2018 IEEE International Conference on Data Mining

Tracking and Forecasting Dynamics in Crowdfunding: A Basis-Synthesis Approach

Xiaoying Ren, Linli Xu*, Tianxiang Zhao, Chen Zhu[1], Junliang Guo, Enhong Chen
Anhui Province Key Laboratory of Big Data Analysis and Application, Baidu Talent Intelligence Center[1]
School of Computer Science and Technology, University of Science and technology of China
wjyren@mail.ustc.edu.cn, *linlixu@ustc.edu.cn, fykd2@mail.ustc.edu.cn
zhuchen02@baidu.com, guojunll@mail.ustc.edu.cn,cheneh@ustc.edu.cn

***Abstract*—Crowdfunding is an emerging online fundraising mechanism for creators to launch campaigns (projects) to solicit funds or expand their influence. Tracking the dynamics, i.e., daily funding amounts can be of great help to campaign creators as well as contributors. Previous works on this subject either fit the fluctuations of time-series with predefined stochastic process or apply a regularization term to constrain learned tendencies, resulting in limited generalization abilities. Patterns of funding-amount sequences in crowdfunding are often exclusive and non-linear, making previous predictors suboptimal. To tackle this problem, we propose a novel method based on synthesized bases which can be composed into arbitrary patterns. Concretely, we build a large set of candidate basis from which we select based on reliability, diversity and latent structures. We use representations of sequences in this basis space as a predictor, and adopt a dual-graph to exploit neighbouring information to enhance its prediction quality. Experimental results demonstrate the effectiveness of our method.**

***Keywords*—crowdfunding; time-series forecasting**

I. INTRODUCTION

As an emerging Internet-based fundraising mechanism, crowdfunding provides a revolutionary way of raising monetary contributions and attracting publicity through collecting many small amounts of money from the crowds. Generally it has three sorts of actors: the campaign creators who propose the campaigns (projects) to be funded, campaign contributors who give the campaign financial support (contributions), and crowdfunding platforms. While maintaining a strong growth momentum in fundraising, crowdfunding also motivates a surge of interest in the context of technology [4] and medical treatment [8], etc.

Since most of current crowdfunding platforms only enable creators to solicit funds in a certain time duration, many works focus on predicting the success or failure of campaigns [10]. However, almost all of them do not tackle the problem of tracking the dynamics, i.e., daily funding amounts in crowdfunding, which is beneficial to all the three actors in the sense that it can provide a guidance for creators to adjust their campaign settings in time, for backers to make a comprehensive decision, and for platforms to make better recommendations to potential customers.

The problem of tracking the dynamics for specific campaigns is first studied in [17], which employs the exponential function to model the rise and the Hazard function to model the decay of the funding dynamics, and uses 'clustering' techniques by learning several sets of parameters, one for each cluster to improve the prediction accuracy. However, it could be inadequate to fit a single pattern to approximate daily funding amounts for different campaign tasks, considering the following three reasons: (1) Different campaigns can exhibit quite different patterns, making it hard to use one simple synthetic influence/decay function to depict them; (2) Other than the overall patterns, it also has noisy fluctuations caused by incidents or just irregular outliers, which further complicates the problem; (3) Contribution records of many campaigns are sparse. Trading amounts can be zero for many days, making extracted patterns unstable and unreliable, with poor generalization ability.

Despite these difficulties, the ability to discover the intrinsic patterns is crucial to accurately predict the future of a funding-amount sequence. Considering the challenges in directly modeling various funding-amount trends, we propose to handle it from a reverse view, through synthesizing a set of basic styles, which are easy to predict and can be composed together into arbitrary patterns. However, implementing the above idea in funding amount forecasting is not trivial with three problems to solve: (1) choosing a proper non-linear projection which can extract arbitrary patterns from precedent funding amounts, (2) choosing proper basic styles that are both expressive and are easy to predict, (3) refining the forecasting results as it is possible that tendencies are not perfectly captured.

In this paper, we adopt the Fourier transformation as the projection function considering its effectiveness and simplicity. To select a proper set of basic types that can be expressive as well as discriminative to capture the trends of different patterns, we build a large candidate set and select from it. After the basic types are chosen, we decompose the dynamics of crowdfunding campaigns into the selected bases, and make predictions with them. We also employ a tightness dual graph to measure the similarities among different campaigns and leverage the information from neighbors to refine the forecasting results. The main contributions of this paper are as follows:

- We propose a novel idea to cast the challenging problem of dynamic tracking and pattern discovery as basis synthesis.
- As a concrete realization of this idea, we show how to select good bases for pattern modeling. We propose an

978-1-5386-9159-5/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDM.2018.00161

effective formulation taking into account the reliability and the diversity, as well as the latent structures of the selected bases simultaneously.
- Interestingly, the design of decomposition for each campaign can be regarded as feature representation and used to measure the similarity among campaign tasks. And we use that to further improve the prediction accuracy. Experiments validate the effectiveness of this method.

