Propositions pour améliorer une méthode de prédiction du succès d'une campagne de financement participatif

Alexandre Blansché, Xavier Mazur

Propositions pour améliorer une méthode de prédiction du succès d'une campagne de financement participatif

Alexandre Blansché et Xavier Mazur

LORIA (UMR 7503), Campus Scientifique B.P. 239
54506 Vandœuvre-lès-Nancy Cedex, France
alexandre.blansche@univ-lorraine.fr

Résumé. Le financement participatif est un mode de financement d'un projet faisant appel à un grand nombre de personnes qui 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 plusieurs pistes d'amélioration pour la prédiction du montant levé lors d'une campagne de financement participatif en utilisant l'algorithme k-NN. La première proposition consiste à utiliser un algorithme de *clustering* afin de segmenter l'ensemble d'apprentissage et faciliter le passage à l'échelle. La seconde proposition consiste à extraire des caractéristiques pertinentes depuis les séries temporelles et les informations sur les campagnes pour avoir une représentation vectorielle.

1 Introduction

En quelques années seulement, le financement participatif a connu une forte croissance. Ce phénomène émergent est encore peu étudié, mais soulève de nombreuses questions dans divers champs scientifiques. L'une des principales interrogations est de connaître à l'avance le succès d'une campagne de financement participatif par des techniques de prévisions.

Nous proposons dans cet article de reprendre la méthode de prédiction du montant final levé lors d'une campagne de financement participatif utilisant l'algorithme k-NN proposée dans Blansché et al. (2017), et d'y apporter deux améliorations afin de faciliter le passage à l'échelle sur des données de plus grande taille et d'accroître les performances des prédictions. La première amélioration proposée consiste à découper l'ensemble d'apprentissage en groupes homogènes à l'aide d'un algorithme de classification non supervisée. La recherche des k plus proches voisins pourra alors se limiter à un faible nombre de *clusters* plutôt qu'à l'ensemble d'apprentissage complet.

La seconde proposition consiste à chercher des attributs, pour décrire les campagnes de financement participatif, permettant de réaliser des prédictions plus précises.

Dans la section 2, nous présentons plus en détail la problématique qui nous intéresse. Dans la section 3, nous présentons notre première proposition, sur l'utilisation de la classification non supervisée et dans la section 4 décrivons la seconde, sur l'utilisation d'attributs extraits des séries temporelles. Enfin, dans la section 5, nous apportons une conclusion à notre travail et proposons plusieurs perspectives de recherches futures.

2 Problématique

2.1 Contexte

Le financement participatif (*crowdfunding*), qui consiste à faire appel à un grand nombre de personnes pour financer un projet (contrairement aux modes de financements traditionnels), a connu une forte croissance avec l'émergence d'Internet et des réseaux socionumériques. Il existe plusieurs formes de financement participatif, selon qu'il y ait une récompense ou non pour les contributeurs, que celle-ci prenne la forme de dividendes en cas de succès ou d'un produit livré, mais également selon la durée des campagnes de financement, limitée ou non. Kickstarter [1] est un site de financement participatif parmi les plus populaires. La durée des campagnes est limitée dans le temps, mais peut varier d'un projet à l'autre. Le créateur d'un projet choisit un seuil de financement en-dessous duquel le projet est considéré comme un échec, mais au-dessus duquel c'est un succès : plus de 60 % des projets ne sont pas financés (le montant levé n'atteint pas le seuil fixé par le créateur). Le seuil de financement est parfois inférieur au seuil de rentabilité du projet : le créateur peut vouloir afficher 100 % de financement rapidement pour augmenter sa notoriété (les contributeurs privilégient souvent les campagnes réussies) sans que cela ne 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 et le créateur ne reçoit aucun fond : celui-ci peut alors s'assurer un remboursement partiel en choisissant un seuil de financement inférieur au seuil de rentabilité.

On peut décomposer l'évolution d'une campagne en $n+1$ états répartis dans le temps. On notera $t_i(c)$ l'horodatage du i-ième état de la campagne c, en normalisant de sorte que $0 \leq t_i(c) \leq 1$. Ainsi $t_0(c) = 0$ représente le début de la campagne et $t_n(c) = 1$ représente la fin de la campagne. On notera alors $m_i(c)$ le montant levé à l'état i d'une campagne c.

