Etude du comportement des consommateurs à l'aide des réseaux récurrents

Kais LARIBI

20 avril 2017

1 Introduction

Depuis quelques années, le nombre de transactions électroniques n'a cessé d'augmenter. Les cartes bancaires sont désormais très utilisées dans les commerces ce qui favorise la traçabilité et la collecte des données de payement. La disponibilité d'une quantité colossale de données de payement à pousser les scientifiques à réfléchir à des modèles pour analyser et prédire le comportement futur des clients. Leurs recherches sont d'une importance cruciale dans le monde du commerce vu qu'elles permettent de comprendre les besoins des clients et de mieux cibler leurs offres et leurs stratégies pour les maintenir. On se propose dans ce rapport, d'étudier le comportement des clients et d'expliquer un modèle de prédiction et de classification en s'appuyant sur les réseaux récurrents.

2 Etat de l'art sur l'analyse de comportement

Les réseaux récurrents ont montré leur efficacité dans la prédiction des comportements futurs des clients. Ce sont des réseaux qui intègrent la notion du contexte et de l'évolution séquentielle de l'information dans leurs décisions. En effet, plusieurs modèles à base de ces réseaux sont apparus dans la littérature pour traiter des sujets dont les données sont évolutifs tel que la prédiction des séries temporelles, le traitement de langage naturel ou encore l'analyse du comportement des consommateurs...
En ce qui concerne l'analyse de comportement, un modèle d'estimation de la préférence des clients d'un restaurant est décrit dans [1], il intègre les données de géolocalisation, l'historique de préférence, l'influence sur les réseaux sociaux ainsi que les trajectoires décrits... Le deep learning a été utilisé également pour classifier les clients qui sont susceptibles de quitter ou se désabonner d'une offre de téléphonie par exemple [2].
Ensuite, les indices R, F, M étaient utilisés dans [3] pour prédire les motifs d'achats des clients d'un magasin... Dans ce qui suit on s'intéressera à ces indices et à l'implémentation d'un modèle de classification de clients reposant sur leurs évolutions.

3 Indices économiques R, F, M

Dans notre sujet, à savoir l'analyse de comportement des clients, les jeux ce données disponibles sous forme bruts sont des transactions relatives à des achats faits par la clientèle d'un magasin. Cependant pour donner un sens à ces données et pouvoir entraîner des modèles d'apprentissage dessus ils faut bien les traiter pour en extraire des variables qui identifient un comportement. Dans ce contexte, On étudie les caractéristiques R,F et M (la récence, la fréquence et la quantité d'argent dépensé par un individu à une date donnée), on répond ainsi à trois questions sur le client : Quand est ce qu'il a acheté pour la dernière fois ? Combien de fois a t-il acheté ? Et quelle somme a t-il payé ?...

4 Approche considérée

L'étude des paramètres R, F et M et leurs évolutions au cours du temps permet de caractériser la relation avec un client. L'objectif de ce projet est donc d'implémenter un algorithme de classification de clients en se basant sur leurs historiques de transactions. L'algorithme en question sera entraîné

avec des données capturées dans des séquences de temps T_0, T_1,...,T_{n-1} pour prédire la classe à une date T_n. Les entrées du modèles récurrents sont séquentiels et prennent en compte l'évolution temporelles des paramètres R, F et M.

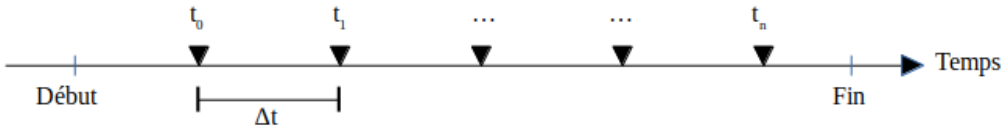

FIGURE 1 – Figure illustrant la segmentation de l'intervalle d'étude

5 Description de la Dataset

Dans ce projet on s'est servi d'une partie d'un jeu de données disponible sur kaggle dans la compétition 'acquire valued shoppers challenge'. Elle fait environ 22gb de taille et contient environ 350 millions de transactions. L'objectif étant de classifier en catégories les clients en considérant les paramètres R,F et M, on s'est intéressé alors qu'aux variables indispensables aux calculs de ces indices à savoir l'Identifiant du client, la date de la transaction et son montant. Pour des contraintes matérielles on a décidé aussi d'utiliser une version réduite qui fait 1,3gb de taille (Elle peut être extraite avec un script disponible sur kaggle également).

