CROSS DEVICE TRACKING: MATCHING DEVICES AND COOKIES

Roberto Díaz-Morales - roberto.diaz@treelogic.com

TREELOGIC
Department of Research and Development

November, 2015

Table of Contents

1. Cross-Device Tracking
   - Multiple Devices
   - The Problem
   - Cross-Device Tracking
   - Addressing the problem
2. The Challenge
   - The ICDM 2015 Challenge
   - The Dataset
   - The Goal
3. Algorithm
   - Preprocessing
   - Initial Selection of Candidates
   - Dataset
   - Supervised Learning
   - Bootstrap Aggregating
   - Semi-Supervised Learning
   - Post-Processing
4. Results
5. Conclusions

Multiple Devices

- Some years ago the PC was the only device with internet access.
- In the late 90s, some people started to use dedicated PDA devices.
- In 2007 Apple introduced the iPhone and in 2008 Android was released. Smartphones became widespread very quickly.
- The term smartTV began to be formally used in 2010 for a television set with internet and Web 2.0 features.
- Currently people use many different devices to access internet.

The Problem

- The internet consumptions habits have changed.
  - People can access internet anytime.
  - People can access internet anywhere.
  - Even in parallel using multiple devices at the same time.
- Behaviors have become increasingly complex.
  - People can watch a film on a SmarTV while is giving his opinion about it in social networks using a smartphone.
  - People can look an online shop with the phone and buy some products after arriving home from the laptop.

- The data used to understand their behavior are fragmented.
- The identification of users is a challenge.

Cross-Device Tracking

- Cross-Device targeting or tracking is to know if the person using computer X is the same one that uses mobile phone Y and tablet Z.

- It is an important emerging technology:
  - Internet advertising move a lot of money.
  - A correct identification can provide ads more efficiently.
  - Very valuable for marketers.

Addressing the problem

Attending to the way of addressing the problem, we can find different groups of solutions:

- Some companies offer a service to track users that are signed in their websites and apps.
  - Very simple and efficient.
  - This requirement is not met in several cases.
- Using deterministic information and exact match rules:
  - Using the information personal information in forms (It uses credit card numbers, email ...)
  - Limited to a reduced number of situations and platforms.
- Machine Learning to create predictive models:
  - Open to any informative feature.
  - Applicable to any situation.
  - It is an uncharted territory.

The ICDM 2015 Challenge

The ICDM 2015 Cross-Device Connections challenge was organized by Drawbridge.

drawbridge

The Challenge was hosted by Kaggle

kaggle

It took place from June 1st 2015 to August 25th 2015 and it brought together 340 teams.

Given usage data and IDs, the goal was to determine which cookies belong to an individual user using a device.

The Dataset

- The dataset contains relational information about devices, cookies, IP addresses and behavior.
- Tables with devices and cookies provided high level information regarding the device and cookie.
- Another table describes the joint behavior of device or cookie on IP addresses.
- We have information that describe each IP across all the devices or cookies.
- Two tables describe properties and categories of the website / mobile app.

Evaluation metric

The objective was to get the classifier with the highest $\mathcal{F}_{0,5}$ score.

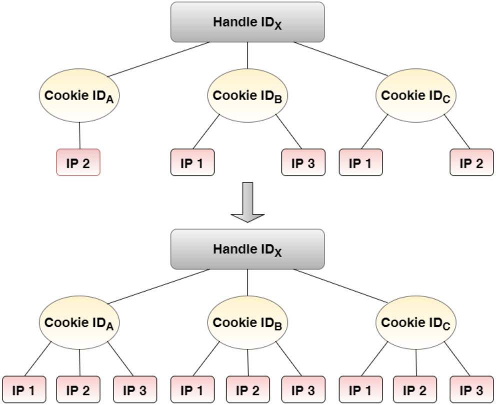
$$\begin{aligned} \mathcal{F}_\beta &= (1+\beta^2)\frac{pr}{\beta^2 p + r}, \\ p &= \frac{tp}{tp+fp}, \\ r &= \frac{tp}{tp+fn} \end{aligned} \tag{1}$$

By using $\beta = 0{,}5$ the score weighs precision higher than recall. The score is formed by averaging the individual $\mathcal{F}_{0,5}$ scores for each device in the test set.

treelogic

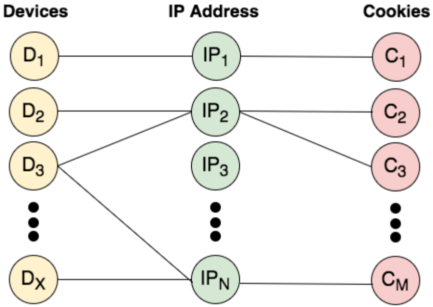
Preprocessing

Sharing information about IPs in cookies with the same handle.