Beaucoup de publications (Etter et al., 2013; Mitra et Gilbert, 2014; Li, 2016) s'intéressent à la prédiction du succès ou de l'échec d'une campagne de financement participatif (classification supervisée). Dans cet article, nous nous intéressons à un problème de régression : l'objectif est de prédire la valeur finale $m_n(c_0)$ d'une campagne c_0 en cours, en ne connaissant que les valeurs $m_0(c_0)$ à $m_i(c_0)$, avec $i < n$, mais en disposant néanmoins d'un historique de campagnes passées. Ce problème est très peu traité jusqu'à maintenant, mais pour le créateur d'un projet, il est pourtant avantageux d'avoir une estimation précise $\hat{m}_n(c_0)$ de $m_n(c_0)$ le plus tôt possible dans la campagne

1. `https://www.kickstarter.com/`

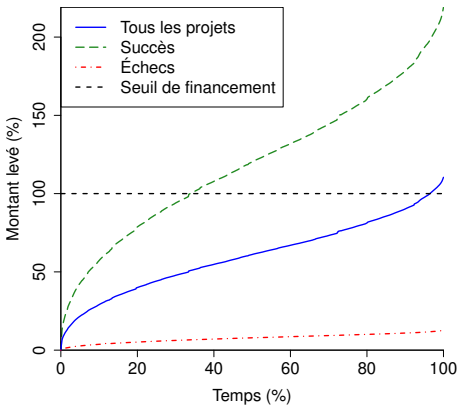

FIG. 1 – *Montant levé selon l'avancement dans la campagne*

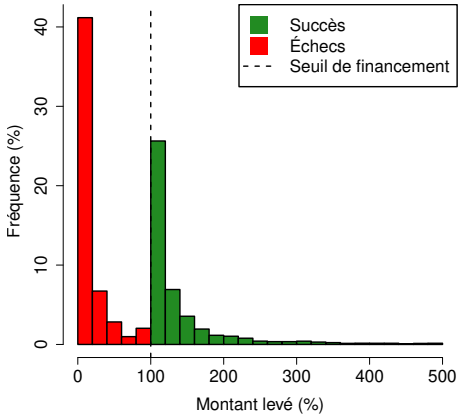

FIG. 2 – *Répartition du montant levé final*

(avec un i le plus petit possible), afin de pouvoir réagir rapidement (restructuration, communication) si les résultats sont inférieurs aux attentes.

2.2 Données utilisées

Dans Etter et al. (2013), les auteurs ont constitué une base de données portant sur 16042 campagnes Kickstarter datant de 2012 et 2013. Les séries temporelles ont été normalisées par un ré-échantillonnage en 1000 états. Il s'agit d'un petit échantillon : Kickstarter compte un total de plus de 375000 campagnes lancées depuis sa fondation en 2009 (58074 campagnes pour l'année 2016).

Sur la figure 1, on observe l'évolution du montant levé moyen au cours du temps. On remarque, en particulier pour les campagnes réussies, que la pente de la courbe est plus forte au début et à la fin de la campagne qu'au milieu de celle-ci. Sur la figure 2, on observe la répartition des montants levés à la fin d'une campagne. On remarque que la plupart des échecs sont très loin du seuil de financement et que la plupart des projets financés avec succès sont à peine au-dessus de ce seuil, bien que certains projets peuvent le dépasser très largement (plus de 500 % du montant demandé).

2.3 État de l'art

2.3.1 Méthodes « classiques »

Si les sommes levées étaient uniformément réparties durant la campagne, une approche naïve consisterait à faire une approximation linéaire. On définit alors $\hat{m}_n(c_0) = t_n(c_0) \times \frac{m_i(c_0)}{t_i(c_0)}$. Cette hypothèse n'est cependant pas vérifiée. Sur la figure 3 (évolution des contributions selon le temps) on remarque que les apports en début et fin d'une campagne sont plus importants qu'au milieu (surtout en cas de succès). On peut également tenter de prédire la valeur finale en construisant un modèle classique

de régression linéaire ou polynomial. Les méthodes d'autorégression (Taylor, 2008) consistent à expliquer une variable numérique par ses valeurs précédentes plutôt que par d'autres variables. Ces méthodes s'appliquent à beaucoup de problèmes d'analyse de séries temporelles telles que la prédiction boursière ou météorologique. La méthode ARIMA (AutoRegressive Integrated Moving Average) est l'une des approches d'autorégression les plus utilisées.