II. RELATED WORK

Crowdfunding Analysis The research on crowdfunding is still in its infancy. Most of the previous works can be grouped into three categories: predicting the funding results, i.e., whether a campaign will succeed or not [10], identifying the influential factors [3] and designing product supplies [11], etc. Recently, some researchers begin to pay attention to the dynamics in crowdfunding. For example, [9] studies the fundraising dynamics of campaigns in their complete funding durations from statistical and empirical perspectives. The authors of [17] first study the problem of tracking the dynamics for specific campaigns and further develop a regression method accommodating to predefined rise and decay functions, and attempt to enhance the prediction performance through learning a set of parameters for each cluster of campaigns and fitting the model on two levels before fusing them together. However, patterns of funding amounts are often complex and non-linear, straying a lot from predefined patterns, which limits their effectiveness on real-world data. Besides, to the best of our knowledge, the literature on tracking the dynamics in crowdfunding platforms from the data-driven perspective is relatively sparse.

Time Series Forecast Time series forecast is a classical machine learning problem and has been explored extensively [5, 20]. In this field, stochastic processes, as a powerful mathematics tool, has been applied in many works and achieves good performances. For example, Gaussian process regression (GPR) models are applied in [7] to predict the cost of a complete trajectory. [12] designs a new scheme where the combined effects of the past events on future events can be superadditive, subadditive, or even subtractive. These methods presume an underlying pattern revealed in time series, and learn model parameters from the given distribution. Consequently, they can only extract a certain tendency variation, which may fail to generalize to realistic data or do not work well when the given distribution is not aligned with the real distribution.

Another important group of time-series forecasting methods focus on incorporating a specific regularizer [18, 19] to capture temporal variations. For example, considering that the significant temporal transitions in traffic usually appear in a few specific time periods, [6] adopts the $l_{1,\infty}$ norm to depict the spurs in travel time prediction problem. [14] assumes that the future predictor is generated from a set of ensemble predictors, and employs a graph Laplacian regularizer to ensure consistency of the predicted time series. However, these methods have certain limitations in capturing latent patterns of time series, and also lack the capacity to generalize to aperiodic forecasting problems. The clear distinction between the proposed method and previous techniques lies in that the existing methods either fit the fluctuations of time-series with predefined stochastic processes or apply a regularizaton term to constrain the learned tendencies, while we propose a base-synthesized way to capture the latent tendencies.

III. PROPOSED METHOD

A. Preliminary

For the task of funding amount forecasting, N campaigns $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ are collected where $\boldsymbol{x}_i \in \mathbb{R}^M$ represents the sequence of funding amounts for campaign i within M funding days. The problem can then be then formally defined as predicting the funding amounts in the following days given the historical daily funding amounts $\{\boldsymbol{x}_i^1, ..., \boldsymbol{x}_i^M\}$.

To effectively model the dynamics of crowdfunding, several challenges need to be tackled. Specifically, it is necessary to model various dynamic patterns of crowdfunding tasks, to capture the overall dynamic trends while filtering out the irregular noisy fluctuations, and to cope with issue of data sparsity. To handle the above challenges, we creatively propose a BAsis- Synthesis Approach (BASA) to discover the dynamic patterns of funding-amount sequences, along with their amplitude weights dominances. Concretely, to select a proper set of bases, we leverage the reliability and diversity as well as the structural information of its members. When predicting, a funding-amount sequence is first projected into the basis space to extract it as a feature extraction process, forecasting is then achieved by performing predictions on each basis, the results are next recalibrated by leveraging the neighboring information.

B. Synthesizing Bases

There are significant difficulties in extracting typical patterns from historical sequences without prior knowledge or pattern-type assumptions. Here we tackle the problem in a novel way by encoding the shape of arbitrary patterns in the basis space. Inspired by the Fourier transformation [1], a frequency representation that can recover any original time-series sequences from basic trigonometric functions, we introduce a projection operator, $\Phi(\cdot)$: $\mathbb{R}^M \to \mathbb{R}^K$ to decompose the historical sequences into a set of K discrete candidate frequencies $\{w_k = \frac{2\pi K}{k} | k = 1, 2, ...K\}$ as

$$\boldsymbol{A}_i \triangleq \Phi(x_i) = \sum_{t=1}^{T} \boldsymbol{x}_i(t) \odot \begin{bmatrix} \cos(\omega_1 t) \\ \sin(\omega_1 t) \\ ... \\ \cos(\omega_{K/2} t) \\ \sin(\omega_{K/2} t) \end{bmatrix} \tag{1}$$

where $A_i \in \mathbb{R}^K$ is the $i\text{-}th$ column of matix A, corresponding to the representation of the $i\text{-}th$ time-series of the crowdfunding task in the bases space, and $\odot$ is the Kronecker product.

