Représentation de données temporelles par un modèle à échelle logarithmique

Brieuc Conan-Guez, Alain Gély, Lydia Boudjeloud-Assala, Alexandre Blansché, Dario Compagno, Arnaud Mercier

Représentation de données temporelles par un modèle à échelle logarithmique

Brieuc Conan-Guez*, Alain Gély*, Lydia Boudjeloud-Assala*, Alexandre Blansché*,
Dario Compagno**, Arnaud Mercier***

*LITA, université de Lorraine,
{brieuc.conan-guez, alain.gely, lydia.boudjeloud-assala, alexandre.blansche}@univ-lorraine.fr
http://www.lita.univ-lorraine.fr/
**CIM, Université Sorbonne Nouvelle - Paris 3
***CARISM, université Panthéon-Assas - Paris 2

Résumé. Dans ce travail, nous nous intéressons aux tweets associés à un événement source (publication d'articles en ligne), et nous représentons la distribution temporelle des tweets par un modèle à résolution variable : haute résolution au début de la série temporelle et plus grossière à la fin. À cet effet, nous proposons d'utiliser un modèle à noyaux basé sur une transformation logarithmique des données.

1 Introduction

Dans ce travail, nous nous intéressons à des familles de séries temporelles particulières, initiées par la survenue d'un événement « source ». Il ne s'agit donc pas d'étudier des séries temporelles sur une fenêtre arbitraire (un mois, un an, etc.) mais d'observer la survenue d'un ensemble d'événements en réponse à l'événement initial. Ce travail a d'abord été motivé par l'étude de la dissémination d'information sur les réseaux sociaux. Plus particulièrement, on s'intéressera à la vie d'un article de presse sur Twitter. L'événement source, la publication d'un article en ligne, étant suivie par une série temporelle d'autres événements : les tweets mentionnant cet article de presse. Cependant, ce travail est transposable à de nombreux cas. On pourra par exemple penser à la dynamique des échanges sur une liste de diffusion (l'événement initial étant la diffusion d'un message polémique, les événements suivants les réponses et commentaires sur ce message) ou encore aux occurrences d'interventions d'équipes de secours (police, samu, etc.) après une crise. (Chae et al., 2014) ont d'ailleurs étudié les occurrences de tweets en réponse à une catastrophe naturelle.

Tous ces exemples ont pour point commun d'avoir un événement initial (l'événement source) qui va engendrer une succession d'autres événements (événements réponses). Dans tous les cas, on peut penser que l'observation des événements réponses proches de la source (de l'ordre de la minute) n'aura pas le même poids dans l'analyse que celle des événements lointains (de l'ordre du mois).

Afin de chercher à dégager les grandes familles de comportements de partage d'articles, nous nous proposons d'utiliser un algorithme de classification non supervisée. Pour ce faire,

il faut donc trouver une représentation des données qui capture l'information finement lorsque l'on est proche de l'événement initial, tout en synthétisant l'information pour les survenues plus tardives, et cela de façon à avoir une description simple et compacte (vecteur de description) de la série temporelle.

Nous proposons une transformation des données de type logarithmique de façon à avoir un niveau de détail d'autant plus élevé que les événements sont proches de leurs sources, tout en permettant de garder trace des événements réponses plus lointains. La transformation logarithmique des données est une approche classique, qui est notamment évoquée dans (Silverman, 1986), quant au modèle à noyaux que nous nous proposons d'utiliser pour modéliser la distribution temporelle des événements réponses, ses propriétés théoriques sont étudiées en détail dans (Charpentier et Flachaire, 2014).

2 Représentation des données

Dans ce travail, nous nous intéressons à un corpus d'articles publiés entre avril et septembre 2014 (soit sur une durée de 5 mois) dans trois media différents : le Monde, le Figaro et Libération. À partir d'avril 2014 et ceci pendant 6 mois (soit jusqu'à un mois après la publication du dernier article), l'ensemble des tweets se rapportant à ces articles a été collecté. Cette durée de collecte assure que pour chaque article, nous disposons d'au moins un mois de données (pour les articles parus fin septembre), et d'au plus 6 mois (pour les articles parus au début de la collecte). Comme nous ne souhaitons analyser que les articles ayant eu un fort impact sur Twitter, les articles pour lesquels le nombre de tweets est inférieur à 100 ont été écartés de l'étude. Le corpus compte donc finalement 6 177 articles pour un total de 822 775 tweets.