2.3.2 Utilisation de k-NN

La première approche proposée dans Etter et al. (2013) est basée sur la méthode k-NN (Cover et Hart, 2006) : on considère un ensemble de s séries temporelles, chacune correspondant à l'évolution du montant levé lors d'une campagne de financement participatif. Les campagnes, ayant des durées variables, sont normalisées par un réechantillonnage en un nombre 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 (sur $i+1$ états : entre 0 et i) sont déterminées et le succès d'une campagne est estimé selon le succès de ses « voisines » (vote à la majorité). Néanmoins l'approche ne prédit pas le montant levé lors d'une campagne mais prédit le succès ou l'échec de celle-ci.

Dans Blansché et al. (2017) nous avons proposé d'étendre cette approche pour estimer le montant levé final, en conservant donc la même méthodologie, mais en faisant de la régression par k-NN. Nous déterminons, parmi les s séries temporelles de l'historique, les k plus proches voisins (c_1 à c_k) d'une campagne c_0 sur les $i+1$ premiers états des campagnes. Notre objectif est d'utiliser les états finaux $m_n(c_1)$ à $m_n(c_k)$ pour obtenir une estimation $\hat{m}_n(c_0)$ de $m_n(c_0)$, l'état final de la campagne c_0. Pour cela, on définit $\mu_i = \frac{1}{k}\sum_{l=1}^{k} m_i(c_l)$ la moyenne des montants levés par les k campagnes voisines à l'état i et $\alpha_i = \frac{m_i(c_0)}{\mu_i}$, le rapport entre le montant levé à l'état i par la campagne c_0 et la moyenne des voisins au même état. Enfin, on nous définissons notre estimation $\hat{m}_n(c_0) = \frac{1}{k}\sum_{l=1}^{k} m_n(c_l) \times \alpha_i$. Dans Etter et al. (2013), pour déterminer les k voisins les plus proches, les auteurs comparent les campagnes en utilisant la distance euclidienne sur les états 0 à i. Cependant, nous avons montré que nous obtenons des prédictions de qualité équivalente en calculant la distance selon l'état i uniquement.

2.3.3 Expérimentations

Nous avons évalué l'efficacité à prédire le montant final des campagnes de financement participatif des différentes méthodes mentionnées sur les données de Etter et al. (2013). Pour comparer ces méthodes, nous avons utilisé la mesure RMSE (*Root Mean Square Error*), très souvent utilisée en régression (Hyndman et Koehler, 2006). Plus la valeur est faible, meilleure est la prédiction.

Dans cette expérimentation (et dans les suivantes), nous réalisons des prédictions pour chaque campagne après avoir produit un modèle sur un ensemble d'apprentissage. Nous avons ordonné les campagnes de financement participatif par date de fin. Pour la prédiction du montant final d'une campagne donnée, l'algorithme k-NN utilise un ensemble d'apprentissage constitué des 1000 dernières campagnes complètes. Ainsi nous

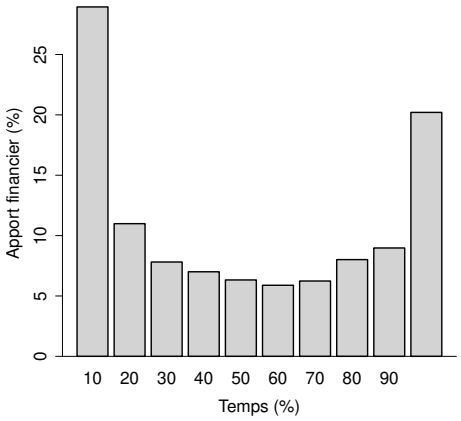

FIG. 3 – *Moyenne de l'apport financier à chaque état (campagnes normalisées en 10 états)*