6 Traitement des données et Implémentation

6.1 Exploitation de la dataset

En vue d'étudier les comportements des clients et leurs évolutions dans le temps on définira sur la période d'étude des intervalles de temps de tailles égales ΔT (selon le problème étudié les subdivisions peuvent être quotidienne, hebdomadaire, mensuelle...). On itère ensuite sur ces subdivisions en considérant à chaque fois les transactions effectuées avant une date t_j et on calcule les valeurs R_{ij},F_{ij} et M_{ij} relatives à chaque client i.

Pour un client i :
— R_{ij} l'écart entre la dernière transaction effectuée et t_j.
— F_{ij} le nombre de transactions jusqu'à t_j
— M_{ij} la somme dépensée jusqu'à t_j
Ces données sont ensuite rangées sous forme d'un tenseur de dimension (n, p, q) qui servira à alimenter l'algorithme.
n : nombre de clients de la Dataset
p : nombre de paramètres à entrer (ici 3 : $R, FetM$)
q : nombre d'intervalles dans la période d'étude : E=$\frac{t_{n-1}-t_0}{\Delta T}$

Les $n-1$ premiers blocs sont utilisés pour entraîner le modèle. Les valeurs de R,F et M à T_n sont utilisées pour définir l'appartenance de chaque client à une classe. En effet, selon ces valeurs on classifie les clients en groupes ijk où $i, j, k \in [1..4]$. A titre indicatif, une valeur de récense élevée signifie que le client n'était pas vu depuis une période quasiment longue, il est donc classifier dans la catégorie 4 pour R. Par contre des valeurs élevées de F, M sont appréciées, elles traduisent la fidélité et les dépenses importantes des clients, ces derniers sont dans les catégories 4 pour F et M.

6.2 Modèle proposé

Le modèle proposé est constitué d'un <input layer> de type réseau récurrent, une couche cachée <dropout> et un <output layer> avec la fonction $softmax$ formé par 64 neurones correspondant aux 64 combinaisons possibles de R,F et M. Le modèle est entraîné à classifier les séquences de

taille $p \times q$ formées par les valeurs R_i,F_i et M_i à chaque date T_j aux 64 groupes en minimisant l'entropie croisée.

