PROJECT PROPOSAL

VIVIDH TALESARA | HARSHAL GIRME | AKASH BANGERA

10/05/2017

Human or a Robot: Predict If an Online Bid is made by Machine or Human

Clear Description of the Topic

In this Project research, we'll be chasing down robots for an online auction site. It is found that Human bidders on the site are frustrated with their inability to win auctions vs. their software-controlled counterparts. As a result, usage from the site's core customer base is plummeting. In order to rebuild customer happiness, we need to eliminate computer generated bidding from their auctions.

The goal of our Project is to identify online auction bids that are placed by "robots", helping us to easily flag these users for removal from the site to prevent unfair auction activity.

Background research of related work

We found that their attempt at building a model to identify these bids using behavioral data, including bid frequency over short periods of time, has proven insufficient. So we will be using multiple classification approaches to get the best optimized solution for identifying the “bots” and “humans”.
We will be using two datasets:

1. Bidder dataset that includes a list of bidder information, including their id, payment account, and address.
2. Bid dataset that includes 7.6 million bids on different auctions. The bids in this dataset are all made by mobile devices.
3. Training and Testing sets for sample calculations.

The online auction platform has a fixed increment of dollar amount for each bid, so it doesn't include an amount for each bid. We will be learning the bidding behavior from the time of the bids, the auction, or the device that will lead us to much more accurate results.

Data Sources

We have used Kaggle Competition Datasets: https://www.kaggle.com/c/facebook-recruiting-iv-human-or-bot

1. sampleSubmission.csv
2. bids.csv
3. test.csv
4. train.csv

Algorithms & Code Sources Used

1. Random Forest Classifier

- A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

2. Feature generation

- Generate a new feature matrix consisting of all polynomial combinations of the features with degree less than or equal to the specified degree. For example, if an input sample is two dimensional and of the form [a, b], the degree-2 polynomial features are [1, a, b, a^2, ab, b^2].

3. Cross Validations

- A another part of the original dataset can be held out as a so-called “validation set”: training proceeds on the training set, after which evaluation is done on the validation set, and when the experiment seems to be successful, final evaluation can be done on the test set. However, by partitioning the available data into three sets, we drastically reduce the number of samples which can be used for learning the model, and the results can depend on a particular random choice for the pair of (train, validation) sets. A solution to this problem is a procedure called cross-validation (CV for short).

References

- https://www.kaggle.com/c/facebook-recruiting-iv-human-or-bot/discussion/14628
- https://github.com/mikegloudemans/kaggleBotOrNot
- https://github.com/rushter/Facebook-Recruiting
- http://blog.kaggle.com/2015/06/19/facebook-iv-winners-interview-2nd-place-kiri-nicholaka-small-yellow-duck/
- https://github.com/rushter/Facebook-Recruiting