Prédiction du montant levé lors d'une campagne de financement participatif par la méthode des plus proches voisins

Alexandre Blansché, Dylan Da Conceicao et Dylan Koby

Laboratoire LITA (EA 3097)
Université de Lorraine
Île du Saulcy 57045 Metz, France
alexandre.blansche@univ-lorraine.fr
http://www.lita.univ-lorraine.fr/

Résumé. Le financement participatif est un mode de financement d'un projet faisant appel à un grand nombre de personnes, contrairement aux modes de financement traditionnels. Il a connu une forte croissance avec l'émergence d'Internet et des réseaux sociaux. Cependant plus de 60 % des projets ne sont pas financés, il est donc important de bien préparer sa campagne de financement. De plus, en cours de campagne, il est crucial d'avoir une estimation rapide de son succès afin de pouvoir réagir rapidement (restructuration, communication) : des outils de prédiction sont alors indispensables. Nous proposons dans cet article une méthode de prédiction du montant final levé lors d'une campagne de financement participatif utilisant l'algorithme k-NN : en utilisant l'historique de campagnes passées, nous déterminons celles qui sont les plus similaires à une campagne en cours. Nous utilisons alors les montants finaux pour faire une estimation. Nous comparons plusieurs mesures de distance pour déterminer les plus proches voisins. Nos résultats indiquent que le dernier état d'une campagne seul est suffisant pour obtenir une bonne prédiction.

1 Introduction

Le financement participatif (*crowdfunding*), qui consiste à faire appel à un grand nombre de personnes pour financer un projet (contrairement aux modes de financement traditionnels), a connu une forte croissance avec l'émergence d'Internet et des réseaux sociaux. Kickstarter (`https://www.kickstarter.com/`) est un des sites de financement participatif les plus populaires. Depuis sa création en 2009, il a permis de lever plus de deux milliards de dollars américains dans des domaines variés. Cependant le nombre de participants d'une campagne est incertain et plus de 60 % des projets ne sont pas financés, il est donc important de bien préparer sa campagne de financement pour réaliser son projet. De plus, en cours de campagne, l'utilisation d'outils de prédiction est nécessaire pour avoir une estimation rapide de son succès afin de pouvoir réagir rapidement (restructuration, communication).

Nous proposons dans cet article une méthode de prédiction du montant final levé lors d'une campagne de financement participatif utilisant l'algorithme k-NN : en utilisant une base de données contenant l'historique de campagnes passées, nous déterminons celles qui sont les plus similaires à une campagne en cours. Nous utilisons alors les montants finaux pour faire une estimation. Lors des expérimentations, nous comparons plusieurs méthodes d'estimation ainsi que plusieurs mesures de distance.

Par la suite, nous présentons un état de l'art des méthodes d'analyse de séries temporelles (section 2). Nous expliquons ensuite notre approche (section 3) puis présentons nos résultats expérimentaux (section 4) avant de conclure (section 5).

2 État de l'art

Il existe beaucoup de méthodes d'analyse de séries temporelles, il s'agit souvent de prédire l'évolution d'une série en fonction de son historique par l'analyse de l'autocorrélation comme ARIMA (Taylor, 2008), mais aussi les réseaux de neurones (Frank et al., 2001) et le *Deep Learning* (Prasad et Prasad, 2014). Concernant le financement participatif, les méthodes se limitent souvent à prédire le succès ou l'échec des projets (Li, 2016; Etter et al., 2013) alors que le seuil de financement est parfois inférieur au montant espéré par le créateur du projet : celui-ci peut vouloir afficher 100 % de financement rapidement pour augmenter sa notoriété (les contributeurs privilégient souvent les campagnes réussies) sans que cela corresponde à ses besoins. De plus, en cas d'échec d'une campagne, aucune somme n'est versée et le capital investi en amont est perdu. Le créateur peut s'assurer un remboursement partiel en baissant son seuil de financement.

On peut décomposer l'évolution d'une campagne en n états uniformément répartis dans le temps. On notera $t_i(c)$ le montant levé à l'état i d'une campagne c. L'objectif est de prédire la valeur finale $t_n(c_0)$ d'une campagne c_0 en cours en ne connaissant que les valeurs $t_1(c_0)$ à $t_i(c_0)$, avec $i < n$. Pour le créateur d'un projet, il est avantageux d'avoir une estimation précise $\hat{t}_n(c_0)$ de $t_n(c_0)$ le plus tôt possible dans la campagne (avec un i le plus petit possible).

Si les sommes levées étaient uniformément réparties durant la campagne, une approche naïve consisterait à faire une approximation linéaire. On définit alors $\hat{t}_n(c_0) = 100 \times \frac{t_i(c_0)}{i}$. Cette hypothèse n'est cependant pas vérifiée. Sur la figure 1 (évolution des contributions selon le temps) on remarque que les apports en début et fin d'une campagne sont plus importants que pendant son déroulement (surtout en cas de succès).