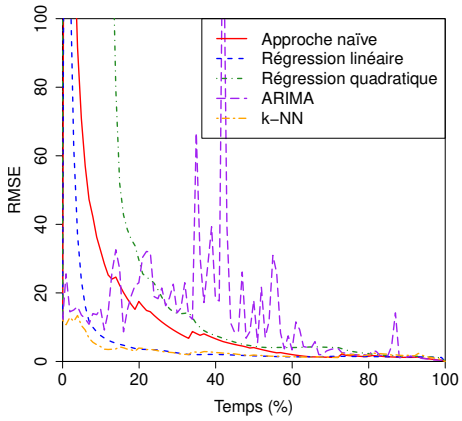

FIG. 4 – *Évolution dans le temps de la performance des méthodes selon le RMSE*

avons réalisé les prédictions sur 12670 campagnes. Le paramètre k de k-NN a été déterminée par validation croisée.

Nous avons calculé les prédictions des différents modèles de régression à partir de différents états de la campagne en cours afin de pouvoir observer l'évolution des performances dans le temps (figure 4). On remarque que les plus mauvaises performances sont obtenues par la méthode d'autorégression ARIMA. L'approche naïve et la régression polynomiale ont également des résultats peu satisfaisants. Étonnamment, les performances de la régression linéaire sont correctes, mais c'est l'approche proposée dans Blansché et al. (2017) qui produit les meilleures prédictions. Ce résultat semblent indiquer que la fouille de données, en exploitant un historique de campagnes passées, est une voie à privilégier pour prédire le montant final d'une campagne.

3 Classification non supervisée

3.1 Méthodologie

L'approche proposée est prometteuse, mais nous ne l'avons appliquée que sur une base de donnée de 16042 campagnes. Or il ne s'agit que d'un petit échantillon des campagnes lancées sur Kickstarter. Depuis sa création, plus de 375000 projets ont été lancés sur la plateforme et il peut y avoir simultanément plus de 4000 projets en cours de financement. De plus, pour chaque campagne, il est nécessaire de répéter la méthode de prédiction pour chaque nouvel état enregistré, afin d'avoir en permanence des prédictions à jour selon les données que l'on possède. La méthode proposée utilise le principe de l'algorithme k-NN et donc, pour chaque prédiction, il est nécessaire de calculer la distance par rapport à toutes les entrées de la base de données.

Nous proposons de réduire le nombre de calculs de distance en utilisant un algorithme de classification non supervisée, en s'inspirant d'une approche proposée dans Wang (2011). L'ensemble d'apprentissage est découpé en groupes homogènes (*clusters*) par un algorithme de classification non supervisée. Chaque *cluster* peut alors être résumé par un individu représentatif, qui minimise la somme des carrés des distances avec les objets qui composent le *cluster*. Pour identifier les k plus proches voisins d'une nouvelle campagne, on détermine d'abord les k' *clusters* les plus proches. On restreint alors l'ensemble d'apprentissage, et donc la recherche des k plus proches voisins, à l'union des observations contenues dans ces k' *clusters*.

3.2 Nombre de points dans les séries temporelles

Dans les données de Etter et al. (2013), chaque série temporelle comporte 1000 états uniformément répartis. On peut légitimement se demander si ce nombre d'états est nécessaire ou s'il est possible d'avoir une représentation fidèle avec moins d'états.

Pour réduire le nombre d'états d'une série temporelle, il existe plusieurs techniques. La plus simple est de faire un ré-échantillonnage avec un nombre d'états réduits, pris uniformément dans temps. D'autres approches, comme le Piecewise Aggregate Approximation (Keogh et al., 2001), consistent à découper la série temporelle en segments et de calculer une moyenne pour chaque segment. Certaines méthodes cherchent des points d'intérêt de la série temporelle tels que les pics les plus saillants (Chung et al., 2001), mais cette approche ne convient pas aux séries temporelles de financement participatif, car ces séries sont souvent monotones croissantes. Les transformées de Fourier (Bloomfield, 2004) sont très utilisées pour traiter des séries temporelles bruitées comme les signaux sonores, mais ont peu d'intérêt pour notre problème.

