Abstain Classifiers for Auditing Online Advertisement Privacy Settings

Hsin Miao, Jagannathan Chakravarthy

{*hsinm,jchakrav*}@*andrew.cmu.edu*

Abstract—**We propose an auditing framework that can determine whether online advertisement privacy policies are violated. The framework utilizes abstaining classifiers, whether they can not only output positive and negative results but also "I don't know" predictions. Such classifiers are useful to reduce the potential cost of privacy policy violation because they can reduce incorrect predictions by reporting 'I don't know" outputs to human for auditing. We implemented two abstaining classifier algorithms: Ensemble Enumeration and Ensemble Relaxed Enumeration. We found that compared to traditional classifier selection algorithms, the performance of the predictions are higher by abstaining classifiers. This project is the first-ever study that combines abstaining classifiers and auditing applications. Further explorations including more data and more complex models can be utilized to get more insight into this field.**

I. Introduction

THE effectiveness of ad campaigns have inflated due to targeted marketing, which has been made possible through websites which track user data and thereby predict a user's interests. This raises a lot of concerns for the user, who would like to know which set of personal information they provide is used for targeted advertising. For this reason, the websites that acquire or collect the personal information from users tell them which set of this information will be used for targeting and which set is not used through their privacy policy. However, we cannot directly determine whether these policies are complied. They can be ascertained by conducting indirect experiments such as analyzing the ads that the user encounters on the internet.

For the companies that provide targeted advertisements, it is impossible to determine whether they violate their privacy policies by human. Machine learning algorithms help the companies automatically audit their privacy settings.

Our work is motivated by Datta *et al.* [1] in which experiments were conducted on Google ads. They developed a tool "AdFisher" that can automatically collect advertisements on the Internet. They found some correlations between the user provided data, the websites the user visited and the advertisements they see. The result implies possible violations of Google's privacy policy with respect to use of sensitive information for targeted advertising.

In this project, we proposed a automated auditing framework using abstain binary classifiers, which belong to online machine learning algorithms. Online machine learning algorithms learn one datum at a time, and predict a label from the datum. The prediction hypothesis is refined if the true label of the datum is learned. Online machine learning models are suitable for the application of auditing online advertisement privacy settings because the cost of frequently training the classifiers for auditing is quite huge. By online learning algorithms, the classifiers can be refined (or learned) in the auditing process.

Abstaining binary classifiers belong to online machine learning models, where they are allowed to output "I don't know" and refrain from predicting the class label of the given data. They may abstain from making a prediction if they are not confident enough for the prediction. Abstaining classifiers are essential in applications of medical diagnosis and fraud risk assessment because incorrect predictions may contribute to fatal consequences[2]. By utilizing the concept of abstaining classifiers, incorrect predictions are reduced because the classifiers can directly output "I don't know" if they do not have enough confidence. After obtaining the "I don't know" results, expertise can make accurate judgements.

Abstaining classifiers are also suitable for the application on auditing because the incorrect predictions should be reduced. If the violations of privacy policies are detected internally, it is easy to fix the algorithms that provide targeted advertisements;

however, such violations found by the users might contribute to huge costs.

We utilized abstaining classifiers to perform the analysis on the data obtained using AdFisher in previous experiments conducted by Datta *et al.* [1]. Abstaining classifiers are integrated into AdFisher to analyze whether the advertisement settings violate their privacy policies. The paper is organized as follows. Section II reviews some literatures about the experiments on online advertisement settings, and abstaining classifiers. Section III described the algorithms we implemented. We will report the experimental results in Section IV, and state conclusions and future works in Section V.

II. Related Work

Analysis on Google Ads has been extensively studied by many researchers. Our work was inspired by [1] in which a tool named AdFisher is used to collect Google ads from different websites by exposing the user to different treatments and analyzing how the treatments correspond to the Ads shown for the users. They conclude that Google Ad-settings is opaque about some features and transparent about other features and these variances lead to seemingly discriminatory ads. Their main results were that when a user visits websites related to substance abuse, he gets significantly different ads as compared to a user who does not visit these websites although the Google Ad-settings page does not show any changes reflecting to the user's interest. Also, when male and female users are exposed to the same treatment, they found that male users get significantly more ads related to high-paying jobs than their female counterparts.