Nous disposons de la date et de l'heure de la publication de chaque article ainsi que de celles de chaque tweet. Afin d'étudier les usages de publication de tweets une fois l'article paru, nous nous intéresserons dans la suite de l'étude non pas à la date des tweets, mais à l'intervalle de temps entre la date de publication des articles et la date de publication des tweets associés. On pose T^j la variable aléatoire modélisant la distribution temporelle des tweets se rapportant à l'article j. Chaque article est donc représenté par un vecteur de k_j réalisations $(t_1^j, \ldots, t_i^j, \ldots, t_{k_j}^j)$ de T^j, où k_j est le nombre de tweets se rapportant à l'article j. La valeur de k_j est dans la pratique comprise entre 100 et 200.

Comme indiqué dans l'introduction, nous souhaitons obtenir une représentation de chaque article qui réponde aux deux exigences suivantes :

- la représentation doit être compacte, i.e. chaque article ne doit être décrit que par un petit nombre de valeurs (de l'ordre d'une vingtaine). Ce qui facilitera premièrement les interprétations, et autorisera l'utilisation de méthodes dont le coût est très dépendant de la dimension des données (calcul de la déformation temporelle dynamique (DTW) entre deux séries par exemple) ;
- l'autre exigence à laquelle nous devons répondre nous est imposée par la nature même du problème : nous souhaitons garder un haut niveau de détail lors des premiers instants suivant la parution de l'article (résolution d'analyse de l'ordre de la minute), alors qu'une description relativement grossière est souhaitée dés lors que nous nous intéressons à des instants suffisamment éloignés de la date de parution de l'article (résolution d'analyse de l'ordre du mois pour la fin des séries temporelles).

Afin de répondre à ces deux exigences, nous nous proposons de représenter la distribution temporelle des tweets de chaque article par un modèle à noyaux à échelle variable : le modèle autorisera une représentation fine (resp. grossière) du début (resp. fin) de la série temporelle.

Le modèle que nous utilisons dans cette étude a été proposé dans (Charpentier et Flachaire, 2014) pour la modélisation de distributions à queue non exponentiellement bornée (*heavy-tailed distribution*). Les auteurs citent comme exemple de telles distributions la distribution des revenus des ménages au Royaume-Uni.

Pour construire ce modèle, nous considérons une fonction croissante h qui effectue un changement d'échelle non uniforme de l'axe temporel : contraction peu importante du temps au début de la série, contraction du temps importante vers la fin de la série. Dans la pratique, une fonction logarithmique $h(t) = \log(t+1)$ sera utilisée afin d'obtenir les résolutions d'analyse souhaitées. La construction du modèle est alors simple : estimer un modèle à noyaux non pas sur les données initiales, les $(t_i^j)_{1 \leq i \leq k_j}$, mais sur les données transformées $(h(t_i^j))_i$, puis reprojeter cette estimation dans l'espace initial.

Nous notons $f^j(t)$ la fonction densité de la variable aléatoire T^j. Pour produire une estimation de cette densité, nous nous intéressons à la distribution de la variable aléatoire $h(T_i^j)$, que nous estimons grâce à un modèle à noyaux classique $\hat{n}^j(.)$. Une estimation de la densité $f^j(t)$ est alors donnée par (Charpentier et Flachaire, 2014) :

$$\hat{f}^j(t) = \frac{\hat{n}^j(h(t))}{|(h^{-1})'(h(t))|} = \hat{n}^j(h(t))h'(t)$$

où h^{-1} est la fonction réciproque de h, et h' sa dérivée. Pour $h(t) = \log(t+1)$, on obtient donc $\hat{f}^j(t) = \hat{n}^j(\log(t+1))/(t+1)$.

On peut noter qu'une approche relativement équivalente, mais ne nécessitant pas d'estimation dans un espace transformé est envisageable. Il est en effet possible de construire un modèle à noyaux dans l'espace initial avec des espacements entre noyaux et des largeurs de noyaux variables. Cette approche nécessite cependant un développement informatique spécifique, alors que l'approche basée sur la transformation des données par h permet de s'appuyer sur des bibliothèques informatiques éprouvées.

Dans la pratique, nous souhaitons obtenir une représentation vectorielle de cette distribution. Nous discrétisons chaque $\hat{f}^j(t)$ sur l'intervalle $[0, t_{sup}]$ en 25 points (t_{sup} peut être choisi comme la valeur du dernier tweet ou par un quantile). Pour obtenir les points de discrétisation de $\hat{f}^j(t)$, l'intervalle $[0, h(t_{sup})]$ est tout d'abord uniformément discrétisé en 25 points, puis chacun de ces points est reprojeté dans l'espace initial par la transformation réciproque h^{-1}.