Nous nous sommes finalement intéressés à l'approche la plus simple et avons étudié l'impact de la réduction du nombre d'état par ré-échantillonnage.

3.3 Mesure de dissimilarité

Dans un deuxième temps, nous nous sommes intéressés au choix de la mesure de dissimilarité. De nombreuses mesures ont été développées pour traiter ce type de données (Giusti et Batista, 2013; Pereira et R.F. de Mello, 2013; Aghabozorgi et al., 2015). Les plus populaires restent la distance euclidienne et la dissimilarité *Dynamic Time Warping* (DTW), développée spécifiquement pour les séries temporelles (Sakoe et Chiba, 1971). La distance euclidienne nécessite d'avoir des séries de longueur identique. La mesure DTW est une mesure « élastique » qui permet d'associer des points particuliers des séries malgré un décalage temporelle. Cette mesure est cependant très coûteuse en temps de calcul.

3.4 Expérimentations

Pour étudier l'impact de la longueur des séries temporelles, nous avons appliqué l'algorithme k-means sur les séries complètes (1000 états) et les séries réduites par un ré-échantillonnage uniforme en 100 états et 10 états, ainsi que des séries représentées uniquement par l'état final. Nous avons utilisé la distance euclidienne et la mesure

	1000 états	100 états	10 états	1 état
1000 états	1	1	0,95	0,79
100 états		1	0,95	0,79
10 états			1	0,78
1 état				1

TAB. 1 – *Comparaison des* clusters *obtenus par la distance euclidienne selon la longueur des séries (κ de Cohen)*

DTW pour construire les *clusters*. Le nombre de *clusters* a été déterminé selon l'indice de Calinski-Harabasz (Caliński et Harabasz, 1974). Pour chaque configuration, nous avons appliqué l'algorithme k-means 100 fois, en six *clusters*, et conservé le meilleur résultat en terme d'inertie intra-classe. Les *clusters* ont été comparés entre eux en utilisant l'indice κ de Cohen. En raison du coût de calcul très élevé de la mesure DTW, nous nous sommes limités à un échantillon de 1000 séries choisies aléatoirement.

Nous avons choisi d'utiliser l'algorithme k-means qui est simple à mettre en œuvre. Le calcul du centroïde, l'individu représentatif au centre d'un *cluster*, va dépendre de la mesure de dissimilarité utilisée. Si la distance euclidienne est utilisée, le barycentre une simple moyenne sur chaque état de la série. Si la mesure DTW est utilisée, le calcul du centroïde n'est pas trivial : l'algorithme *DTW Barycenter Averaging* (DBA) est alors la méthode la plus couramment employée (Petitjean et al., 2011). Pour une utilisation à grande échelle, l'algorithme k-means ne sera pas une solution envisageable. En effet, il sera nécessaire de mettre à jour les *clusters* régulièrement (chaque fois qu'une campagne se termine) et l'utilisation de k-means serait alors trop coûteuse en temps de calcul et nous perdrions le bénéfice de l'approche. Ainsi, l'algorithme k-means devra être remplacé par une approche incrémentale (Ning et al., 2010) qui sera en mesure de mettre à jour les *clusters* à faible coût lorsque de nouvelles campagnes viendront s'ajouter à l'ensemble d'apprentissage.

Les tableaux 1, 2 et 3 montrent les résultats obtenus. On remarque qu'en utilisant la distance euclidienne, les *clusters* obtenus sont très semblables quelque soit la longueur des séries temporelles (1000, 100 ou 10 états). La différence s'accentue un peu quand on n'utilise qu'un seul état. En utilisant la mesure DTW, on observe que le nombre d'états a un impact beaucoup plus fort sur les *clusters*. On remarque également que moins il y a d'états dans les séries temporelles, moins il y a de différences entre la distance euclidienne et la mesure DTW.