Some literatures regarding abstaining classifiers are introduced as follows. Li *et al.* proposed a learning framework "Know What it Knows" [3] for online learning The learning framework is designed to make only accurate predictions although the algorithm can output "I don't know" for uncertain results. Theoretical bound on the number of abstaining is discussed by the authors because such algorithm should satisfy the requirements of accurate predictions and minimum abstaining outputs.

In the survey paper written by Balsubramani [2], the utility of abstaining in binary classifiers is explored in which the paper concentrates on the tradeoff between avoiding incorrect predictions and making sufficiently predictions. The paper [4] also studies about the tradeoff between ”don't know” predictions and making mistakes. It presents an algorithm that minimizes the number of ”don't know” predictions bounded by the number of allowed mistakes. In addition, Blaszczyski *et al.* proposed a framework of ensembles of abstain classifiers by rule sets [5]. Moreover, the paper [6] talks about a method to build a specific type of abstaining binary classifier using ROC analysis. The classifiers are built on cost based model, bounded-abstention model and bounded-improvement model. The paper demonstrates the use of these models on reducing misclassification costs.

Although algorithms and theoretical bounds of abstain classifiers are discussed in those literatures, there are few applications based on abstaining classifiers. To best of our knowledge, this project is the first one that integrate abstain classifiers into auditing applications.

III. Method

A. Automated Auditing Framework for Online Advertisement Settings

In this project, we extended the tool "AdFisher" that abstaining classifiers are integrated. This section first introduces the framework of AdFisher. AdFisher is a tool that can automatically collect online advertisements written by Selenium API [7] in Python. Users can first set the treatments for some browser agents. For example, if we want to determine whether Google Ad-setting is related to discrimination on gender, we can first set different genders in Google Ad-setting page. After setting treatments, those browser agents then collect online advertisements on websites.

Analyses are then performed to determine whether privacy policies are violated. The collected advertisements for a browser agent are encoded into one agent vector. The label of an agent vector is the type of treatment. (e.g. control group or experimental group.) Machine learning algorithms are utilized to train a model by training data of the vectors, where the learning models can be evaluated by the accuracy of test data. The next step is to determine whether the prediction results form two different groups. AdFisher performs permutation tests on the test data, and then report p-value that specifies whether the different treatments contribute to different groups of results. After the analyses processes, advertisement companies can use the results to audit their privacy policies.

B. *Abstaining Classifiers*

The main contribution of the project is to implement abstaining classifiers in AdFisher framework. We implemented the classifiers based on "Know What it Knows" model and the concept of ensemble classifiers. There are two algorithms in this project: Ensemble Enumeration and Ensemble Relaxed Enumeration, which will be described in the following paragraphs.

1) Ensemble Enumeration: The Ensemble Enumeration works when the hypothesis set $\mathcal{H}$ is finite. We first divided the data into training data and testing data, where testing data contains $\delta\%$ of the entire data set. The training data is used to generate $|\mathcal{H}|$ classifiers. Logistic Regression model is chosen in our experiments because Datta *et al.* [1] found that such model is suitable in online advertisement applications. (Note that our abstain classifier framework is flexible that other machine learning models can be easily replaced.) In order to obtain the hypothesis set, $|\mathcal{H}|$ regularization parameters in logistic regression are set to generate classifiers.

The Ensemble Enumeration algorithm is described after the hypothesis set $\mathcal{H}$ is generated.

Algorithm 1 Ensemble Enumeration

```
$\mathcal{V} \leftarrow \mathcal{H}$
**for** $i = 1, 2, ...$ **do**
    **if** $h(x_i)$ is the same $\forall h \in \mathcal{V}$ **then**
        Choose any $h \in \mathcal{V}$ and predict $h(x_i)$
    **else**
        Output "I don't know"
        Predict with the majority, i.e. $\hat{y}_i = argmax_{\lambda \in \{-1,+1\}} \{h \in \mathcal{V} : h(x_i) = \lambda)\}$
        $\mathcal{V} \leftarrow h \in \mathcal{V} : h(x_i) = \hat{y}_i$
    **end if**
    **if** $|\mathcal{V}| = 1$ **then**
        Terminate with final output $\mathcal{V}$
    **end if**
**end for**
```

The algorithm works depending on the predictions by the classifiers in the hypothesis set $\mathcal{H}$. If the predictions are all identical, the abstaining classifier is confident that the result is correct, so it outputs the prediction directly. However, if there is no consensus among the predictions, the abstaining classifier will output "I don't know", and update the hypothesis set $\mathcal{H}$ by ensemble that keeps the hypotheses with majority predictions.