Pour le paramètre de lissage (largeur des noyaux dans l'espace transformé), nous avons choisi une valeur unique pour tous les modèles à noyaux. La régle de Silverman (voir (Silverman, 1986), équation (3.31), page 48) a été utilisée pour déterminer une largeur adaptée pour chaque modèle, puis une valeur proche de la médiane a finalement été sélectionnée.

Sur la figure 2, on retrouve pour quatre articles la densité estimée par le modèle. En abscisse, la valeur des 25 points de discrétisation est indiquée (la première valeur n'est pas reportée). Les données sont décrites avec une résolution de l'ordre de la minute pour les premières variables, et de l'ordre du mois pour les dernières (56 jours séparent les deux derniers points de discrétisation).

Chaque point rouge au dessous des courbes représente un tweet. Une échelle logarithmique est utilisée pour représenter les données sur l'axe des abscisses. De fait, sur cette visualisation,

une répartition de tweets peut paraître uniforme, mais il n'en est rien : deux graduations successives ne représentent pas un même intervalle de temps en fonction de leur distance à l'origine.

3 Classification non supervisée

Après la discrétisation des modèles à noyaux, nous obtenons un tableau de données de 6 177 individus (les articles) et de 25 variables (les intervalles de temps). Nous avons appliqué une classification ascendante hiérarchique sur les articles afin d'identifier des comportements types de circulation d'événements en réponse à la publication d'un article. La distance euclidienne ainsi que le critère de Ward ont été utilisés. À partir de cette classification hiérarchique, une partition à 12 classes a été extraite. Le choix de 12 classes a permis d'obtenir une bonne homogénéïté dans chaque classe.

La figure 1 permet de juger de la qualité de représentation des données, et d'analyser les résultats de la classification. Sur cette *heatmap*, les valeurs jaunes correspondent à des valeurs élevées de la densité, alors que les valeurs rouges correspondent à des valeurs proches de 0. On distingue des différences de comportements à différentes échelles. Tout d'abord, des articles pour lesquels l'activité Twitter est concentrée sur les premiers instants suivant la publication de l'article (approximativement durant la première heure). Ensuite, les articles pour lesquels l'activité s'étend sur plusieurs heures, voire même sur plusieurs jours.

Afin de mieux décrire les différents comportements, nous proposons les *heatmaps* restreintes à certaines classes. La figure 3 décrit les classes 9, 3, 7 et 1 dont les centroïdes sont représentés par la figure 2. Pour ces classes, on peut extraire les caractérisations suivantes :

- le centroïde de la classe 9 montre un pic unique. L'échelle logarithmique sur l'axe des abscisses masque l'aspect asymétrique de la distribution. Cette classe se caractérise par une majorité de tweets publiés très tôt après la publication de l'article associé : la moyenne des médianes des articles se situe à 1h. La dispersion est très faible : l'écart interquartile moyen est de 2h ;
- le centroïde de la classe 3 montre une dispersion plus importante que celui de la classe 9. Cette classe se caractérise par une arrivée un peu plus tardive des tweets : la médiane moyenne est à 6h. La dispersion est relativement importante : l'écart interquartile moyen est de 20h ;
- le centroïde de la classe 7 montre l'extrême sensibilité aux occurrences en début de distribution : un seul tweet provoque un pic malgré le lissage du modèle à noyaux. Après ce pic initial, l'activité Twitter lié à l'article se maintient plusieurs heures. La classe 7 se caractérise par une majorité de tweets publiés tardivement : la médiane moyenne se situe à 11h. La dispersion est importante : l'écart interquartile moyen est de 27h ;
- la classe 1 se caractérise par de nombreux articles publiés vers minuit (au moins un quart). Les premiers tweets sont publiés plusieurs heures plus tard : en moyenne presque 6h après la publication de l'article. La médiane moyenne se situe à 33h, valeur beaucoup plus importante que pour les autres classes, ce qui indique une activité très décalée dans le temps. La dispersion est très importante : l'écart interquartile moyen est de 42h.