Note that we choose to use the Fourier basis because of its simplicity and effectiveness, but our method is not confined to it, any basis with sufficient expression abilities can be used.

After building the candidate basis set, to capture the overall dynamic trends and filter out noisy fluctuations, we propose a method to select bases guided by its structural information.

As Eq.(1) indicates, time-series of campaign tasks can be decomposed into a set of K candidates, which can recover their tendency inversely as $\Phi_k^{-1}(\Phi_k(x_i))$. Different campaign tasks exhibit various patterns while each pattern has its own decompositions. So, we propose to select those bases that can help us differentiate between various patterns and preserve the main tendency of each pattern while removing noise. First, we evaluate each basis according to its ability in modeling crowdfunding dynamics and define a ranking matrix $\boldsymbol{R}$ to represent the confidence score of campaign tasks for each basis. The i-th column of $\boldsymbol{R}$ is learned by solving the following equation:

$$\begin{aligned} \min_{\boldsymbol{R}_{\cdot,i}} \quad & ||\boldsymbol{x}_i - \sum_k R_{k,i}\Phi_k^{-1}(\Phi_k(\boldsymbol{x}_i))||_F^2 \\ \text{s.t.} \quad & \boldsymbol{R}_{\cdot,i} \succeq 0, \boldsymbol{R}_{\cdot,i}^T \mathbf{1}_K = 1, \end{aligned} \tag{2}$$

in which $\boldsymbol{R}_{k,i}$ is a scoring value, indicating the inclination of x_i to choose basis k. $\mathbf{1}_K$ is a vector with K elements, and Φ_k refers to the k-th dimension of Φ. According to the constraint $\sum_k R_{k,i} = 1$, Eq.(2) can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{R}_{\cdot,i}} \quad & \boldsymbol{R}_{\cdot,i}^T \boldsymbol{G}^i \boldsymbol{R}_{\cdot,i} \\ \text{s.t.} \quad & \boldsymbol{R}_{\cdot,i} \succeq 0, \boldsymbol{R}_{\cdot,i}^T \mathbf{1}_K = 1, \end{aligned} \tag{3}$$

where $\boldsymbol{G}^i = [G_{mn}^i] \in \mathbb{R}^{K\times K}$ is the local gram matrix for $\boldsymbol{x}_i$ with elements $G_{mn}^i = (\boldsymbol{x}_i - \Phi_m^{-1}(\Phi_m(\boldsymbol{x}_i))^T(\boldsymbol{x}_i - \Phi_n^{-1}(\Phi_n(\boldsymbol{x}_i))$. Eq.(3) is a standard quadratic programming (QP) problem whose optimal solution can be obtained by any off-the-shelf QP solver.

Further, to make sure the chosen bases are diverse enough, we construct a regularization term $\boldsymbol{R}_{\cdot,i}^T \boldsymbol{W} \boldsymbol{R}_{\cdot,i}$, where $\boldsymbol{W} \in \mathbb{R}^{K\times K}$ is the similarity matrix, in which each element $w_{j,h}$ encodes the correspondence between two bases, and the diagonal elements are set to zero. Intuitively, when the similarity $w_{j,h}$ is large, assigning large values to $r_{j,h}$ and $r_{h,i}$ simultaneously implies a large loss. As a consequence, this regularization term encourages diverse bases to be selected. The formulation can then be augmented as

$$\begin{aligned} \min_{\boldsymbol{R}_{\cdot,i}} \quad & \boldsymbol{R}_{\cdot,i}^T \boldsymbol{G}^i \boldsymbol{R}_{\cdot,i} + \alpha \boldsymbol{R}_{\cdot,i}^T \boldsymbol{W} \boldsymbol{R}_{\cdot,i} \\ \text{s.t.} \quad & \boldsymbol{R}_{\cdot,i}^T \mathbf{1}_K = 1, \boldsymbol{R}_{\cdot,i} \succeq 0. \end{aligned} \tag{4}$$

To get a better measurement of basis correspondence, we leverage our non-linear transformation operator $\Phi(\cdot)$, which can be regarded as a feature extractor. Given that the matrix $\boldsymbol{A}$ is the feature representation for all tasks with $A_{j,i} = \Phi_j(x_i)$, the similarity between two bases can be measured as $W_{j,h} = \sum_i A_{j,i} \cdot A_{h,i}$. And the formulation becomes

$$\begin{aligned} \min_{\boldsymbol{R}_{\cdot,i}} \quad & \boldsymbol{R}_{\cdot,i}^T \boldsymbol{G}^i \boldsymbol{R}_{\cdot,i} + \alpha \boldsymbol{R}_{\cdot,i}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{R}_{\cdot,i} \\ \text{s.t.} \quad & \boldsymbol{R}_{\cdot,i}^T \mathbf{1}_K = 1, \boldsymbol{R}_{\cdot,i} \succeq 0. \end{aligned} \tag{5}$$

