Rapport hebdomadaire de stage

Alexis Blanchet

25 avril 2018

1 INTRODUCTION

Arrivé lundi en stage chez SFR, cette semaine a été dédiée a l'exploration des données fournies, du matériel et du lieu. On présentera ici le stage afin d'introduire de manière claire les problématiques qui seront traitées ultérieurement.

le stage consiste a détecter automatiquement le passage de publicité a la télévision afin d'extraire des informations concernant les habitudes et les sensibilités des utilisateurs aux publicités.

On dispose pour cela des actions de chaque utilisateur : à chaque fois que le décodeur de la télévision est sollicité, nous recevons une fichier qui indique entre autre la date exacte de l'action et la nouvelle chaîne vers laquelle l'utilisateur a zappé. On peut donc a partir de ces données extraire des tendances qui permettront la détections de publicités (en théorie).

1.1 SOCIOLOGIE DU ZAPPING

Le premier travail consiste a chercher les raisons qui poussent les utilisateurs a changer de chaîne. La croyance populaire voudrait que les utilisateurs changent de chaîne a l'apparition d'une publicité. Divers articles, par exemple [1] ou [2], datant des années 1990/1992, démontrent que l'apparition de la télécommande doublée de l'augmentation du nombre de chaînes entraîne une zapping plus important de la part des utilisateurs. Il s'agit en effet du pouvoir d'action de ce dernier, de la « lutte du Zappeur avec la publicité ».Cependant des études plus récentes ([3] du 05/11/2010) montrent que les utilisateurs regardent majoritairement les publicités et que les publicités touchent 85% des adultes quotidiennement (des 376 adultes qui ont participé a l'étude). En moyenne, la population est touchée par 26 publicités chaque jour. Fait notable pour la suite des travaux, le zapping juste avant la pub est de 11%, juste après la pub de 13%, et pendant la pub de seulement 14%.

1.2 Définition du problème et présentation des données

Prenons le problème simplement : chaque utilisateur a une décision binaire a prendre a chaque minute. Soit il change de chaîne soit il reste sur celle-ci. On cherchera plus tard étudier si la chaîne d'arrivée a une importance dans le zapping(on peut par exemple imaginer que le début d'un programme sur une autre chaîne justifie un changement). Une fois que l'utilisateur a pris la décision de changer de chaîne il faut classer la raison de ce changement de façon a détecter les publicités. Pour cela on dispose de différents jeux de données : tout d'abord les programmes TV fixés a l'avance, et ensuite les données des utilisateurs par box. Il s'agit d'un historique de toutes ls actions effectuées a partir du décodeur télé.

1.3 Principales difficultés

La difficulté majeur consiste en l'absence totale d'historique : on ne sait jamais si une personne a changé de chaine a cause d'une publicité. On ne sait en fait meme pas quand sont effectivement les publicités. Nous n'avons ainsi AUCUN labels permettant un apprentissage. On va donc avant toute chose devoir créer un tel historique en apprenant au système a détecter de potentielles publicités et en les annotant ensuite de façon a affiner la précision de nos algorithmes. Il s'agira donc de construire un système qui se doit d'être précis sur un ensemble de données bruitées et incomplètes.

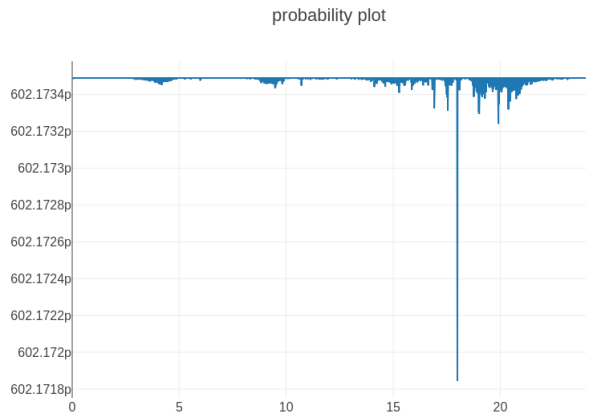
2 Pistes pour résoudre le problème

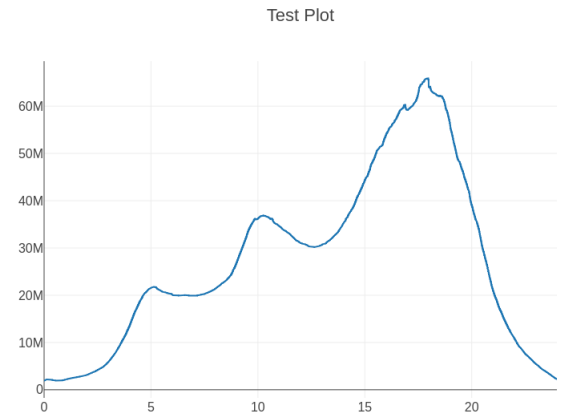
Nous allons séparer les données en trois flux afin de résoudre ce problème. Chaque flux apportera des features au problème.