Some properties of the algorithm is described as follows. The abstention bound in the algorithm is $|\mathcal{H}| - 1$ because each time the algorithm outputs an "I don't know", at least one hypothesis in $\mathcal{V}$ is removed. The ensemble step in the algorithm is to aggregate the prediction results by the hypotheses, and remove some unwanted hypotheses. Although the Ensemble Enumeration algorithm is quite general and simple, the time and space complexity of the algorithm is huge because the size of the hypothesis set $|\mathcal{H}|$ might be very large.

2) Ensemble Relaxed Enumeration: The assumption of "Know What it Knows" model is that the learner makes no mistakes, which is a extreme restriction in many circumstances. To mitigate the restriction, the enumeration algorithm was extended to allow up to k mistakes in Ensemble Relaxed Enumeration algorithm.

Algorithm 2 Ensemble Relaxed Enumeration

```
$\mathcal{V} \leftarrow \mathcal{H}$, $s \leftarrow |\mathcal{H}|^{k/k+1}$, $m \leftarrow 0$
**for** $i = 1, 2, ...$ **do**
    $\mu = min_{\lambda \in \{-1,+1\}} |\{h \in \mathcal{V} : h(x_i) = \lambda)\}|$, so that $\mu$ is the number of hypotheses in $\mathcal{V}$ predicting the minority label
    **if** $\mu \leq s$ **then**
        Predict with the majority, i.e. $\hat{y}_i = argmax_{\lambda \in \{-1,+1\}} \{h \in \mathcal{V} : h(x_i) = \lambda)\}$
    **else**
        Output "I don't know"
        $\mathcal{V} \leftarrow h \in \mathcal{V} : h(x_i) = \hat{y}_i$
    **end if**
    **if** $|\mathcal{V}| = 1$ **then**
        Terminate with final output $\mathcal{V}$
    **end if**
**end for**
```

The Ensemble Relaxed Enumeration algorithm tolerates some incorrect predictions. Theoretical bound of the algorithm can be found in the paper [4].

C. *Traditional Classifiers*

This section described how traditional classifiers work. We compared the abstaining classifiers to traditional model selection frameworks in the project. Traditional model selection frameworks utilize cross validation to select the best model for testing process. Each model is trained and evaluated the prediction results by cross validation, and the hypothesis with the highest accuracy is selected to be the final classifier. In our implementation, we used model selection framework to select one of twenty logistic regression classifiers.

IV. Experimental Results

In the section, we will first introduce the data we used, and then show the experimental results of abstaining classifiers.

A. Data Collection and Evaluation Metric

We analyzed the data collected by AdFisher. Several treatments on browser agents were set to collect online advertisements, including *mental disorder*, *gender effects on job search*, *cars*, and *dating websites*. A treatment is that the browser agents visit some specific websites and then collect online advertisements on websites such as bbc.com. For example, the *cars* treatment is that the browser agent first visits some car websites, and then collects online ads.

We evaluate the accuracy of the results predicted by classifiers. There are two possibilities of high accuracy. The first one is that the treatment is effective, so browser agents for different treatments in "advertisement space" can be separated clearly by simple models. The second one is that the model is suitable for such kind of data, so although browser agents for different treatments in "advertisement space" are not separated clearly, we can still use the model to separate the data.

We also evaluate p-value of permutation tests. The p-value can be used to determine whether the effect of the treatments is significant. If the p-value is smaller than 0.05, we can reject the null hypothesis that the results of two treatments are identical. In other words, we can conclude that we have some evidence showing that different treatments contribute to different results by the p-value which is smaller than 0.05.

B. Comparisons between Abstaining Classifiers and Traditional Classifiers

TABLE I
Prediction accuracy and p-value for different algorithms in mental disorder treatment

Algorithm	Accuracy	p-value
Ensemble Enumeration	0.505	0.4665
Ensemble Relaxed Enumeration	**0.5475**	**0.0235**
Traditional Classifiers	0.505	0.4479

TABLE II
Prediction accuracy and p-value for different algorithms in dating treatment

Algorithm	Accuracy	p-value
Ensemble Enumeration	**0.59**	**0.0068**
Ensemble Relaxed Enumeration	0.56	0.055
Traditional Classifiers	0.555	0.0704

Table I and II show the prediction accuracy and p-value for different algorithms in the *mental disorder* and *dating* treatments. We found that the prediction accuracy in traditional classifiers is low, which is near random prediction. The p-value is high in traditional classifiers. Although by the abstaining classifier we merely obtained slightly higher prediction accuracy, we achieved the p-value smaller than 0.05, so we have some evidence that the *mental disorder* and *dating* treatments contribute to different set of online advertisements.