In the meantime, considering that funding-amount sequences can be very sparse, which may lead to an unstable ranking score matrix $\boldsymbol{R}$. On the other hand, time series of similar campaign tasks exhibit similar tendencies which inspire us to boost the performance from related tasks. We further incorporate the campain correlations into consideration and refine $\boldsymbol{R}$ by leveraging the shared structures among campaign tasks. Intuitively, campaign tasks with similar tendencies should also have aligned ranking scores. To encourage ranking scores from campaigns with similar patterns to be clustered together, we integrate a trace norm [15, 16] to constrain the structure of $\boldsymbol{R}$. We assume that $\boldsymbol{R}$ can be decomposed into groups of low-rank bases, which implies a block structure

$$\begin{aligned} \min_{\boldsymbol{R}} \quad & \sum_{i=1}^N \boldsymbol{R}_{\cdot,i}^T \boldsymbol{G}^i \boldsymbol{R}_{\cdot,i} + \alpha \sum_{i=1}^N \boldsymbol{R}_{\cdot,i}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{R}_{\cdot,i} + \eta ||\boldsymbol{R}||_* \\ \text{s.t.} \quad & \boldsymbol{R}_{\cdot,i}^T \mathbf{1}_K = 1, \boldsymbol{R}_{\cdot,i} \succeq 0. \end{aligned} \tag{6}$$

The intuition above should also work for matrix $\boldsymbol{A}$, as campaign tasks with similar tendency variations should also have aligned features. Therefore, we utilize the supervision in $\boldsymbol{A}$ to guide $\boldsymbol{R}$ as $\boldsymbol{R} = \sum_z \beta_z \boldsymbol{A}_z$, where $\boldsymbol{A}_z = \boldsymbol{u}_z \boldsymbol{v}_z^T$ and $\boldsymbol{u}_z$, $\boldsymbol{v}_z$ are the z-th left and right singular vector of $\boldsymbol{A}$ respectively, with $\boldsymbol{A} = \sum_z \lambda_z \boldsymbol{u}_z \boldsymbol{v}_z^T$. To ensure that $\boldsymbol{R} \succeq 0$ and constrain the complexity of $\boldsymbol{R}$, we require $\beta_z^k \geq 0$ and introduce the l_2 norm constraint on $\boldsymbol{\beta}$. The final optimization problem for the basis selection procedure can be expressed as

$$\begin{aligned} \min_{\boldsymbol{R},\boldsymbol{\beta}} & \sum_{i=1}^N \boldsymbol{R}_{\cdot,i}^T \boldsymbol{G}^i \boldsymbol{R}_{\cdot,i} + \alpha \sum_{i=1}^N \boldsymbol{R}_{\cdot,i}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{R}_{\cdot,i} + \eta ||\boldsymbol{R} - \sum_z \beta_z \boldsymbol{A}_z||_F^2 \\ \text{s.t.} \ & \boldsymbol{R}_{\cdot,i}^T \mathbf{1}_K = 1, \boldsymbol{R}_{\cdot,i} \succeq \mathbf{0}, \boldsymbol{\beta} \succeq 0, ||\boldsymbol{\beta}||_F^2 \leq C. \end{aligned} \tag{7}$$

where α and η are the two parameters to control the strength of the two regularization terms.

C. Funding Prediction

After obtaining the ranking matrix $\boldsymbol{R}$, we choose bases according to it. When dealing with a new campaign, we first get its representation on selected bases, then we proceed to make the prediction in the bases space before performing an inverse-transformation predictor Φ^{-1} to get the actual prediction of the following k-th day:

$$x_i^{'}(t+k) = (\Phi^{-1}(\Phi(\boldsymbol{x}_i)))(t+k). \tag{8}$$

Despite the effectiveness of this method, there are still extreme cases when trading amounts are too sparse or contain too much noise that the projection fail to fully discover its intrinsic patterns. Therefore, we further exploit the neighboring information to refine the predictions. As Φ can also be treated as a step of feature extraction, it can be used to measure the correlations among campaigns. If we denote the similarity matrices $\boldsymbol{W}^X$ and $\boldsymbol{W}^A$ as the cosine similarity among campaigns in the original space and the feature space respectively, we can integrate them by taking the average of the two matrices $\boldsymbol{W}^C = \frac{1}{2}\boldsymbol{W}^X + \frac{1}{2}\boldsymbol{W}^A$ for simplicity. We localize $\boldsymbol{W}^C$ and normalize it for each campaign. Then, we fuse the neighboring information into forecasting with a trade-off parameter μ

$$\boldsymbol{x}_i(t+1) = (1-\mu)(\Phi^{-1}(\Phi(\boldsymbol{x}_i)))(t+1)+\mu(\boldsymbol{W}^C_{i,:}\boldsymbol{X}_{-i}(t+1)). \tag{9}$$

IV. Optimization

Apparently, the objective function in Eq.(7) is not convex regarding the variables $\boldsymbol{R}$ and $\boldsymbol{\beta}$ simultaneously. Here we derive an iterative optimization algorithm. In each iteration, only one variable is updated while the others remain unchanged.