3 Méthode proposée

La première approche proposée dans Etter et al. (2013), basée sur la méthode k-NN (Cover et Hart, 2006), est efficace, simple et n'utilise qu'une seule information (l'évolution du montant levé). Néanmoins l'approche ne prédit pas le montant levé lors d'une campagne mais prédit le succès ou l'échec de celle-ci. Dans cet article, nous proposons d'étendre l'approche pour obtenir une méthode de prédiction du montant levé en conservons donc la même méthodologie. On considère un ensemble de m séries temporelles, chacune correspondant à l'évolution du montant levé lors d'une campagne

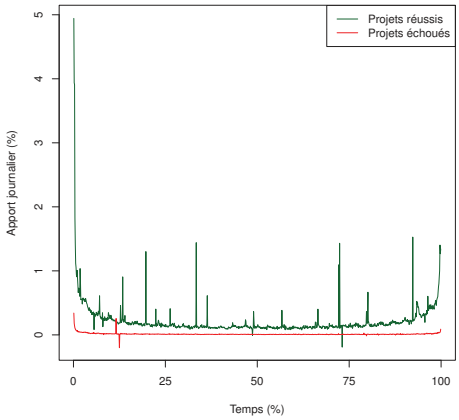

FIG. 1 – *Moyenne de l'apport financier à chaque état*

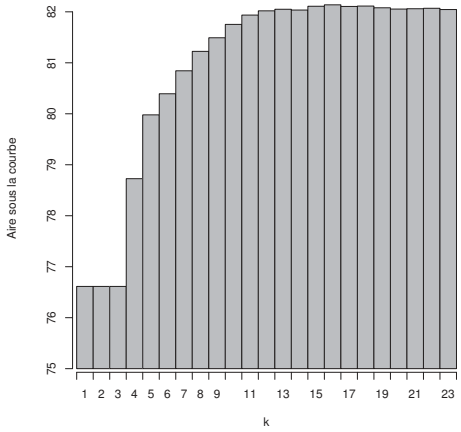

FIG. 2 – *Détermination de k*

de financement participatif. Les campagnes, ayant des durées variables, sont normalisées par un ré-échantillonnage en un nombre (n) fixe d'états équitablement répartis. Le montant levé est également normalisé en divisant par le seuil de financement. Pour une campagne en cours (à l'état i), les k campagnes les plus proches sur la période correspondante (entre 1 et i) sont déterminées et le succès d'une campagne est estimé selon le succès de ses « voisines » (vote à la majorité).

Nous proposons d'utiliser cette approche pour estimer le montant levé final en faisant de la régression par k-NN (Altman, 1992). Nous déterminons les k plus proches voisins (c_1 à c_k) d'une campagne c_0 sur les i premiers états des campagnes et nous utilisons les états finaux $t_n(c_1)$ à $t_n(c_k)$ pour obtenir une estimation $\hat{t}_n(c_0)$ de $t_n(c_0)$.

Une approche simple consiste à calculer une moyenne des valeurs $t_n(c_1)$ à $t_n(c_k)$. On définit alors $\hat{t}_n(c_0) = \frac{1}{k}\sum_{l=1}^{k} t_n(c_l)$. Cependant si la campagne c_0 surpasse toutes les campagnes c_1 à c_k à l'état i, il est probable qu'elle les surpasse à l'état n. La moyenne ne semble donc pas être un estimateur efficace et deux autres propositions vont tenter de corriger cela. On définit $\mu_i = \frac{1}{k}\sum_{l=1}^{k} t_i(c_l)$ la moyenne des montants levés par les k campagnes voisines à l'état i. La deuxième proposition (*shift*) consiste à calculer $\delta_i = t_i(c_0) - \mu_i$. On définit alors $\hat{t}_n(c_0) = \frac{1}{k}\sum_{l=1}^{k} t_n(c_l) + \delta_i$. La troisième proposition (*coeff*) consiste à calculer $\alpha_i = \frac{t_i(c_0)}{\mu_i}$. On définit alors $\hat{t}_n(c_0) = \frac{1}{k}\sum_{l=1}^{k} t_n(c_l) \times \alpha_i$.

Nous pouvons également nous interroger sur la mesure de distance à employer pour déterminer les k plus proches voisins. Dans Etter et al. (2013), les auteurs comparent les campagnes en utilisant la distance euclidienne sur les états 1 à i des séries temporelles. Cependant, l'évolution d'une campagne étant irrégulière, il peut être préférable d'utiliser une métrique plus flexible comme le *Dynamic Time Warping* (Sakoe et Chiba, 1971) qui s'est montré efficace pour comparer des séries temporelles dans des domaines variés. Enfin, on peut se demander s'il est nécessaire de mesurer la distance depuis le début de la série : nous pouvons calculer la distance selon l'état i uniquement.

