MASTER DATA SCIENCE UNIVERSITÉ PARIS SACLAY - TÉLÉCOM PARISTECH

Rapport hebdomadaire de stage

Alexis Blanchet

8 juin 2018

1 Introduction

Les deux dernières semaines ont été employés a la recherche d'un moyen sur de reconstituer les programmes télévisés en utilisant les Change Point détectés précédemment. On va donc utiliser le contexte du point (différents features qui seront explicités ci-dessous) pour tenter de trouver a quoi correspond le-dit point. La difficulté ici repose sur la linéarité et l'interdépendance des prédictions. On prédit un point après l'autre en fonction du point précédant. C'est en effet ce dernier point qui permet de déterminer le contexte du nouveau point. Afin de parvenir a nos fins, nous commencerons tout d'abord par créer un historique fiable du contexte pour les points importants (Change Points). Pour ce faire, nous établirons a la main un arbre de décision qui sera testé sur les jours constituant l'historique sur. L'algorithme sera ensuite testé sur tous les jours du mois de décembre afin de voir s'il se comporte bien sur des jours qu'il n'a jamais vu (le week end par exemple). Une fois cet algorithme établi on cherchera a le remplacer par un algorithme de machine learning plus classique (on privilégiera au début des Tree Boosting comme XGBoost et CatBoost) cette seconde étape dépendant en majeur partie de la première on veillera a obtenir de bons résultats et a contrôler de manière précise les résultats intermédiaires.

2 Arbre de Décision

Nous avons Donc d'un côté les Programmes Télévisés officiels (PTV) d'un autre Chacun des 1437 points constituant la journée (les 3 premières minutes servent d'étalon) avec a chaque fois si c'est un Change Point ou non et sa probabilité de l'être. On va directement accoler a ce point son contexte :

1. Tout d'abord nous considèrerons les features liés uniquement à l'heure :
   - l'heure exacte (en minutes depuis 3 :00) à laquelle on se trouve.
   - Sa place dans la journée (fin de nuit,début de matinée,matinée,fin de matinée,midi,début d'après-midi,fin d'après-midi,soirée,Prime Time,Fin de soirée,nuit)
   - On aura aussi, mais pas pour le moment, les features liés a la journée : Vacances,Week-end,Journée Spéciale.
2. On considère ensuite les features liés au programme :
   - le type de programme (Jeu,magazine,Journal,Série)
   - la durée du programme
   - le nombre de Publicités que l'on pourrait rencontrer dans le programme
   - la fin prévisionnelle du programme
3. et enfin le plus difficile à obtenir sans un programme détaillé :
   - le pourcentage du programme déjà visionné
   - la partie du programme dans laquelle on se trouve (début, milieu,fin)
   - de quand date la dernière pub
   - de quand date la dernière fin de programme

- de quand date le dernier Change Point
- combien de Pubs il y a déjà eu depuis le début du programme
- combien de pubs il y a eu dans l'heure

Ces features sont sujets a évolution avec l'agrandissement régulier de la taille de l'arbre de décision afin d'assurer une meilleur précision. L'utilisation croisée de ces différents features permet de retrouver le programme télévisé tel que réellement diffusé.

l'arbre de Décision est construit sur les 4 jours disponibles ou je sais exactement ce qu'il s'est passé et sur lesquels je peux vérifier directement s'il y a des erreurs. La cross-validation se fait par la recherche d'incohérence sur les programmes télévisés prédit pour le mois de décembre de l'année dernière.

3 Data Exploration

Au cours de l'établissement de l'arbre de décision, il est apparu importants de noter pour la suite quelques remarques :

