2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Implementation of Chaotic Analysis on Retweet Time Series

Yuanyuan Bao[1], Chengqi Yi[2], Jingchi Jiang[2], Yibo Xue[1], Yingfei Dong[3]
[1] Tsinghua National Lab for Information Science and Technology, Tsinghua University, Beijing, China
[2] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
[3] Department of Electrical Engineering, University of Hawaii, Honolulu, USA
(by51800@163.com, garnettyige@163.com, jiangjingchi0118@163.com, yiboxue@tsinghua.edu.cn, yingfei@hawaii.edu)

***Abstract*—Retweet has become one of the most prominent feature on social networks and an important mean for secondary content promotion. Most existing investigations of retweet behaviors on social networks are conducted based on empirical studies or information diffusion models (such as stochastic process or cascading model). To the best of our knowledge, such a retweet process has not been investigated as a chaotic process. In this paper, we have first examined that retweet time series by 0-1 test where the results provide identification of chaotic behaviors. Furthermore, taking into account of the proven chaotic characteristic, chaos LS-SVM prediction method is applied to form predictions using only a small fraction of the retweet time series. Our evaluation on Sina Weibo dataset and comparisons with a bayesian model and strawman modal show that this nonlinear prediction method can translate to good step ahead forecasts and perform high accuracy in retweet prediction.**

***Keywords*—Retweet, Chaotic Analysis, Nonlinear Prediction, LS-SVM, Social Network**

I. INTRODUCTION

Today's constant connectivity and mobile devices have dramatically changed our lives, and we have witnessed the explosive growth of social networks as an innovative communication medium. Social network allows users to freely spread real-time information through both the WWW and the mobile phone networks, where each piece of information may further be tagged for searching and grouping. Because of these features, social network has successfully distinguished itself from traditional media under a number of events by being more spontaneous, mobile, and disseminative. Main social network providers (such as Twitter, Facebook, and Sina Weibo) have heavily affected users' lives in many different ways recent years. Facebook has over 1.3 billion monthly active users who generate more than 47 billion pieces of contents per month. Twitter has over 0.3 billion monthly active users who generate 0.5 billion tweets every day. In China, the most famous social network site--Sina Weibo has more than 0.2 billion monthly active users who generate more micro blogs than Twitter.

A social network is a social structure consisting of many actors tied together based on common interests. There are three main actions on social networks: *post*, *retweet*, and *mention*, which contribute to information diffusion. The actions are all related to the contents with a limited size (up to 140 characters in case of Sina Weibo and Twitter) generated by users. Among these actions, retweet is the most powerful mechanism to diffuse information via social network. When a user finds a tweet worth sharing, he could copy the entire post and retweet it to his followers. The information could thus reach to the entire social network while the content remains relatively intact. The dynamic of retweet in social networks is of potential implications for understanding the spread of broader ideas, trends in social networks and also revenue models for both individuals who "sell tweets" and for those looking to monetize their reach. What's more, it also contributes to develop better control of illegal information like rumors. Hence, the retweet process needs correct measurement, analysis, and reliable estimates. The purposes of this paper are to investigate the question whether retweet exhibits chaotic behavior and obtain reliable prediction.

There have been some studies on the retweet prediction in micro-blog network. Many of them consider the prediction problem as a classification problem, which predicts whether a content will be retweeted or not and determines which features dominate the prediction accuracy. These methods all need large amounts of data collections, including the follower graph, content of tweet and user profiles and consume a lot of API calls. There are a few other works adopting linear prediction model to estimate retweet number directly on retweet time series. However, there is no systemic investigation on whether retweet process is chaotic or random. Based on our investigation, we have found that the complexity of such process is mainly affected by tweet poster, tweet topic, and post time. In other words, the state of such process is highly depended on initial conditions, which is a basic characteristic of chaos systems. And this uncertainty will make the linear prediction model be less accurate. Moreover, most of the work focus on English micro-blog network like Twitter, Youtube and Facebook and few studies are on Chinese micro-blog network.

In this paper, we firstly analyze the chaotic characteristic of retweet time series in real Sina Weibo dataset and point out it is not rational to directly conduct linear prediction on retweet time series. Then we conduct phase space reconstruction on retweet time series and implement least squares support vector machine (LS-SVM) on the new built phase space to predict the retweet number in every time interval only using information of the early retweet time series. We aim to make these predictions very early in the lifetime of the tweet, sometimes within minutes of it being posted.

The main contributions of this paper are:

ASONAM '15, August 25-28, 2015, Paris, France

 ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2808881

- We implement 0-1 test for chaos on retweet time series of real Sina Weibo dataset, and determine significant chaotic characteristic of retweet time series.
- Based on obtained time delay and embedding dimension of retweet time series, we construct phase space reconstruction and implement least squares support vector machine prediction model to predict the retweet number.
- We conduct experiments of Chaos LS-SVM prediction model on Sina Weibo dataset and make comparisons with bayesian model and strawman model, which all show that Chaos LS-SVM prediction model can help us effectively predict retweet number in every time interval.