A. Computing $\boldsymbol{R}$

First, we fix $\boldsymbol{\beta}$ and update $\boldsymbol{R}$. when $\boldsymbol{\beta}$ is fixed, Eq.(7) is only related to $\boldsymbol{R}$. Abbreviating $\boldsymbol{R}_{\cdot,i}$ as $\boldsymbol{r}_i$, we need to solve the following problem

$$\begin{aligned}&\min_{\boldsymbol{R}}\sum_{i=1}^{N}\boldsymbol{r}_i^T\boldsymbol{G}^i\boldsymbol{r}_i+\sum_{i=1}^{N}\alpha\cdot\boldsymbol{r}_i^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{r}_i+\eta||\boldsymbol{R}-\sum_z\beta_z\boldsymbol{A}_z||_F^2\\&\text{s.t.}\boldsymbol{r}_i^T\boldsymbol{1}_K=1,\boldsymbol{r}_i\succeq 0.\end{aligned} \tag{10}$$

Leaving the third term $\eta||\boldsymbol{R}-\sum_z\beta_z\boldsymbol{A}_z||_F^2$ for later consideration, Eq.(10) can be decomposed into N subproblems. We employ the augment Lagrange method (ALM) [2] to rewrite the $i\text{-}th$ subproblem as

$$\begin{aligned}\min_{\boldsymbol{r}_i}\quad&\boldsymbol{r}_i^T(\boldsymbol{G}^i+\alpha\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{r}_i+\boldsymbol{\nu}_{i_1}^T(\boldsymbol{r_i}-\boldsymbol{s_i})+\frac{\tau}{2}||\boldsymbol{r_i}-\boldsymbol{s_i}||_F^2\\&\nu_{i_2}(\boldsymbol{r}_i^T\boldsymbol{1}_K-1)+\frac{\tau}{2}\upsilon_{i_2}(\boldsymbol{r_i}\boldsymbol{1}_K^T-1)^2\\\text{s.t.}\quad&\boldsymbol{s}_i\succeq 0.\end{aligned} \tag{11}$$

where $\boldsymbol{s}_i$ is an auxiliary vector, τ is a scalar, $\boldsymbol{\nu}_{i_1}$ and ν_{i_2} are the Lagrange multipliers in terms of the $i\text{-}th$ subproblem.

Update $\boldsymbol{r}_i$. Notice that when fixing $\boldsymbol{s}_i$, Eq.(11) can be formulated as an unconstrained QP problem, i.e., $\min_{\boldsymbol{r}_i}\boldsymbol{r}_i^T\boldsymbol{H}\boldsymbol{r}_i+\boldsymbol{r}_i^T\boldsymbol{b}$, where $\boldsymbol{H}=\boldsymbol{G}^i+\alpha\boldsymbol{A}^T\boldsymbol{A}+\frac{\tau}{2}I_K+\frac{\tau}{2}\boldsymbol{1}_K\boldsymbol{1}_K^T$ and $\boldsymbol{b}=-\tau\boldsymbol{s}_i-\tau\boldsymbol{1}_K+\boldsymbol{\nu}_{i_1}+\nu_{i_2}\boldsymbol{1}_K$. The solution of this QP problem can be represented as a linear equation: $2\boldsymbol{r}_i^T\boldsymbol{H}=\boldsymbol{b}$. Note that $\boldsymbol{H}$ is a positive definite matrix, so we employ the algorithm proposed in [13] to give a nearly linear convergence solution. Concretely, the gradient of $\boldsymbol{R}$ related to Eq.(10) is the combination of N subproblems plus the gradient of the third term.

Update $\boldsymbol{s}_i$. Denote $\boldsymbol{s}_i=\{s_i^1,...,s_i^K\}$, fixing the other variables, the objective function w.r.t. each $\boldsymbol{s_i}$ is

$$\min_{\boldsymbol{s}_i\succeq 0}\frac{\tau}{2}(\boldsymbol{r}_i-\boldsymbol{s}_i)^2+\boldsymbol{\nu}_{\boldsymbol{i_1}}^T(\boldsymbol{r}_i-\boldsymbol{s}_i), \tag{12}$$

and the solution to this problem is

$$s_i^k=\max(0,r_i^k+\frac{\nu_{i_1}^k}{\tau}). \tag{13}$$

B. Computing $\boldsymbol{\beta}$

Next, we fix $\boldsymbol{R}$ and update $\boldsymbol{\beta}$. When we fix $\boldsymbol{R}$ and ignore the constant terms in Eq.(7), we have

$$\min_{\boldsymbol{\beta}}\eta||\boldsymbol{R}-\sum_z\beta_z\boldsymbol{A}_z||_F^2\quad\text{s.t.}\ \ \boldsymbol{\beta}\succeq 0,||\boldsymbol{\beta}||_F^2=C. \tag{14}$$

The above QP function can also be solved by the ALM algorithm or other existing convex optimization packages.

