication date: July 31st 2020

1. Introduction

After the big explosion in technology, and the rapid flow of web activities, online advertisements become more relevant in the advertisement market, even that nowadays there is no way to browse the internet without observing them (Ratliff 2010). Most advertisers or advertising companies prefer online ads because of its relatively inexpensive, reaches a wide audience, and more targeted audience (Deshwal 2016). Consequently, in recent years it has gradually been adopted. . In this industry, there are several pricing models, currently, pay per click is dominating the market (Zhang 2008). This model involves four parties: The user who views the ad, the advertiser who creates his/her ad and tries to be displayed on a publisher's website/application, the advertising networks, which represents coordinator between publisher and advertiser, and the publisher who invests his/her website or application to get revenue (Oentaryo 2014) (Daswani 2008).

To understand how does click fraud happen, we must know that it can be carried automatically or manually.

- Automatically: click fraud happens using bots, (bots are computer programs simulates a human activity) using to do illegal tasks over the network (Wang 2010). In our case, these bots are programmed to repeatedly click advertisements. However, it's easier to detect than the manual one.
- Manually: It is carried out by humans, like click farms (people are working as advertisements clickers to gain money) (Xu 2014).

In some cases, fraud clicks may be competition clicks, when unethical advertisers make numerous clicks on a competitor's ad to deplete his budget, or forced click as ads covering content that lead users to click on ads without giving an opportunity to them (Pechuán 2014). Anyway, As ad clicks increase, the publisher's profit increases that may encourage dishonest publishers to generate fake clicks on their websites/apps. In this case, advertiser spends more money without any interest in his ad, so this poses a major threat to the online advertising industry, it also will destroy the trust between publisher and advertiser (Taneja 2015). So we must find a solution to this growing problem. In this paper, we review the effect of feature engineering and dimensionality reduction to develop a new approach solving the problem of malicious clicks in advertisements. The remainder of our paper is organized as follows, we mention some statistics in Section II. After that we survey related work in Section III and Section IV to explain what is feature engineering and dimensionality reduction and why we need them in this problem. We present our proposed system in Section V. Finally, we conclude the paper with the impact of used features preprocessing besides classification algorithms.

2. Statistic of Pay Per Click Advertising and Click Fraud

Digital advertisements attract more fraudsters to do illegal activities and gain money through fraud clicks, which negatively effect on this market and make advertisers annoyed from this industry. Despite these troubles, PPC has gained popularity as a dynamic online ads model (Asdemir 2012).

- A Washington Post article reminded that around 40% of online ads in 2006 were PPC model. Estimates indicated that the attribution grew more than 50% in 2007 and it became approximately 60% in 2008. In general, the percentage of malicious clicks is hard to determine, most researchers estimate that 10-20% of ad clicks are fake (Kshetri 2010).
- Many researches and reports have indicated to the source of fake clicks, oftentimes "click farms" are based in developing countries, Times of India (May 3, 2004) and New York Times (May 12, 2009)

mention that there are groups of people responsible for click fraud operations, most of them from the following countries : South Africa, Albania, Brazil, Mexico, Angola, Bosnia/Herzegovina, Botswana, Mongolia, Egypt, Algeria, Sudan, Ukraine and other former Soviet Union economies, Nepal, Honduras, Vietnam, Indonesia, Philippines, Thailand, Syria, Lebanon, and many others (Kshetri 2010). However, malicious clicks problem is increasing worldwide,and according to statistics, it becames necessary to take preventative measures to protect online ads.

3. Literature Review

Click fraud in Pay_Per_Click model represents a big threat on online ads, as this problem increases, ads network have developed their own systems to detect fraud and reimburse users who endure from fraudulent clicks like Google AdWords. This problem has been solved using classification algorithms, but there is a gap to detect duplicate clicks in real world, it's still a constant hard conflict. Iqbal et la (2016) presented a technique for detecting automated clicks from the user side, using five classification algorithms SVM, 2KNN, 5KNN, C4.5, Naïve Bayes, Random Forest, they created HTTP request trees to generate features and used machine learning to detect ad requests. They achieved high accuracy, with tolerable false positive rate of all the captured network traffic. Taneja et la (2015) proposed a unique system to classify fraudulent publishers in mobile advertising using Hellinger Distance Decision tree algorithm and Recursive Feature elimination RFE (selection method to get the best features through removing the weak one after each step). RFE was used to generate valuable attributes before classification, and it gave better results compared to wrapper methods. They also compared five classification algorithms (Logitboost, REP tree, Random forest, J48 and HDDT) with applying both different types of selection methods (wrapper and filter) to get the most important features from the dataset. The result achieved from RFE+HDDT, was the best compared to other used algorithms.