Initial Selection of Candidates

- It is not possible to create a training set containing every combination of devices and cookies. Some basic rules were created to select a reduced number of eligible cookies for every device.



- The rules are based on the IP addresses that both device and cookie have in common and how frequent they are in other devices and cookies.

Initial Selection of Candidates

- At the beginning, if an IP address has less than X links with devices and Y links with cookies then all the cookies are candidates of the devices.

- For every device, we iteratively increase the values of X and Y until it has eligible cookies.

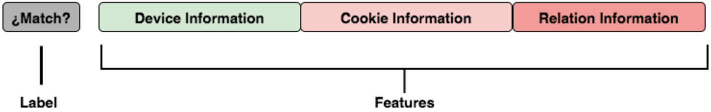
- If we cannot find candidates with these rules, every cookie that shares an IP address with the device is a candidate.

- In 98.3 % of the devices, the set of candidates contains the deviceś cookies.

treelogic

Creating a dataset

- In the training set, every sample represents a pair device/eligible cookie pair and is composed by a total of 67 features.



- There are three main groups of features:
  - Features with information about the device (Operating System, Country,...)
  - Features with information about the cookie (Cookie Browser Version, Cookie Computer OS, ...)
  - Features with relational information (Number of IP addresses in common, the average features of these IP addresses,...)

treelogic

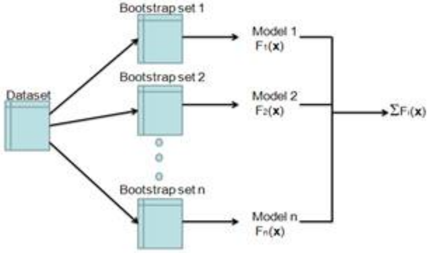
Supervised Learning

- Algorithm: Regularized Boosted Trees.
- Cost Function: Logistic.
- Software: XGBoost (Parallel on a multi-threaded CPU)
- Hyperparameter selection: 10 fold cross validation.

treelogic

Bootstrap Aggregating

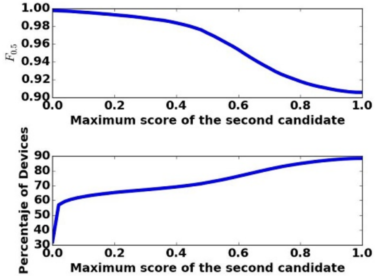
- Bootstrap aggregating, also calling bagging, is a technique to improve the stability of machine learning algorithms.



- It consists on the generation of different training sets sampling from the original one and then training the algorithm with all of them and average results.
- We have used eight baggers.

Semi-Supervised Learning

- Semi-supervised learning makes use of unlabeled data.
- In cases where the first candidate obtains a high score and the second one obtains a low score it is very likely that the first one is a match.
- We have taken the devices of the test where the first candidate scores higher than 0.4 and the second candidate scores less than 0.05 and we have considered pairs device-cookie and we have included them in the training set to recalculate the model.



Post-Processing

- We take the candidate cookie with the highest score.
- If this score is higher than a threshold then we choose this cookie and other ones with the same handle as device's cookies.
- If the score is very low, we choose a new set of eligible cookies selecting every cookie that shares an IP with the device.
- We use different thresholds to continue including the second and third cookies with higher score.

Results

- The score of this competition was evaluated using a test set of 61156 devices.
- During the challenge the leaderboard was calculated on 30 % of the test set (public leaderboard).
- After the competition the final result was evaluated on the other 70 % (private leaderboard).

Results

Sel = Initial selection of candidates
SL = Supervised Learning
B = Bagging
SSL = Semi-Supervised Learning
PP = Post Processing

Procedures	Public Leaderboard	Private Leaderboard
Sel	0.498	0.5
IL + SL	0.872	0.875
Sel + SL + B	0.874	0.876
Sel + SSL +B + PP	0.878	0.88

Conclusions

- This is a simple way to deal with the problem of cross-Device tracking.
- It has obtained an $F_{0,5}$ score of 0.88.
- The good performance has been tested finishing 3rd of 340 teams.
- Future Research Lines:
  - Exploring new features like time series of visited addresses.
  - Using an ensemble of different algorithms could obtain better results if there is diversity among the classifiers.

Acknowledgements

My gratitude:

- To the organizers.
- To everyone who participated in the challenge.

Thank you!!!

treelogic