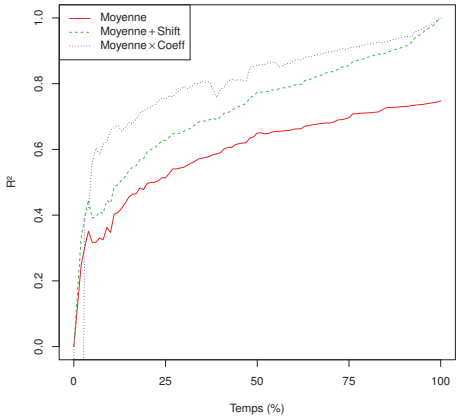

FIG. 3 – *Méthode d'estimation pour la prédiction du montant levé*

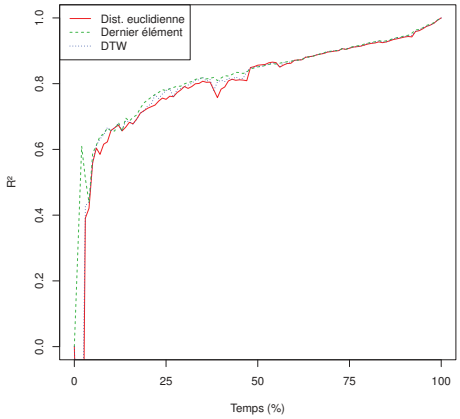

FIG. 4 – *Mesure de distance pour la prédiction du montant levé*

4 Expérimentations

Dans Etter et al. (2013), les auteurs ont constitué une base de données portant sur 16042 campagnes de financement datant de 2012 et 2013. Les séries temporelles ont été normalisées par un ré-échantillonnage en 1000 états. S'agissant d'un travail préliminaire et exploratoire, nous avons travaillé sur un échantillon de 500 campagnes sur 100 états, afin de réduire les temps de calcul. Nous avons mené plusieurs expérimentations afin d'évaluer par *bootstrap* l'efficacité de notre approche et l'influence de plusieurs paramètres de la méthode : nombre de voisins k de k-NN, méthode d'estimation du montant levé, mesure de distance. Toutes les combinaisons de paramètres ont été testées mais par soucis de clarté, nous présentons des expérimentations séparées.

Nous avons d'abord fait varier k entre 1 et 23 afin de déterminer la valeur optimale. Nous avons évalué le coefficient de détermination (R^2) pour chaque état de la campagne (entre 1 % et 100 % d'avancement) pour une valeur k donnée et calculé l'aire sous la courbe obtenue : une aire est importante indique un meilleur résultat. Sur la figure 2, on observe une amélioration de l'évaluation en augmentant k jusqu'à un plateau ($k > 10$). La meilleure valeur est obtenue pour $k = 16$, nous utilisons cette valeur dans la suite.

Nous avons comparé les trois propositions pour estimer la valeur finale d'une campagne (moyenne, *shift* et *coeff*). On voit que l'utilisation de la moyenne produit (comme prévu) des résultats de moindre qualité (selon le R^2) et que l'utilisation d'un coefficient multiplicateur est plus efficace tout au long de la campagne (fig. 3). Concernant l'impact de la mesure de distance, nous observons qu'elle produisent toutes des estimations de même qualité, les trois courbes étant presque confondues (fig. 4). Ces résultats nous poussent à choisir la méthode *coeff* pour l'estimation (meilleure précision) et de n'utiliser que le dernier état connu pour la distance (temps de calcul réduit).

Nous avons comparé notre approche avec la méthode naïve présentée dans la section 2. On constate sur la figure 5 que l'approche naïve produit de mauvais résultats en