Mouawi et la (2018) built a system to classify malicious publishers, Trusted by advertisers and advertising networks, through adding a new party to manage the process of click fraud detection, this system collects click data from advertising network and user activity data from advertiser then combine them to get trusted information. They used the following classification algorithm: KNN, ANN and SVM. Those methods gave convenient results and low values of the FPR. However, for used dataset KNN was the best classifierinput. Jiarui et la (2016) proposed a clustering framework to detect groups of malicious clicks, based on analyzing Crowdsourcing factors as the variance between normal and fraudulent traffic, the clicks which happened in a comparatively same time, and the denseness of clicks like click farms, where users represent a group related to a particular type of ad. Although this method is highly accurate, it requires a lot of time in addition to the high complexity. Zhang et la (2018). Used an effective type of Neural Network (Cost-sensitive Back Propagation) with wrapper methods to solve the problem of click fraud with considering the cost of misclassification for normal publisher and fraud one. This paper presented the importance of Artificial Bee Colony algorithm to avoid overfitting and reduce model complexity, it also gave better results compared with other wrapper algorithms. They achieved high accuracy by using the minimum number of features, with acceptable fault, and Strong robustnesso.

4. Feature preprocessing methods

In many classification problems, input data plays a critical role in model performance, regardless of the used algorithm we can notice that if we use raw datasets the outcomes will be non-logic. Here feature engineering comes into play, to preprocess existing factors and reduce their dimensionality (Marsland 2014). The dimensions of variables in the dataset, influence the machine learning model because of correlated features, redundant or categorized sometimes (Liu 2017). So that we are looking for the most favorable features, which conform to the assumptions of the model, Hence, transformations are essential and affect the training process (Zheng 2018). Dimensionality reduction includes two types: feature selection and feature extraction, some information can be lost during selection process as of some of the irrelevant attributes drop from dataset. Whereas, feature extraction is to generate a new set of features with captures the most important information of the original feature space (Khalid 2014).

Feature reduction methods require extensive calculations which lead to high complexity. However, they are used for some purpose, such as face recognition, speech recognition, biomedical engineering, marketing, wireless network, software fault detection, internet traffic prediction, etc (Ghojogh 2019). also for any application has noisy, uncompleted, relevant or redundant raw data.

5. Approach

Our object is to solve click fraud detection on advertisement using derived attributes which have a critical function to improve system effectiveness and avoid overfitting. Our proposed system includes three steps illustrated in Figure 2 the most important one is detecting related features to extract new one and drop useless one using dimensionality reduction algorithm. We are going to study clicks repletion opportunity which happens

from the same device, same operating system, and application at the same click time.The main idea is about finding correlation coefficient for attributes, so we will generate new features which in fact represent the relationship between the old one. The second step is applying a suitable classification algorithm on processed attributes and, the last one is detecting click fraud.

6. Conclusion

Data in real world are complex, changeable and do not hold enough information for classification. In a problem like click fraud we have so many correlated variables, also high-entropy features together with low-entropy ones, which need to preprocess to get satisfied outcomes.After surveying previous studies, we found some weaknesses like high false positive rates or less accuracy. However researches using reduction algorithms gave better results than others so it is necessary to focus on data preprocessing which represents the key to improve learning model results', and make it easier to understand.

References

Asdemir, K., Kumar, N., & Jacob, V. S. (2012). Pricing models for online advertising: CPM vs. CPC. Information Systems Research, 23(3 PART 1), 804–822. https://doi.org/10.1287/isre.1110.0391

Daswani, N., Mysen, C., Rao, V., Weis, S., Gharachorloo, K., & Ghosemajumder, S. (2008). Online Advertising Fraud. C, 1–28.