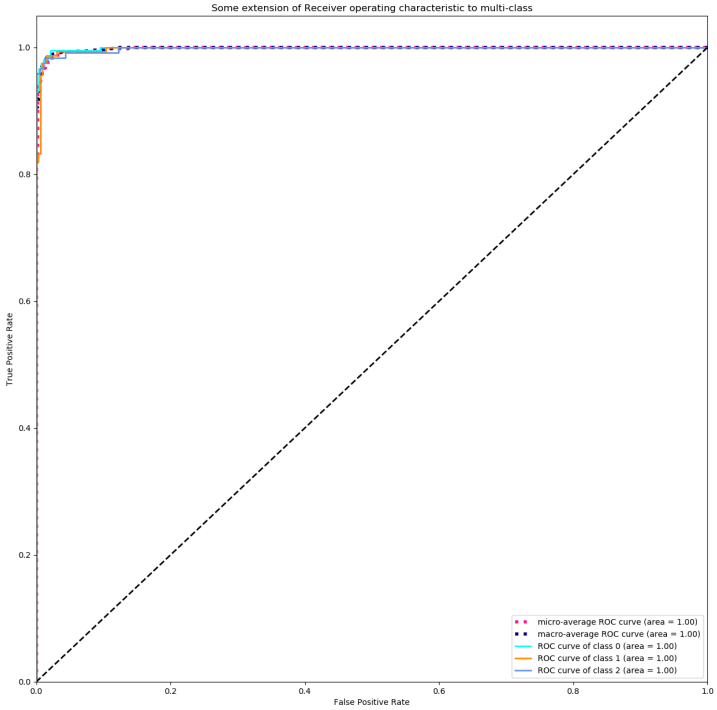
- Si l'on a une précision importante au détriment du recall, la reconstruction est impossible
- Si le recall est important au détriment de la précision, la cohérence finale du programme est faible
- entre les deux défauts ci-dessus il vaut mieux avoir un recall important
- Tout au long de la journée sera ajouté des mécanismes de rappels permettant de corriger, et au besoin de redémarrer l'algorithme afin que les erreurs ne s'accumulent pas. On pourra par exemple ajouter manuellement des points dans la liste des CP pour forcer une détection, Stopper l'algorithme et le redémarré s'il n'y a pas eu de détection depuis un certain temps, ou encore se baser sur des programmes phares et réguliers pour vérifié la cohérence du programme télévisé (par exemple les JT pour TF1)
- Les caractéristiques des nouveaux programmes ne sont pas connues a l'avance(nombre de pubs, fin en avance régulier,...) et ainsi seront sujets a des variations qu'il sera bon de suivre.
- Lors de certaines journées particulière les Journaux TV sont absents du programme Télévisé.
- Pour la Coupe du Monde de Football, en cas de prolongation, cette dernière sera précédée d'une publicité qui sera détectée comme une fin de programme (qui correspondra parfaitement à celle prévu) et ainsi il y aura changement de programme avec ajout d'un nouveau programme "Prolongation" non notée dans le programme TV
- L'algorithme est entrainé pour avoir une précision sur la journée. Il peut se tromper et compenser et cela n'est pas corrigé par mes soins. Ainsi, si le matin il y a un problème mais que le JT n'est pas décalé et qu'il n'y a pas d'incohérence majeure sur les programmes phares le précédent (12 coups de midi), alors rien n'est fait pour augmenter la précision ce qui reviendrait a faire de l'overfitting.

- Le Nouveau programme Télévisé repose entièrement sur la possibilité d'extraire un nombre suffisent d'informations de chacun des programmes. Une méconnaissance des caractéristiques du programme entraîne une variance énorme dans les résultats. En particuliers il est VITALE de connaître a l'avance le nombre de Pub qu'il y aura dans chaque programme. Cela est possible sur TF1 grâce au site de vente des espaces publicitaires qui donne les programmes avant et après chaque espace publicitaire.

4 SECOND LAYER OF MACHINE LEARNING

De l'algorithme ou sort un tableau contenant chaque Change Point détecté par la première couche avec son contexte et son label (0 faux CP, 1 publicité, 2 fin de programme). Il s'agit maintenant de faire découvrir a l'ordinateur la structure inhérente au tableau. On utilisera ici de nouveau des arbres de Décisions : 2 XGBoost, 2 CatBoost et trois algorithmes de Sklearn(DecisionTreeClassifier, MLPClassifier, RandomForestClassifier). A ce stade, seul l'exploration des algorithmes disponibles et la recherche des meilleurs paramètres est en cours. Classiquement, on en fera stack et une logistic regression sera sans doute appliquée a l'ensemble. Les premiers résultats sont présentés ici :

	classe 0	classe1	classe 2	overall	FP	FF
XGB	0.99 | 0.97 | 0.97	0.96| 0.98 | 0.98	0.95 | 0.96 | 0.96	0.97	3	2
CatBoost	0.99 | 0.95 | 0.96	0.96 | 0.98 | 0.97	0.93 | 0.98 | 0.97	0.97	1	3
DecisionTreeClassifier	0.94 | 0.97 | 0.97	0.97| 0.92 | 0.93	0.91 | 0.92 | 0.92	0.94	3	7
RandomForestClassifier	0.98 | 0.96 | 0.96	0.95| 0.96 | 0.96	0.94 | 0.96 | 0.96	0.96	3	4



Il est cependant important de noter que la présence de la moindre erreur dans le programme se répercute jusqu'au prochain Reset. Ainsi, quand bien même les résultats sont ici satisfaisants, ils peuvent engendrer si l'on fait la même erreur chaque jour une détérioration rapide des capacités prédictives de l'algorithme.