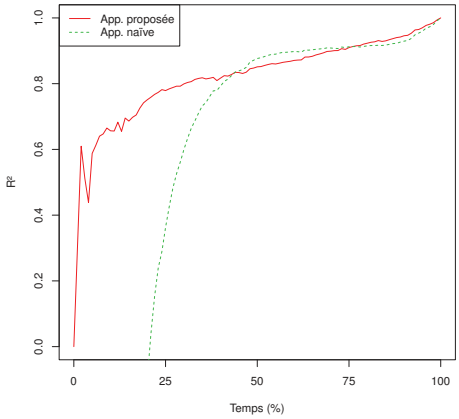

FIG. 5 – *Comparaison avec l'approche naïve*

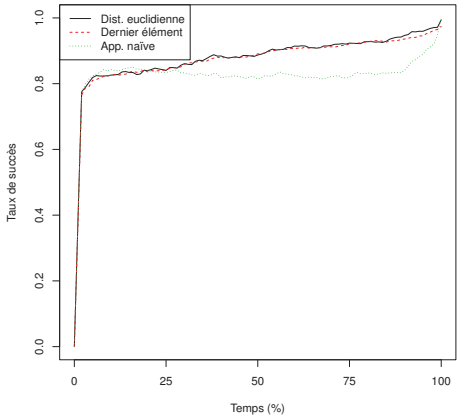

FIG. 6 – *Prédiction du succès d'une campagne*

début de campagne. Notre méthode surpasse l'approche naïve sur la première moitié de la campagne, puis les deux approches deviennent comparables. Nous montrons ainsi que l'approche proposée est efficace en début de campagne, l'objectif étant d'avoir une bonne estimation le plus tôt possible (pour pouvoir anticiper et réagir).

Enfin, nous avons comparé l'approche de classification proposée dans Etter et al. (2013) en utilisant deux mesures de distance (distance euclidienne et distance sur le dernier état connu) et l'approche naïve (pour la classification). On observe que k-NN a rapidement un taux de succès élevé (fig. 6), conformément à Etter et al. (2013), mais aussi que la distance calculée sur le dernier état produit des résultats semblables. Ici également l'utilisation des séries complètes ne semble pas nécessaire. L'approche naïve produit des résultats équivalents au début avant d'être dépassée par k-NN (contrairement au cas de la régression) : un financement initial élevé implique généralement un succès, ce qui permet de prédire aisément la réussite ou l'échec (bons résultats en classification), mais le montant final sera surestimé, puisque que les sommes levées ne se maintiennent pas longtemps à un niveau aussi élevé (mauvais résultats en régression).

5 Conclusion

Dans cet article, nous avons proposé une méthode de prédiction du montant levé lors d'une campagne de financement participatif. Notre approche consiste à rechercher les k campagnes les plus similaires à la campagne en cours et d'utiliser les montants levés par celles-ci pour faire une prédiction (estimation à partir de la moyenne et d'un coefficient multiplicateur). Pour comparer les séries temporelles entre elles, il suffit de comparer les valeurs sur le dernier état connu de la campagne en cours. Les résultats expérimentaux obtenus sont prometteurs et montrent le bien-fondé de la méthode.

Il reste de nombreuses améliorations à apporter. Il est nécessaire de travailler sur le passage à l'échelle, l'approche nécessitant de calculer une distance avec toute les

campagnes de la base de données. Nous pouvons envisager l'utilisation de structures de données performantes comme les arbres k-d ou bien de réduire l'espace de recherche en segmentant les campagnes. Nous pouvons aussi chercher à améliorer les performances (prédictions plus précises, plus tôt) en utilisant des données enrichies (prise en compte de la catégorie du projet, de l'impact des réseaux sociaux sur la campagne, etc.)

Références

Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician 46*(3), 175–185.

Cover, T. et P. Hart (2006). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*(1), 21–27.

Etter, V., M. Grossglauser, et P. Thiran (2013). Launch hard or go home ! Predicting the success of Kickstarter campaigns. In *Proceedings of the first ACM Conference on Online Social Networks*, COSN '13.

Frank, R., N. Davey, et S. Hunt (2001). Time series prediction and neural networks. *Journal of Intelligent and Robotic Systems 31*(1), 91–103.

Li, Y. (2016). Project success prediction in crowdfunding environments. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pp. 247–256. ACM.

Prasad, S. et P. Prasad (2014). Deep recurrent neural networks for time series prediction. *CoRR abs/1407.5949*.

Sakoe, H. et S. Chiba (1971). A dynamic programming approach to continuous speech recognition. In *Proceedings of the Seventh International Congress on Acoustics*, Volume 3, pp. 65–69.

Taylor, J. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science 54*(2), 253–265.

Summary

Crowdfunding is a methodology of funding a project from a large number of people, in opposition with traditional practices. With the Internet and social networking, this type of funding rapidly gained popularity. However more than 60% of projects are not funded, thus it is necessary to prepare carefully the crowdfunding campaign. Moreover, during the campaign, it is critical to be able to estimate the success as soon as possible in order to react adequately (reorganization, communication): prediction tools are then essential. In this article, we propose a prediction method for the final amounts raised during a crowdfunding campaign using the k-NN algorithm: with history of past campaigns, we determine the most similar ones to an ongoing campaign. Then we use the final amounts raised to build an estimation. We compared several distance measure to determine the nearest neighbors. Our experimental results indicate that the last state of a campaign is a good enough information to get an accurate prediction.