V. Experiments

A. Dataset Description and Evaluation criteria

Indegogo.com dataset We retrieve the funding-amount data from indiegogo.com, which enables creators to solicit funds in a pledged time duration. This dataset collects 14,143 launched campaigns from July 2011 to May 2016, soliciting over 18 billion funds from 217,156 investors. Here, we remove unfinished campaigns. The dataset is partitioned into two subsets based on the declared funding days. For 30-day campaigns, we record funding amounts of campaigns in their duration (30 days) and the following 15 days. Similarly, for 60-day campaigns, we keep observations in their time duration (60 days) and the following 30 days. Dataset features are also retrieved for baselines methods. Due to the space limitation, we only present the results with training ratio at 60% and 80%, which are named as $D\#1$ and $D\#2$ respectively. The reported results are averaged over five-round tests.

Considering the sparsity of most campaign tasks, we adopt the same criteria described in [14] to measure the root Mean Squared Error (rMSE) between the predicted funding amounts and the ground truth. The smaller the value of rMSE, the better the performance.

B. Implementation Details

For basis decomposition, we fix the time interval K as 500. For basis selection, the weighting parameters α and η are set to 0.1 and 1 respectively, and remain fixed through out all the experiments. We select bases with top-m ranking scores, where m indicates the number of selected bases. Parameters of the baselines are set as suggested in the corresponding papers. We implement ARMA models using Python statsmodels library[1] and tune the optimal AR degree p and MA degree q for each campaign task. For the GPR algorithm, the bandwidth of the Gaussian kernel is set to 0.2. For the RDMTR, ORION, SR and SWR models, the regularization parameters are set to the default values.

C. Results and Analysis

Comparison with Baseline Methods To validate the effectiveness of BASA, we compare it with the classical time-series methods, including ARMA and GPR models [7]. We also compare to baselines including RDMTR [6], ORION [14] and the state-of-the-art SR and SWR [17] models.

[1] http://www.statsmodels.org/stable/index.html

Model		rMSE_30 days					rMSE_60 days				
		1-day	2-day	3-day	4-day	5-day	1-day	2-day	3-day	4-day	5-day
D#1	ARMA	921.5	901.0	953.7	1046.2	976.2	827.4	856.8	901.6	878.4	838.1
	GPR	805.2	756.4	798.8	851.6	978.1	821.5	736.7	792.0	763.5	790.1
	RDMTR	392.6	425.1	413.7	373.1	392.7	459.1	468.7	374.3	471.6	534.2
	ORION	887.1	910.2	735.6	798.1	812.7	845.8	687.3	752.1	676.4	702.5
	SR	296.9	336.2	358.2	399.4	452.6	389.1	455.9	485.5	518.4	514.5
	SWR	272.0	295.8	304.0	298.3	294.3	281.6	359.9	318.5	402.2	457.6
	BASA	**80.8**	**115.3**	**121.7**	**193.8**	**108.3**	**95.9**	**164.6**	**185.5**	**190.6**	**348.8**
D#2	ARMA	807.9	912.3	954.2	1046.7	1109.2	745.2	832.4	956.9	1331.6	1559.1
	GPR	631.2	703.5	749.1	801.6	883.3	597.4	658.5	792.9	1010.7	1119.6
	RDMTR	345.8	371.4	382.7	431.1	450.3	422.2	357.4	321.0	349.9	427.5
	ORION	723.3	776.1	731.7	791.4	758.2	502.6	524.2	551.8	593.5	612.2
	SR	292.3	307.5	272.7	295.7	299.6	281.9	388.6	435.2	390.3	440.3
	SWR	231.7	213.6	251.2	235.6	264.7	224.3	263.3	245.1	281.6	287.1
	BASA	**121.2**	**97.3**	**123.9**	**189.1**	**153.3**	**82.4**	**220.1**	**163.2**	**136.6**	**260.1**

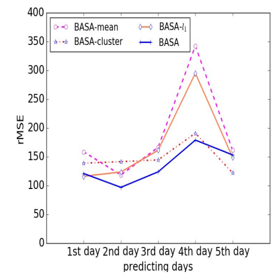
TABLE I: Performance of various algorithms in terms of rMSE on the indiedgogo dataset. Methods with the best performances (measured by paired t-tests at 95% significance level) are bolded.