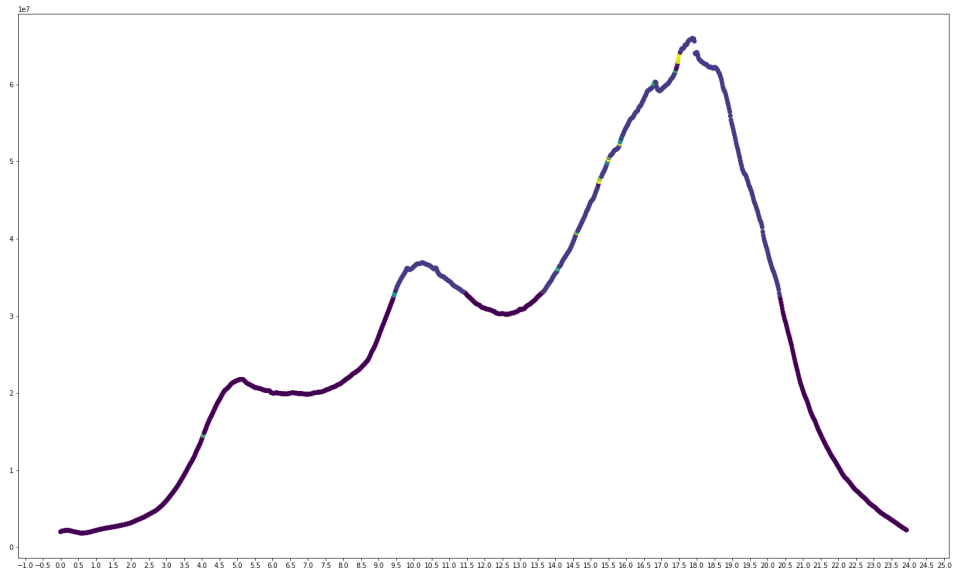
2.1 Real time series

les Real Time Series représente le nombre de téléspectateur devant une chaîne a une minute donnée. De ces RTS nous allons extraire des informations sur les tendances. Les usagers ne seront pas considérés au cas par cas mais comme un tout. Il s'agit donc d'un dataset bruité car on ne connait que les variations du nombre d'usager, et non pas le nombre de départ et le nombre d'arrivée (c'est une possibilité que l'on étudiera plus tard). De plus il s'agit d'une courbe relativement régulière ne contenant que 1400 points et qui présente une telle variation entre les extrema que les sauts (ou chutes) d'audience sont difficilement détectable. De ces RTS on va considérer chaque point en même temps que les trois précédents afin de lui donner du contexte mais pas de points après lui afin de ne pas supposer l'avenir (on pourrait le faire dans le cadre de l'annotation des événements pour détecter les publicités). ce Dataframe connait ensuite diverses modifications pour parvenir a un ensemble de features sur chaque points. On va ensuite mettre en place des thresholds afin de détecter des features anormaux permettant la détection d'événements anormaux. La mise en place de ces Threshold se fait pour le moment manuellement mais a terme devrait être appris de l'historique et l'importance de chaque feature devrait de même être fait automatiquement (ils ont pour

le moment tous le même poids). Une fois l'historique établi, un simple xgboost devrait permettre de tuner l'anomalie détection plus finement
NB : le fichier processing.py du dépôt git présente le code du traitement des RTS.
NB2 : des graphes d'un feature et d'une RTS et d'une RTS annotée







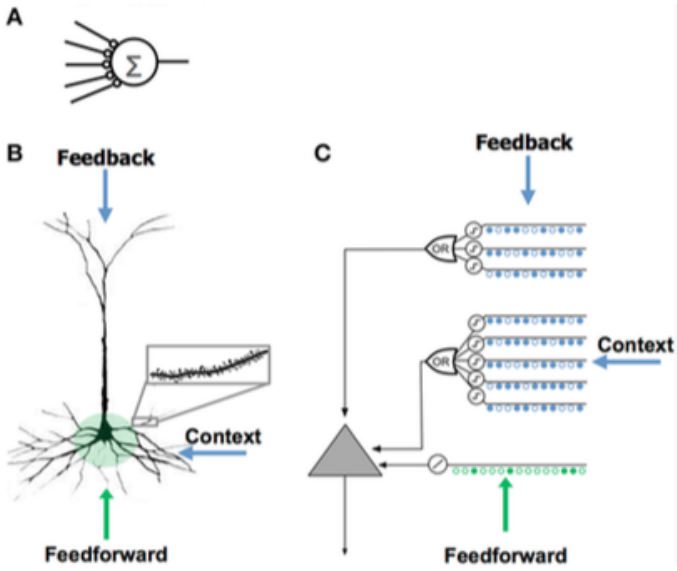
2.2 PROGRAMME TV

n'a pas été traité cette semaine

2.3 HITORIQUE DES UTILISATEURS

inexistant pour le moment

3 RÉCAPITULATIF DU MODÈLE



le modèle sera bâti en concordance avec la dichotomie des flux décidé auparavant : dans le schéma ci-dessus le contexte sera les tendances du moments, les caractéristiques du programme et l'historique d'utilisation du téléspectateur. Le Feedforward sera le flux actions que l'on reçoit en temps réel. Le feedback sera la ré-annotation de l'historique en fonction des décisions prises (algorithme de ré-apprentissage de fréquence lente) La première étape consiste a traiter les trois flux de données afin de repérer les publicités(dans le passé ou en temps réel). Cette détection permettra de mieux entrainer le traitement automatique des flux de données et ainsi une meilleur détection. Il s'agit d'une version customisée de L'EM algorithm et de d'un HTM [4]

Bibliographie

[1] Perin Pascal. Le zapping. In : Réseaux, volume 10, n°51, 1992. Sociologie des journalistes. pp. 117-125

[2] Chateau Dominique. L'effet zapping. In : Communications, 51, 1990. Télévisions / mutations. pp. 45-55 ;

[3] Television and Internet Commercial avoidance - Thales Teixeira University of Michigan 2009

[4] Unsupervised real-time anomaly detection for streaming data - Subutai Ahmad , Alexander Lavin , Scott Purdy , Zuha Agha

[5] Feature-based time-series analysis - Ben D. Fulcher

[6] Special Issue on Learning from Imbalanced Data Sets - Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Kolcz