The structure of this paper is as follows. We will introduce related work in Sec. II. In Sec. III, we provide a description of the data utilized. In Sec. IV, we implement 0-1 test for chaos on retweet time series. And we will introduce the Chaos LS-SVM prediction model in Sec. V and evaluate it with the real-world dataset in Sec. VI. We will conclude the paper in Sec. VII.

II. Related Work

In this section we review the related work on chaos theory and existing projects on information diffusion and retweet prediction on social networks.

A. Chaos Theory and Its Applications

Chaos is an irregular, seemingly random phenomenon, generated by nonlinear interactions in deterministic systems, also named as *deterministic randomness*. Chaotic behaviors can be observed in many natural systems, such as physics [1], hydrology [2], medicine [3], and other fields, where small differences in initial conditions yield widely diverging outcomes, making long-term prediction nearly impossible.

Since Lorenz discovered chaos, it has been widely used for nonlinear analysis in many dynamical systems. Shi [1] presented an improved method to detect chaotic motions for rotor-bearing systems, which introduces the correlation integral function method to simultaneously estimate the embedding dimension and the reconstruction delay. Albostan [2] investigated the chaotic characteristics of all the stations in a river basin. By using three nonlinear data analysis methods, the daily discharge data of four gauge stations is demonstrated chaotic characteristics. Jiang [3] employed power spectrum, Lyapunov exponent, and Kaplan-Yorke dimension to describe chaotic vibrations in the vocal-fold model. In summary, chaos theory studies the behavior of dynamical systems that are highly sensitive to initial conditions, and has made important contributions in nonlinear analysis.

B. Retweet Behavior and Retweet Prediction

Emerging social network tools have successfully distinguished themselves from traditional media, by being more spontaneous, mobile, and disseminative. Meanwhile, the massive amount of human interaction data on social networks makes understanding the process extremely difficult. Although retweeting is the most powerful mechanism for diffusing information via social network, the analysis and prediction of retweets are fairly new issues.

Many existing works have focused on empirical studies and related features of retweet behaviors. Galuba et al. [4] proposed a propagation model to predict which users are likely to mention which URLs. Steeg [5] studied the factors that prohibit the epidemic transmission of popular news posted on Digg. Boyd et al. [6] examine the practice of retweeting as a way by which participants can be "in a conversation" and highlight how authorship, attribution, and communicative fidelity are negotiated in diverse ways. Yang [7] proposed a factor graph model to predict users' retweeting behaviors.

Many of works consider the prediction problem as a two-classification problem, which predicts whether a content will be retweeted or not and determines which features dominate the prediction accuracy. Bakshy et al. [8] tried to predict the existence of a retweet between a particular pair of users. Suh et al. [9] used a generalized linear model to understand what features influenced the chance of a tweet being retweeted. Bandari et al. [10] and Hong et al. [11] used a variety of algorithms not to predict exact number of retweets, but rather a coarse interval for number of retweets of a tweet.

A few recent works propose mathematical models for understanding and predicting retweet on social networks. But these works accomplish the prediction either on topology of social graph, spatial information or content. Wang [12] described information propagation as a discrete process based on the Galton-Watson model which models the evolution of family names and successfully explains the interaction between the topology of social graphs and the intrinsic interests of messages. Wang [13] proposed a Partial Differential Equation (PDE) to model the temporal and spatial characteristics of information diffusion. Yang [14] proposed a Role-Aware Information diffusion model (RAIN) that integrates social role recognition and diffusion modeling into a unified framework. This model not only can be used to predict whether an individual user will repost a message, but also can be used to predict the scale and duration of a diffusion process. Yang [15] developed a linear influence model representing the diffusion by predicting which node will be influenced by other nodes in a network. These methods all need large amounts of data collections, including the follower graph, content of tweet and user profiles and consume a lot of API calls.

In contrast to these previous works, we are trying to predict the exact retweet number in every interval only using information of the early retweet time series. This is similar to Szabo and Huberman [16] who use a linear model to predict the popularity of stories on Digg.com and videos on YouTube by observing their popularity after one hour and one week, respectively. Another similar work is Tauhid Zaman [17] who develops a probabilistic model for the evolution of the retweets using a Bayesian approach, and form predictions using observations on the retweet times and the local network

of the retweeters. There are some differences between our model and these two models. Firstly, Szabo and Huberman [16] assume that retweet time series are linear and use a linear model to predict. However, there is no systemic investigation on whether retweet time series is chaotic or random. So using linear model directly to make prediction is not rational. Secondly, although our prediction goal is similar to [17], Tauhid Zaman makes prediction not only on early time series, but also on retweets of other tweets and follower graph. That is, their method needs much more information than ours.

III. DATA SET

In order to analyze the retweet behavior, we collect tweet posted from 03/2012 to 01/2015, whose topics are politics, technology, economy and entertainment. We choose 300 different tweets with highest retweet counts as our research objectives, which are all called root tweets. We collect the entire retweets of all these root tweets. The retweets of these 300 root tweets are amount to 3,494,198. The data includes retweet time, identity of the users who retweet, content of retweets and identify of each retweet (i.e. ID corresponds to one tweet.)