Table 1 shows the forecasting performance in terms of the following 5 days. Without loss of generality, we only report the results on the 80%-20% data split (D#2). Throughout the experiments, we make the following observations. First, not surprisingly ARMA model achieves the worst performance since it is linear and leads to exponential decay. Second, considering the rise-and-fall patterns can enhance the representation quality in dynamic forecasting. As a consequence, RDMTR, SR, SWR achieve better performance than ORION which only considers the smoothness property of time sequences. Third, exploiting time tendency variations can make the pattern description more discriminative. Time series patterns are preserved in the basis space, and amplitute values of these bases are also encoded for each specific campaign task. This guarantees that our model can manage time series variations flexibly. In addition, BASA does not have a significant performance decay in the five-day case compared to predicting only one day, indicating that our model can cope with more complex tendency variations.

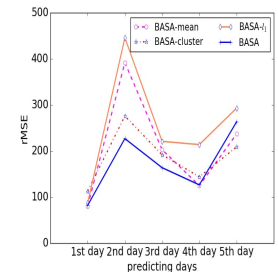
Evaluation of Basis Selection We proceed to investigate the influence of selection methods and number of bases. To demonstrate the effectiveness of our basis selection method, we first compare it with three different variants of BASA:

- BASA-mean: We choose bases evenly at the same interval. To be fair, the number of bases is fixed at 60.
- BASA-cluster: In this variant, we consider the cluster property of bases, i.e., similar frequency of bases can profile similar patterns. We employ K-means to obtain 10 clusters, and choose 6 bases from each cluster as representatives.
- BASA-l_1: In this variant, we incorporate the l_1 norm to control the diversity of chosen bases.

The performance of different model variants are presented in Fig. 1. Obviously, the performance of the other basis-selection methods are inferior and can drop significantly, while BASA and BASA-cluster remain stable. The reason is that BASA-mean and BASA-l_1 only consider recovering previous data, which makes them suboptimal. On the other hand, BASA and BASA-cluster exploit the structure information, which enhances the robustness, especially when the data is sparse.

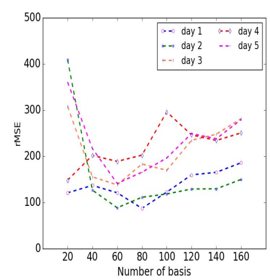


(a) $campaign\_30days$

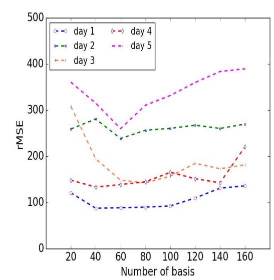


(b) $campaign\_60days$

Fig. 1: Evaluation of basis selection methods



(a) $campaign\_30days$



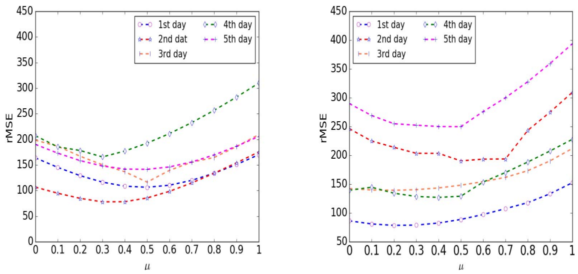
(b) $campaign\_60days$

Fig. 2: Evaluation on the number of chosen bases

In addition, BASA also explicitly takes diversity into consideration, and achieves better performance than BASA-cluster.

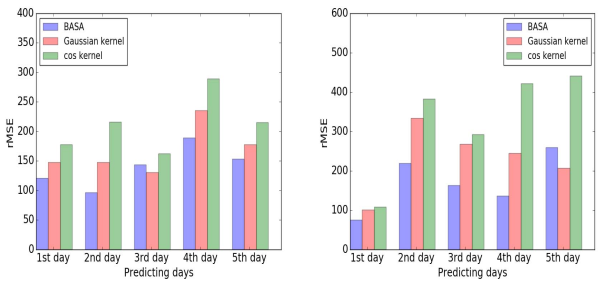
Next, we study the influence of the number of selected bases. Experimental results are presented in Fig. 2. We can observe that when the number of bases is small ($m < 60$), the chosen bases can not fully capture the variation patterns so that the forecasting performance is suboptimal. The best performance can be achieved around $m = 60$. When the number of bases increases, the redundancy of bases weakens the prediction performance as more noisy information is kept which impairs discriminating power.

Evaluation of Model Components To validate the contribution of each component in our approach, we first investigate how our method behaves when varying μ, which is a trade-off parameter between our basis-based predictor and neighbor-



(a) *campaign_30days* (b) *campaign_60days*

Fig. 3: Parameter sensitivity



(a) campaign_30days (b) campaign_60days

Fig. 4: Performance comparison over kernel methods

based predictor. The results w.r.t. different values of μ on the following five days prediction are shown in Fig. 3. To summarize, with μ increasing from 0 to 0.5, the corresponding loss drops dramatically, indicating the tightness graph predictor can introduce more supervised information, guiding the model to produce enhanced results in sparse scenarios. However a larger μ will make the model rely too much on neighbors, which impairs its prediction capability. As a tradeoff, the performance is reasonable when μ is around 0.5.