Table III shows the results in the *cars* treatment. The accuracy of the three algorithms are relatively higher compared to the *mental disorder* treatment, and p-values are very small. The accuracy in the abstain classifier is higher than traditional classifiers, which shows that abstain classifier algorithms can effectively enhance the performance.

Table IV shows the case that the results of prediction accuracy are low, and p-values are high. In this case, we have no evidence that the results of different gender treatments on job search are different. There are two possibilities to interpret the result. The first one is that the collected online advertisements are not related to gender discrimination. The other one is that it is difficult to separate browser agents in different treatments in the "advertisement space".

C. Discussion

The results show that the performance of abstaining classifiers is better than traditional classifiers on the collected data. However, there are some limitations in the project.

The first limitation is the size of data. In fact, there are only hundreds of browser agents because

TABLE III
PREDICTION ACCURACY AND P-VALUE FOR DIFFERENT ALGORITHMS IN CARS TREATMENT

Algorithm	Accuracy	p-value
Ensemble Enumeration	**0.8818**	<0.0005
Ensemble Relaxed Enumeration	0.8454	<0.0005
Traditional Classifiers	0.8545	<0.0005

TABLE IV
PREDICTION ACCURACY AND P-VALUE FOR DIFFERENT ALGORITHMS IN GENDER EFFECTS ON JOB SEARCH TREATMENT

Algorithm	Accuracy	p-value
Ensemble Enumeration	0.515	0.3778
Ensemble Relaxed Enumeration	0.55	0.0973
Traditional Classifiers	0.55	0.0929

it takes a lot of time to collect online advertisements. Because if the small size of data, we cannot conclude that the results can be generalized to all cases. If the size of data can be scaled up, we might obtain significant results and then we will gain much more insights on the usage and the applications of abstaining classifiers.

Second, the design of hypothesis set can be refined. The "Know What it Knows" model is designed to predict bounded incorrect results. However, due to the time limitation, we could only generate the hypothesis set by using different parameters in logistic regression model. We believe that if more complex design of hypothesis set and learning models are utilized, we will obtain much better performances on both prediction accuracy and p-value results.

V. CONCLUSION AND FUTURE WORKS

In the project, we proposed an auditing framework for online advertisement settings by using abstaining classifiers. We extended the tool "AdFisher" by implementing two abstaining classifier algorithms: Ensemble Enumeration and Ensemble Relaxed Enumeration. The results show that the performance on abstaining classifiers are better than traditional model selection algorithms because abstaining classifiers output "I don't know" on uncertain results.

There are some possible future works on the project. First, in order to obtain more significant results, more data on online advertisements should be collected. More browser agents can be used to attain the goal. The other way to refine the work is to design more complex machine learning models. For example, the design of hypothesis set in abstaining classifiers is essential for significant results. Also, the proposed framework is flexible that it can combine different machine learning models in the learning process. By ensemble for different learning models, we believe that we will obtain higher accuracy on the prediction of browser agents, and then we will gain more insights on the application of auditing on privacy policies.

REFERENCES

[1] A. Datta, M. C. Tschantz, and A. Datta, "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination," *arXiv:1408.6491*.

[2] Akshay Balsubramani, "The Utility of Abstaining in Binary Classification," *manuscript*, 2013.

[3] Lihong Li and Michael L. Littman and Thomas J. Walsh, "Knows What It Knows: A Framework For Self-Aware Learning," *Proceedings of the 25th international conference on Machine learning*, 2008.

[4] A. Sayedi, M. Zadimoghaddam, and A. Blum, "Trading off mistakes and don't-know predictions," in *Advances in Neural Information Processing Systems*, pp. 2092–2100, 2010.

[5] J. Blaszczynski, J. Stefanowski, and M. Zajac, "Ensembles of abstaining classifiers based on rule sets," in *Foundations of Intelligent Systems*, pp. 382–391, Springer, 2009.

[6] T. Pietraszek, "Optimizing abstaining classifiers using roc analysis," in *Proceedings of the 22nd international conference on Machine learning*, pp. 665–672, ACM, 2005.

[7] SeleniumHQ, "SeleniumHQ. [online]. available: http://www.seleniumhq.org/."