	minute	TITRE	Change Point	pourcentage vu	Évenement	Heure
0	180	Programmes de la nuit	Non	0.000000	Début de Détection	3:0
1	288	publicité	oui	0.619608	publicité dans un programme	4:48
2	377	Programmes de la nuit	oui	0.968627	fin d'un programme	6:17
3	484	publicité	oui	0.891667	publicité dans un programme	8:4
4	505	TFou	oui	1.066667	fin d'un programme	8:25
5	510	Météo	oui	1.000000	fin d'un programme	8:30
6	523	publicité	oui	0.288889	publicité dans un programme	8:43
7	566	Téléshopping	non	1.244444	fin non détectée d'un programme	9:26
8	572	Météo	non	1.200000	fin non détectée d'un programme	9:32
9	592	publicité	oui	0.666667	publicité dans un programme	9:52
10	608	Petits secrets entre voisins	oui	1.200000	fin d'un programme	10:8
11	621	publicité	oui	0.371429	publicité dans un programme	10:21
12	638	Petits secrets entre voisins	oui	0.857143	fin d'un programme	10:38
13	655	publicité	oui	0.485714	publicité dans un programme	10:55
14	667	Demain nous appartient	oui	0.828571	fin d'un programme	11:7
15	697	publicité	oui	0.545455	publicité dans un programme	11:37
16	709	Les feux de l'amour	oui	0.763636	fin d'un programme	11:49
17	715	Petits plats en équilibre	non	1.200000	fin non détectée d'un programme	11:55
18	749	publicité	oui	0.618182	publicité dans un programme	12:29
19	769	Les douze coups de midi	oui	0.981818	fin d'un programme	12:49
20	775	L'affiche du jour	non	1.200000	fin non détectée d'un programme	12:55
21	780	Journal	non	0.166667	--soft reset to avoid any error--	13:0
22	819	Journal	oui	1.300000	fin d'un programme	13:39
23	835	Petits plats en équilibre	non	1.600000	fin non détectée d'un programme	13:55
24	856	Météo	non	1.400000	fin non détectée d'un programme	14:16
25	893	publicité	oui	0.352381	publicité dans un programme	14:53
26	927	publicité	oui	0.676190	publicité dans un programme	15:27
27	963	En cavale pour Noël	oui	1.019048	fin d'un programme	16:3

	minute	TITRE	Change Point	pourcentage vu	Évenement	Heure
28	1007	publicité	oui	0.488889	publicité dans un programme	16:47
29	1027	publicité	oui	0.711111	publicité dans un programme	17:7
30	1041	Marié avant Noël	oui	0.866667	fin d'un programme	17:21
31	1078	publicité	oui	0.569231	publicité dans un programme	17:58
32	1095	Quatre mariages pour une lune de miel	oui	0.830769	fin d'un programme	18:15
33	1112	publicité	oui	0.261538	publicité dans un programme	18:32
34	1147	publicité	oui	0.800000	publicité dans un programme	19:7
35	1163	Mon plus beau Noël	oui	1.046154	fin d'un programme	19:23
36	1189	Demain nous appartient	oui	0.742857	fin d'un programme	19:49
37	1195	Météo	non	1.200000	fin non détectée d'un programme	19:55
38	1200	Journal	non	0.142857	--soft reset to avoid any error--	20:0
39	1236	Journal	non	1.028571	fin non détectée d'un programme	20:36
40	1242	My Million	non	1.200000	fin non détectée d'un programme	20:42
41	1247	Météo	oui	1.000000	fin d'un programme	20:47
42	1253	Nos chers voisins	non	1.200000	fin non détectée d'un programme	20:53
43	1262	C'est Canteloup	oui	0.900000	fin d'un programme	21:2
44	1309	publicité	oui	0.313333	publicité dans un programme	21:49
45	1344	publicité	oui	0.546667	publicité dans un programme	22:24
46	1371	publicité	oui	0.726667	publicité dans un programme	22:51
47	1393	publicité	oui	0.873333	publicité dans un programme	23:13
48	1411	Enfoirés Kids	oui	0.993333	fin d'un programme	23:31
49	8	publicité	oui	0.296000	publicité dans un programme	0:8
50	35	publicité	oui	0.512000	publicité dans un programme	0:35
51	76	publicité	oui	0.840000	publicité dans un programme	1:16
52	105	Vendredi, tout est permis avec Arthur	oui	1.072000	fin d'un programme	1:45
53	111	Tirage de l'Euro Millions	non	1.200000	fin non détectée d'un programme	1:51

	minute	TITRE	Change Point	pourcentage vu	Évenement	Heure
54	123	publicité	oui	0.041379	publicité dans un programme	2:3