We further compare our kernels W^C with the Gaussian and Cosine kernels to evaluate the effectiveness of using our time series decomposition to measure the similarity between campaigns. Results are illustrated in Fig. 4. We can observe that our kernel achieves better performance in most scenarios. It also indicates that the learned basis space A can represent features of different patterns, which in turn validates its ability of capturing the trends in time-series tasks.

Conclusion

In this paper, we propose a novel method to predict the funding amounts for crowdfunding campaigns. As dynamics in crowdfunding campaigns are diverse, it is hard to directly catch their patterns, so we try to tackle this problem from an inverse view through bases synthesis. We select proper bases that are both expressive and discriminative, then decompose task sequence into those bases and make predictions on each separate basis before composing them together to get the forecasting results. Neighboring information is also exploited to refine the prediction results. Experimental results validate the ability of the chosen bases to catch the dynamic trends and its effectiveness as feature representations.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (61673364, 91746301, and U1605251), and the Fundamental Research Funds for the Central Universities (WK2150110008).

References

[1] David H. Bailey and Paul N. Swarztrauber. A fast method for the numerical evaluation of continuous fourier and laplace transforms. *Siam Journal on Scientific Computing*, 15(5):1105–1110, 1994.

[2] Minming Chen, Arvind Ganesh, Zhouchen Lin, Yi Ma, John Wright, and Leqin Wu. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *Coordinated Science Laboratory Report no. UILU-ENG-09-2214*, 2009.

[3] Jaegul Choo, Daniel Lee, Bistra Dilkina, Hongyuan Zha, and Haesun Park. To gather together for a better world: Understanding and leveraging communities in micro-lending recommendation. In *Proceedings of the 23rd international conference on World wide web*, pages 249–260. ACM, 2014.

[4] Alessandro Cordova, Johanna Dolci, and Gianfranco Gianfrate. The determinants of crowdfunding success: Evidence from technology projects. *Procedia-Social and Behavioral Sciences*, 181:115–124, 2015.

[5] Hao Hu and Guo-Jun Qi. State-frequency memory recurrent neural networks. In *International Conference on Machine Learning*, pages 1568–1577, 2017.

[6] Aiqing Huang, Linli Xu, Yitan Li, and Enhong Chen. Robust dynamic trajectory regression on road networks: A multi-task learning framework. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 857–862. IEEE, 2014.

[7] Tsuyoshi Idé and Sei Kato. Travel-time prediction using gaussian process regression: A trajectory-based approach. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 1185–1196. SIAM, 2009.

[8] Jennifer G Kim, Ha Kyung Kong, Karrie Karahalios, Wai-Tat Fu, and Hwajung Hong. The power of collective endorsements: credibility factors in medical crowdfunding campaigns. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4538–4549. ACM, 2016.

[9] Venkat Kuppuswamy and Barry L Bayus. Crowdfunding creative ideas: The dynamics of project backers in kickstarter. 2015.

[10] Yan Li, Vineeth Rakesh, and Chandan K Reddy. Project success prediction in crowdfunding environments. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 247–256. ACM, 2016.

[11] Qi Liu, Guifeng Wang, Hongke Zhao, Chuanren Liu, Tong Xu, and Enhong Chen. Enhancing campaign design in crowdfunding: a product supply optimization perspective. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 695–702. AAAI Press, 2017.

[12] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6738–6748, 2017.

[13] Daniel A Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2004.

[14] Jianpeng Xu, Pang-Ning Tan, Jiayu Zhou, and Lifeng Luo. Online multi-task learning framework for ensemble forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 29(6):1268–1280, 2017.

[15] Rong Yan, Jelena Tesic, and John R Smith. Model-shared subspace boosting for multi-label classification. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 834–843. ACM, 2007.

[16] Yongxin Yang and Timothy M. Hospedales. Trace norm regularised deep multi-task learning. 2016.

[17] Hongke Zhao, Hefu Zhang, Yong Ge, Qi Liu, Enhong Chen, Huayu Li, and Le Wu. Tracking the dynamics in crowdfunding. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–634. ACM, 2017.

[18] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang Tien Lu, and Naren Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1503–1512, 2015.

[19] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. A multi-task learning formulation for predicting disease progression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, Ca, Usa, August*, pages 814–822, 2011.

[20] Ali Ziat, Edouard Delasalles, Ludovic Denoyer, and Patrick Gallinari. Spatio-temporal neural networks for space-time series forecasting and relations discovery. 2018.