Nous avons enfin étudié l'impact du découpage de l'ensemble d'apprentissage en *clusters* sur la méthode de prédiction du montant levé en utilisant le même protocole expérimental que celui décrit dans la section 2.3.3. Nous avons utilisé deux mesures de dissimilarité pour construire les *clusters* : la distance euclidienne et la mesure DTW. Nous avons également fait varier la taille de l'ensemble d'apprentissage de 1000 à 15000 et mesuré le temps de calcul de différentes configurations :

— prédiction par k-NN, sans découpage en *clusters* ;
— prédiction par k-NN avec un découpage en six *clusters* et recherche des voisins dans le *cluster* le plus proche (noté 6/1) ;

	1000 états	100 états	10 états	1 état
1000 états	1	0,69	0,52	0,58
100 états		1	0,63	0,62
10 états			1	0,8
1 état				1

TAB. 2 – *Comparaison des* clusters *obtenus par la dissimilarité DTW selon la longueur des séries (κ de Cohen)*

1000 états	100 états	10 états	1 état
0,53	0,69	0,79	1

TAB. 3 – *Comparaison des* clusters *obtenus par la distance euclidienne et la dissimilarité DTW selon la longueur des séries (κ de Cohen)*

— prédiction par k-NN avec un découpage en vingt *clusters* et recherche des voisins dans les cinq *clusters* les plus proches (noté 20/5).

Sur la figure 5, nous pouvons voir l'évolution du temps de calcul selon la taille de l'ensemble d'apprentissage pour les trois configurations étudiées (ce temps de calcul ne prend pas en compte le temps de construction des *clusters*). On remarque que, comme nous pouvions nous y attendre, le temps de calcul évolue linéairement selon la taille de l'ensemble d'apprentissage. De plus, on remarque que le découpage en *clusters* améliore nettement le temps de calcul, de l'ordre de 50 % environ avec les configurations choisies.

Sur la figure 6, nous pouvons voir la performance de la prédiction des méthodes comparées, selon l'avancement dans la campagne de financement, avec un ensemble d'apprentissage composées de 1000 campagnes. Nous observons que les performances de notre approche avec et sans *clustering* sont similaires si l'on utilise la distance euclidienne pour créer les *clusters*. On observe même une amélioration en début de campagne (avant 10 % d'avancement). En revanche, en utilisant DTW, les performances sont très instables (une étude plus approfondie des *clusters* obtenus pourra nous permettre de comprendre l'impact négatif de DTW). On remarque également que les résultats sont meilleurs dans les configurations 20/6 que dans les configurations 6/1.

4 Sélection de variables

4.1 Méthodologie

Un des résultats importants de l'article Blansché et al. (2017) est que la recherche des k plus proches voisins à l'état i ne nécessite pas de calculer une mesure de distance sur l'ensemble de la série temporelle. Les prédictions obtenues en utilisant uniquement le dernier état connu de la série sont tout aussi performantes. Ce résultat remet en question la nécessité de représenter une campagne par une série temporelle de l'évolution du montant levé. Nous avons donc tenté une représentation vectorielle, par extraction

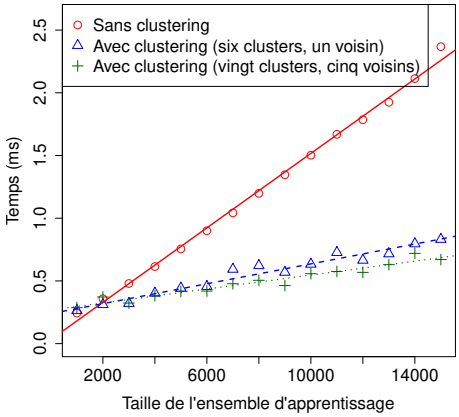

FIG. 5 – *Temps de calcul de la prédiction selon la taille de l'ensemble d'apprentissage*

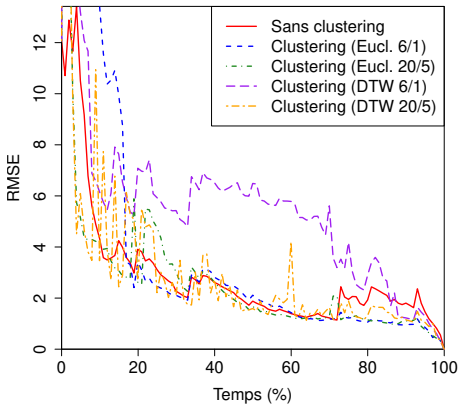

FIG. 6 – *Performance de la prédiction selon le RMSE*

de divers attributs depuis les données fournies par Etter et al. (2013). À noter que des données plus riches permettraient d'extraire un plus grand nombre d'attributs.