Deshwal, P. (2016). Online advertising and its impact on consumer behavior. International Journal of Applied Research, 2(2), 200–204. www.allresearchjournal.com

Ghojogh, B., Samad, M. N., Mashhadi, S. A., Kapoor, T., Ali, W., Karray, F., & Crowley, M. (2019). Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review. May. http://arxiv.org/abs/1905.02845

Iqbal, M. S., Zulkernine, M., Jaafar, F., & Gu, Y. (2016). FCFraud: Fighting Click-Fraud from the User Side. Proceedings of IEEE International Symposium on High Assurance Systems Engineering, 2016-March, 157–164. https://doi.org/10.1109/HASE.2016.17

Jiarui, X., & Chen, L. (2016). Detecting crowdsourcing click fraud in search advertising based on clustering analysis. Proceedings - 2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing, 2015 IEEE 12th International Conference on Advanced and Trusted Computing, 2015 IEEE 15th International Conference on Scalable Computing and Communications, 20, 894–900. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.172

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. Proceedings of 2014 Science and Information Conference, SAI 2014, July, 372–378. https://doi.org/10.1109/SAI.2014.6918213

Kshetri, N. (2010). The economics of click fraud. IEEE Security and Privacy, 8(3), 45–53. https://doi.org/10.1109/MSP.2010.88

Liu, C., Wang, W., Konan, M., Wang, S., Huang, L., Tang, Y., & Zhang, X. (2017). A new validity index of feature subset for evaluating the dimensionality reduction algorithms. Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2017.01.017

Marsland, S. (2014). Machine learning: An algorithmic perspective. In Machine Learning: An Algorithmic Perspective, Second Edition. https://doi.org/10.1201/b17476

Mouawi, R., Awad, M., Chehab, A., El Hajj, I. H., & Kayssi, A. (2019). Towards a Machine Learning Approach for Detecting Click Fraud in Mobile Advertizing. Proceedings of the 2018 13th International Conference on Innovations in Information Technology, IIT 2018, 88–92. https://doi.org/10.1109/INNOVATIONS.2018.8605973

Oentaryo, R., Lim, E. P., Finegold, M., Lo, D., Zhu, F., Phua, C., Cheu, E. Y., Yap, G. E., Sim, K., Nguyen, M. N., Perera, K., Neupane, B., Faisal, M., Aung, Z., Woon, W. L., Chen, W., Patel, D., & Berrar, D. (2014). Detecting click fraud in online advertising: A data mining approach. Journal of Machine Learning Research.

Pechuán, L. M., Ballester, E. M., Manuel, J., & Carrasco, G. (2014). Online Advertising and the CPA Model : Challenges and Opportunities. 3, 324–334.

Ratliff, J. D., & Rubinfeld, D. L. (2010). Online advertising: Defining relevant markets. Journal of Competition Law and Economics, 6(3), 653–686. https://doi.org/10.1093/joclec/nhq011

Taneja, M., Garg, K., Purwar, A., & Sharma, S. (2015). Prediction of click frauds in mobile advertising. 2015 8th International Conference on Contemporary Computing, IC3 2015, 162–166. https://doi.org/10.1109/IC3.2015.7346672

Wang, P., Aslam, B., & Zou, C. C. (2010). Peer-to-Peer Botnets : The Next Generation of Botnet Attacks. Electrical Engineering.

Xu, H., Liu, D., Koehl, A., Wang, H., & Stavrou, A. (2014). Click Fraud Detection on the Advertiser Side. 419–438.

Zhang, L., & Guan, Y. (2008). Detecting click fraud in pay-per-click streams of online advertising networks. Proceedings - The 28th International Conference on Distributed Computing Systems, ICDCS 2008, 77–84. https://doi.org/10.1109/ICDCS.2008.98

Zhang, X., Liu, X., & Guo, H. (2018). A click fraud detection scheme based on cost sensitive BPNN and ABC in mobile advertising. 2018 IEEE 4th International Conference on Computer and Communications, ICCC 2018, 1360–1365. https://doi.org/10.1109/CompComm.2018.8780941

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning. In O'Reilly Media. https://doi.org/10.13140/RG.2.1.3564.3367