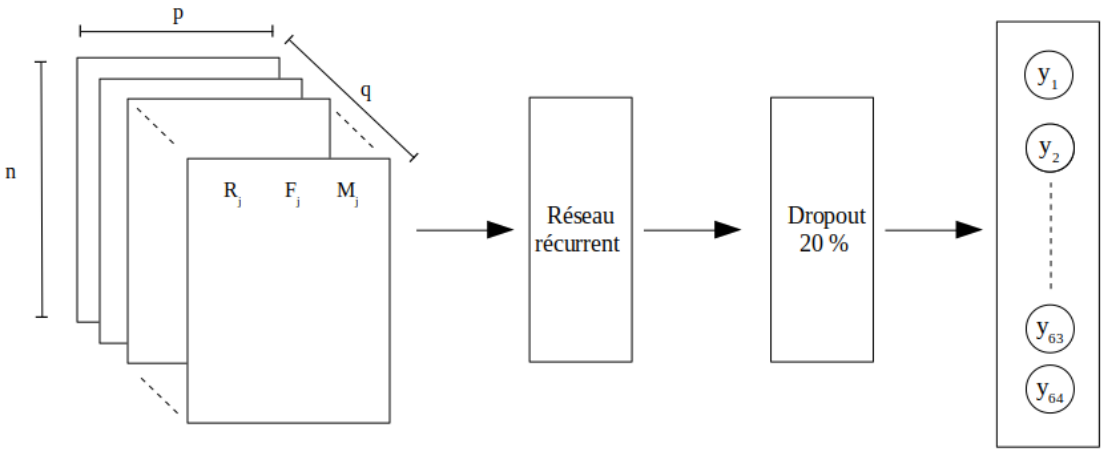

FIGURE 2 – Modèle de classification de clients avec des réseaux récurrents

6.3 Implémentation sous Keras

Keras fournit des classes prédéfinies pour construire des modèles de deep learning. En particulier les classes SimpleRNN, LSTM, GRU accessibles depuis le paquet *keras.layers.recurrent*. La subtilité réside dans la manière avec laquelle keras interprète les données pour faire apparaître et intervenir la notion du temps ou de l'évolution. Le fait de ranger les données en tenseur de taille (n, p, q) selon la description du paragraphe 6.1 permet de considérer l'évolution des p paramètres sur les q intervalles. Finalement, les valeurs des classes sont encodées selon le processus <one hot encoding> et présentées en sortie.

6.4 Evaluation

En premier lieu et lors du traitement de la dataset on a choisi les clients présents depuis le début de l'étude pour avoir des séquences de mêmes longueurs. Mais, en vrai les réseaux récurrent peuvent interpréter l'absence de l'information en les représentant avec des séquences nulles. Il est donc possible d'extraire un nombre plus important de clients à partir de la dataset. Ça serait le focus de la deuxième partie de l'évaluation...

6.4.1 Utilisation des séquences complètes non nulles :

L'algorithme est d'abord entraîné sur 67 % des données de la dataset (environ 153000 exemples). Il est ensuite évalué sur la classification des exemples restants. Le tableau suivant récapitule les résultats obtenus pour les 3 réseaux :

	Simple RNN	LSTM	GRU
Training Acc	89%	92%	89%
Error (Test set)	10%	8%	13%

6.4.2 Utilisation des séquences en partie nulle :

On se propose dans cette partie d'évaluer le modèle sur des séquences de tailles variées et ceci en considérant des clients qui apparaissent notamment à une date ultérieure à la date de début de l'étude. Les données utilisées dans cette évaluation sont générées à partir de la même Dataset de départ. Dans cette évaluation les clients possèdent au maximum 4 séquences nulles c'est à dire qu'ils sont apparus avant la limite supérieure de la 4ème subdivision de l'intervalle d'étude. L'algorithme est entraîné alors sur environ 181000 exemples et évalué sur la classification de 89000 cas.

	Simple RNN	LSTM	GRU
Training Acc	88%	90%	87%
Error (Test set)	7%	29%	30%

Il est remarquable que le modèle LSTM se comporte bien vis à vis des séquences complètes de données. L'erreur de classification devient importante lorsque les séquences sont de tailles différentes ce qui est le cas du GRU aussi. Par contre, le récurrent simple prouve son efficacité dans ce genre de problème de classification en ayant un taux d'erreur de 7% sur l'ensemble de la test set.

6.5 Conclusion :

On a décrit dans ce rapport une approche pour expérimenter les réseaux récurrents et les appliquer sur l'étude du comportement de la clientèle. Ceci en considérant un problème de classification et en étudiant l'évolution des paramètres économique R,F et M. A priori, les résultats obtenus avec le SRN sont prometteuses surtout que les suppositions faites sont très proches de la réalité (prise en compte de l'évolution temporelle, séquences de tailles variées...). D'autres pistes peuvent être explorées dans ce cadre tel que la prédiction des séquences futures de R, F et M et ainsi prédire à plus long terme le comportement des clients.

Références

[1] S.S. Lam S.W. Yoon N.Gnanasambandam B. Zheng, K. Thompson. Customers behavior prediction using artificial neural network. *IIE Annual conference. Proceedings*.

[2] J.Zaratiegui A.Vazquez F.Castanedo, G.Valverde. Using deep learning to predict customer churn in a mobile telecommunication network.

[3] S.Rahnamayan H.Salehinejad. Customer shoppig pattern prediction : A recurrent neural network approach. *IEEE*, 2016.