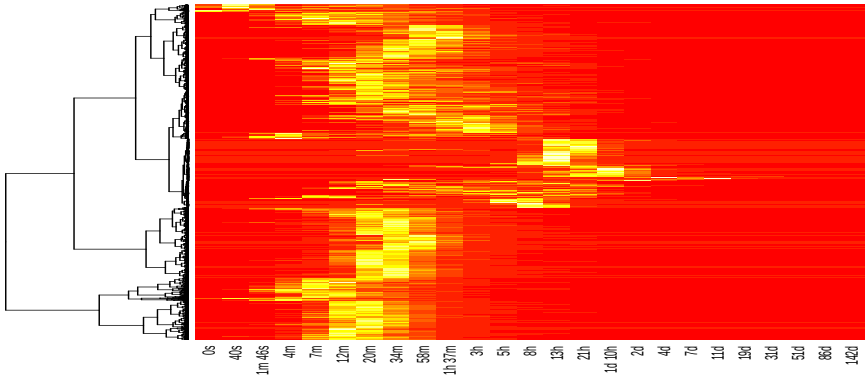

FIG. 1 – *Heatmap de la classification (la couleur jaune correspond à des valeurs élevées)*

4 Conclusion

Les premiers résultats de ce travail exploratoire sont très encourageants. Il reste notamment un travail d'interprétation et de description des classes en intégrant les méta-données associées aux tweets et aux articles. En effet, pour ce travail préliminaire, nous avons uniquement utilisé l'horodatage de publication des articles et des tweets, il serait intéressant de s'appuyer par exemple sur la thématique de l'article ou de différencier les tweets des retweets.

D'un point de vue technique, nous avons testé une représentation des données Twitter basée sur une échelle logarithmique. Celle-ci permet d'adapter la résolution d'analyse de façon à privilégier un niveau de détail plus élevé pour les tweets proches de la publication de l'article. Il nous reste à investiguer l'influence de trois paramètres : le paramètre de lissage des noyaux, la dimension des vecteurs de description, et la métrique utilisée. Actuellement, la classification obtenue utilise une métrique euclidienne, nous envisageons de tester par la suite d'autres dissimilarités, et notamment DTW très adaptée à la comparaison de séries temporelles.

Références

Chae, J., D. Thom, Y. Jang, S. Kim, T. Ertl, et D. S. Ebert (2014). Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics 38*, 51–60.

Charpentier, A. et E. Flachaire (2014). Log-Transform Kernel Density Estimation of Income Distribution. working paper or preprint.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London : Chapman & Hall.

Représentation de données temporelles par un modèle à échelle logarithmique

Summary

In this work, we are interested in tweets associated to a source event (online publication), and we seek to represent the temporal distribution of tweets thanks to a model with a variable resolution: high resolution at the beginning of the time serie, and coarse resolution at the end. For this purpose, we use a kernel model based on a logarithmic transformation of data.