Nous avons construit un ensemble de huit attributs :

— $m_i\,(c)$: montant levé à l'état i ;
— $p_i\,(c)$: pente de la courbe à l'état i (différence avec l'état $i-1$) ;
— $a_i\,(c)$: aire sous la courbe d'évolution du montant levé jusqu'à l'état i ;
— $im_i\,(c)$: impulsion de la série (montant à 1 % d'avancement de la campagne) ;
— $nc_i\,(c)$: nombre de contributeurs à l'état i ;
— $d_i\,(c)$: durée de la campagne ;
— $f_i\,(c)$: seuil de financement en dollars américains ;
— $cm_i\,(c)$: contribution moyenne à l'état i ($\frac{m_i(c)}{nc_i(c)}$).

Les quatre premiers attributs sont calculés uniquement à partir des données sur le montant levé. Les quatre derniers attributs utilisent d'autres données, comme par exemple l'horodatage du début et de la fin de la campagne ou le nombre de contributeurs. Trois de ces attributs n'évoluent pas durant la campagne : la durée de celle-ci, le seuil de financement du projet et l'impulsion. Nous avons cherché quels sont le ou les attributs les plus pertinents pour prédire le montant levé final d'une campagne. Une recherche exhaustive, parmi les 255 combinaisons possibles, a été réalisée.

4.2 Expérimentations

Nous avons appliqué le même protocole expérimental que celui décrit dans la section 2.3.3. Pour chaque combinaison d'attributs, nous avons appliqué notre approche en utilisant la distance euclidienne sur l'espace vectoriel correspondant, centré et réduit, pour chercher les k plus proches voisins. Nous avons ensuite comparé les combinaisons entre elles selon la moyenne de performance sur chaque état. Nous pouvons difficilement

Attribut	Nombre d'occurrences
Aire sous la courbe	2
Pente	6
Nombre de contributeurs	7
Montant levé	11
Contribution moyenne	12
Durée	14
Seuil de financement	21
Impulsion	23

TAB. 4 – *Occurrences des attributs dans les combinaisons les plus performantes*

exposer dans le présent document l'ensemble des résultats obtenus, mais simplement une synthèse des faits les plus remarquables.

Les évaluations sont réparties entre 2,09 et 119,57. Les quartiles sont 2,73 (25 %), 2,85 (50 %) et 5,97 (75 %). La combinaison d'attributs qui a produit la meilleure performance (2,09) est composée d'un unique attribut : l'impulsion. Ce résultat est particulièrement étonnant car il s'agit d'un attribut qui est fixe tout au long de la campagne. Ainsi, ce seront toujours les mêmes k voisins qui serviront de références pour estimer le montant final levé. Ce résultat conforte l'idée que la réussite d'une campagne de financement participatif réside principalement dans son lancement, mais remet en question l'intérêt de traiter les données comme des séries temporelles.

Nous avons également observé les 10 % des combinaisons les plus efficaces (26 combinaisons) et quels attributs composent ces combinaisons : nous avons compté le nombre d'occurrences (cf. tableau 4). On observe que les trois attributs fixes dans le temps sont les trois attributs les plus utilisés parmi les meilleures combinaisons de variables et que l'attribut le plus utilisé est l'impulsion, ce qui confirme son importance.

5 Conclusion

Dans cet article, nous avons présenté deux propositions d'amélioration d'un algorithme de prédiction utilisant les k plus proches voisins. Nous avons évalué l'impact du découpage de l'ensemble d'apprentissage en groupes homogènes pour réduire le temps de calcul et l'utilisation d'attributs à la place de séries temporelles.