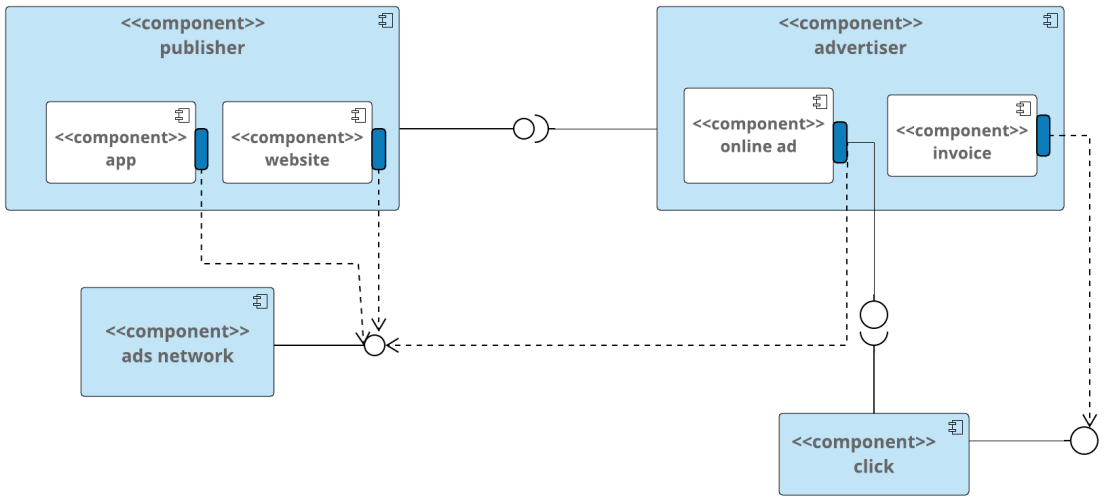

Figure 1.Component diagram for pay per click model

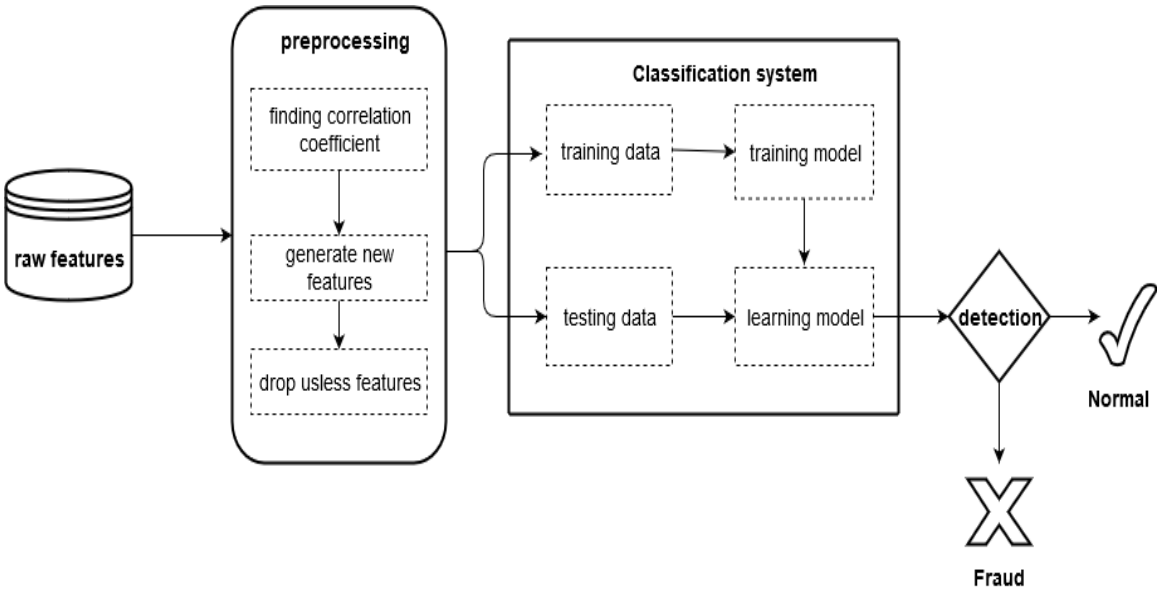

Figure 2. Architecture of proposed system