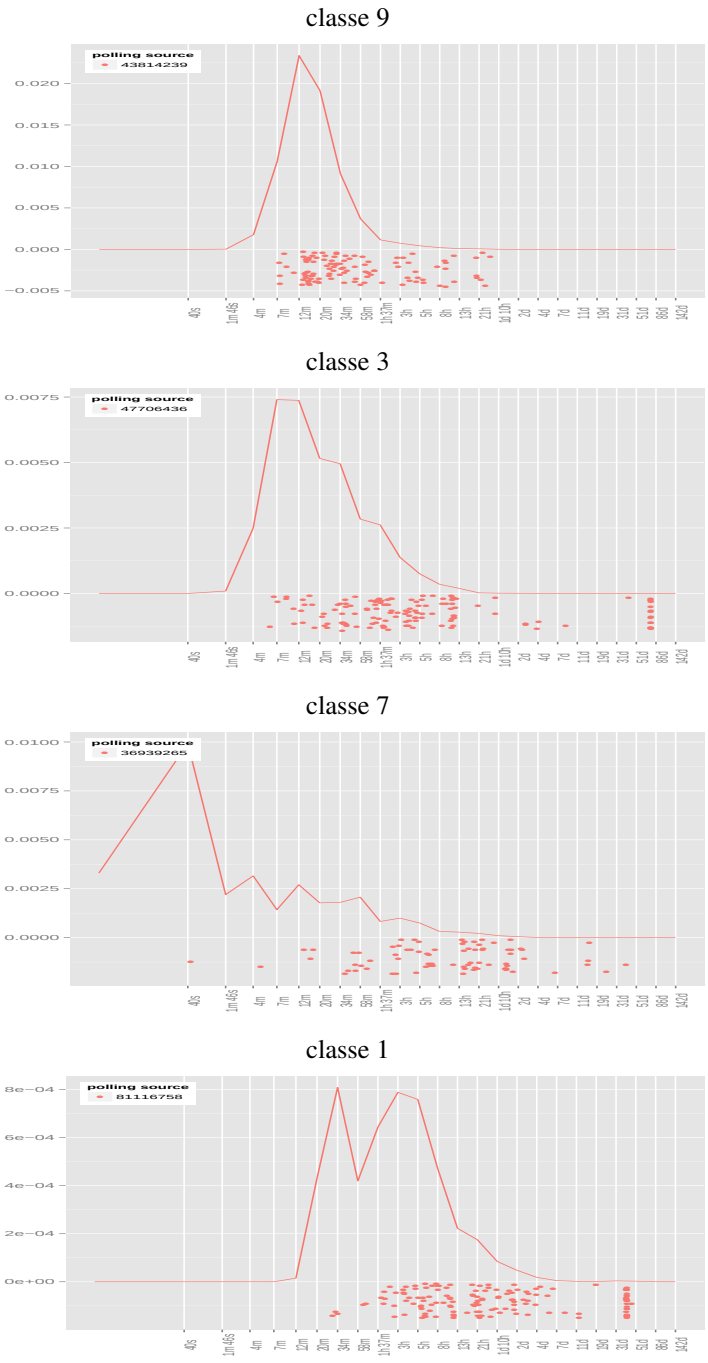

FIG. 2 – *Centroïdes des classes 9, 3, 7 et 1*

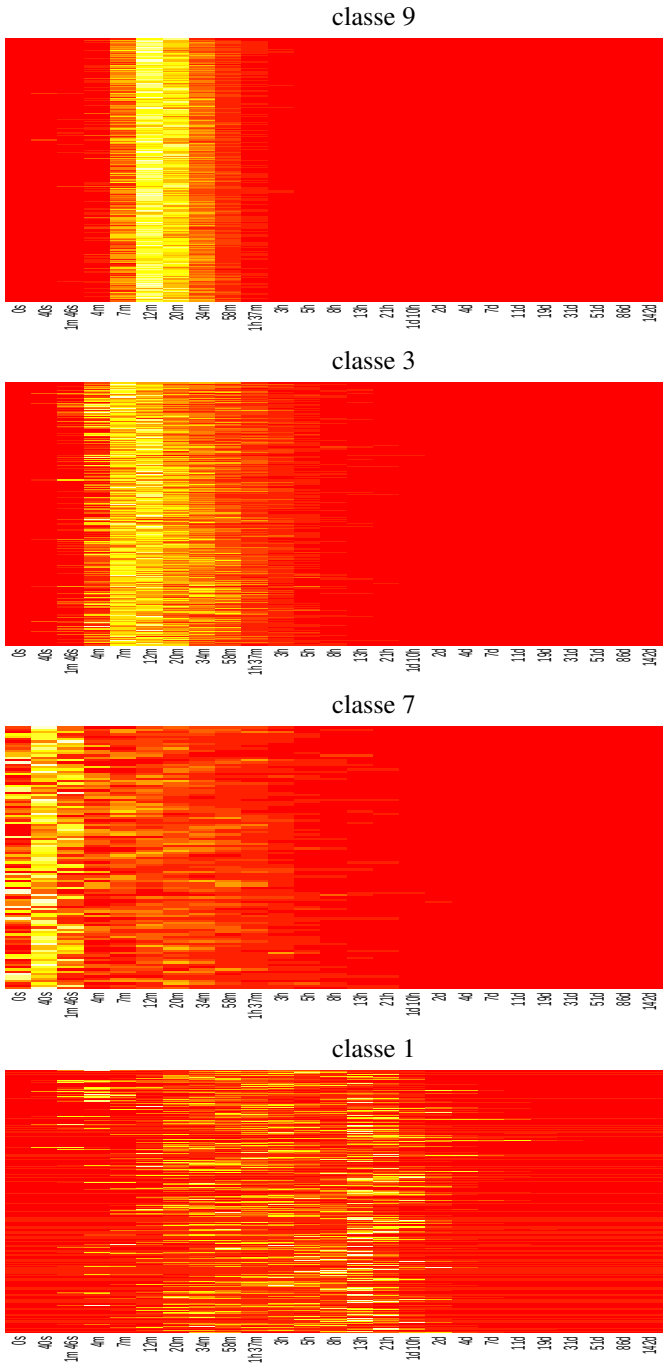

FIG. 3 – *Heatmap des classes 9, 3, 7 et 1*