La plupart des articles de recherche sur la prédiction du résultat d'une campagne de financement participatif, qu'il s'agisse de prédire de son succès ou son échec ou bien le montant levé final, ont tendance à utiliser des séries temporelles (Li, 2016; Zhao et al., 2017; Fan-Osuala et al., 2018). Néanmoins, certains résultats que nous avons obtenus nous font penser que cette approche n'est peut-être pas la bonne. Il y a plusieurs éléments qui mènent à cette conclusion :

- quand on réduit la longueur des séries temporelles de 1000 états à 100, voire à 10 états, les *clusters* obtenus sont presque identiques ;

- l'utilisation de la mesure DTW détériore les performances de l'algorithme de prédiction (par rapport à la distance euclidienne) ;

- l'attribut qui maximise les performances de prédiction ne dépend pas du temps.

Cette conclusion inattendue nous pousse à voir le problème sous un autre angle et à reconsidérer notre méthodologie afin de produire des prédictions plus efficaces.

Notre principale perspective est d'étudier les possibilités d'amélioration des performances (prédictions plus précises, plus tôt) en améliorant notre approche (en explorant les méthodes de raisonnement à partir de cas) ou en développant d'autres algorithmes, mais également en utilisant des données enrichies, internes au site de financement participatif (nombre de contributeurs, de commentaires, etc.) ou portant sur la réputation d'un projet, en particulier sur les réseaux socionumériques. Nous comptons également continuer à travailler sur la classification non supervisée de campagnes de financement, en particulier en utilisant un méthode incrémentale, moins coûteuse en temps de calcul.

Références

Aghabozorgi, S., A. S. Shirkhorshidi, et T. Y. Wah (2015). Time-series clustering - a decade review. *Information Systems 53*(C), 16–38.

Blansché, A., D. Da Conceicao, et D. Koby (2017). Prédiction du montant levé lors d'une campagne de financement participatif par la méthode des plus proches voisins. In *17ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2017, 24-27 Janvier 2017, Grenoble, France*, pp. 387–392.

Bloomfield, P. (2004). *Fourier Analysis of Time Series : An Introduction*. Wiley Series in Probability and Statistics. Wiley.

Caliński, T. et J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation 3*(1), 1–27.

Chung, F., T.-C. Fu, R. Luk, et V. Ng (2001). Flexible time series pattern matching based on perceptually important points. In *Workshop on Learning from Temporal and Spatial Data in International Joint Conference on Artificial Intelligence*.

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, pp. 177–182.

Fan-Osuala, O., D. Zantedeschi, et W. Jank (2018). Using past contribution patterns to forecast fundraising outcomes in crowdfunding. *International Journal of Forecasting 34*(1), 30–44.

Giusti, R. et G. Batista (2013). An empirical comparison of dissimilarity measures for time series classification. In *Brazilian Conference on Intelligent Systems, BRACIS 2013, Fortaleza, CE, Brazil, 19-24 October, 2013*, pp. 82–88.

Hyndman, R. et A. Koehler (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting 22*(4), 679–688.

Keogh, E., K. Chakrabarti, M. Pazzani, et S. Mehrotra (2001). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems 3*(3), 263–286.

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.

Mitra, T. et E. Gilbert (2014). The language that gets people to give : Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pp. 49–61. ACM.

Ning, H., W. Xu, Y. Chi, Y. Gong, et T.S Huang (2010). Incremental spectral clustering by efficiently updating the eigen-system. *Pattern Recognition 43*(1), 113–127.

Pereira, C. et R.F. de Mello (2013). Common dissimilarity measures are inappropriate for time series clustering. *RITA 20*(1), 25–48.

Petitjean, F., A. Ketterlin, et P. Gançarski (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition 44*(3), 678–693.

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.

Wang, X. (2011). A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality. In *IJCNN*, pp. 1293–1299. IEEE.

Zhao, H., H. Zhang, Y. Ge, Q. Liu, E. Chen, H. Li, et L. Wu (2017). Tracking the dynamics in crowdfunding. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 625–634. ACM.

Summary

Crowdfunding is a methodology of funding a project from a large number of people. 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 estimate the success as soon as possible in order to react adequately (reorganization, communication): prediction tools are then essential. In this article, we propose several methods to improve the prediction of the amount raised during a crowdfunding campaign using the k-NN algorithm. The first proposition consists in using a clustering algorithm in order to segment the learning set and to facilitate the scaling for big data sets. The second proposition consists in extracting relevant features from the time series and information on the campaigns, in order to have a vector representation.