We analyzed the distribution of retweet count shown in Fig. 1. It shows the number of retweets of each root tweet. Only a few of root tweets have larger retweet counts, while most of root tweets have smaller retweet counts. For example, the total number of retweet counts of root tweet: 3635531139581716 is 146,838, while the total retweet counts of most of root tweets are only about 12,000. That is, in micro-blog network, retweet count complies with power-law distribution. Because of the huge difference of the retweet counts between root tweets, we should carefully examine the root tweet respectively.

Through the data collection, we obtain entire retweets of each root tweet. Assuming each root tweet be represented by ID_j, j=1, 2, …, 300 and has a time series P={t_1, t_2, …} composed by the posted time of each retweet. Then if we choose a time interval δ, then the retweet count in each time interval δ can form the retweet time series x={x_1, x_2, …x_N}, which is our research objective.

The purpose of this paper is to make early prediction of retweet count of each time interval. Therefore, in the analysis of each root tweet, we only use data of the retweets posted in the first 24 hours after root tweet being published. That is, the length of retweet time series is 24*60/δ. Take root tweet: 3617517265784502 as an example, making the time interval be five minutes and through the summation of retweet count in each time interval, we can obtain retweet time series of root tweet: 3617517265784502 shown in Fig. 2. Through Fig. 2, apparently we can figure out the retweet time series does not follow the linear principle and look like randomly fluctuating. Then the problem whether the retweet behavior is chaotic or random determines this behavior can be predicted or not. Following we will examine whether it is chaotic by 0-1 test.

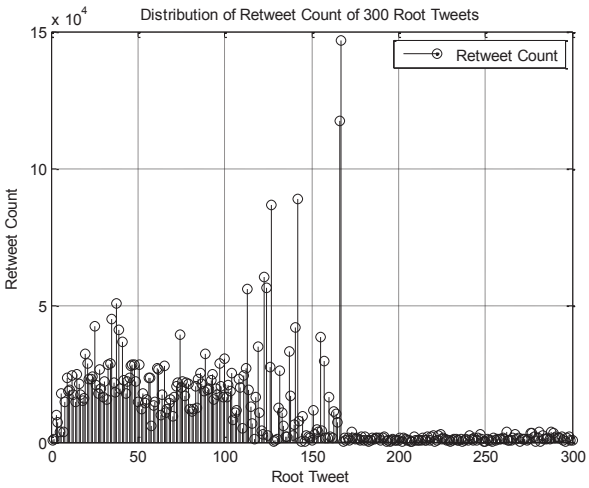

Fig. 1. Plot of distribution of retweet count for different root tweets

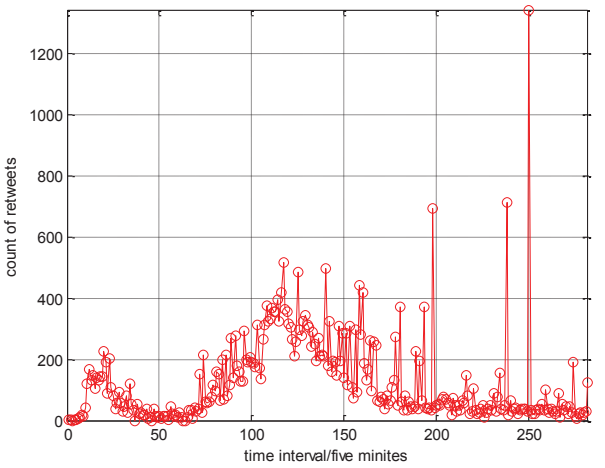

Fig. 2. The Retweet Time Series for Root Tweet: 3617517265784502. The time interva is five minutes. The plot is the total number of retweet count versus time interval for root tweet: 3617517265784502.

IV. CHAOTIC ANALYSIS OF RETWEET TIME SERIES

Chaotic analysis has successfully captured the key characteristics of epidemic spreading. Retweeting on social networks has a striking similarity to epidemic spreading. In this section, we will analyze the chaotic characteristics of retweet time series with a binary test, the 0-1 test which is designed for the analysis of deterministic dynamical system [18]. 0-1 test does not depend on phase space reconstruction but rather works directly with the time series given. Compared with maximum Lyapunov exponent method, we implement 0-1 test for the chaotic analysis.

Given a discrete retweeting time series x={x_n}, n=1,2, …, N, for $c \in (0, 2\pi)$, we compute the translation variables

$$p(n) = \sum_{i=1}^{n} x(i)\cos(ic), \quad n = 1, 2, ...N \tag{1}$$

$$q(n) = \sum_{i=1}^{n} x(i)\sin(ic), \quad n = 1, 2, ...N \tag{2}$$

Based on translation variables p(n) and q(n), we obtain the mean square displacement M(n)

$$M(n) = M_c(n) - (E(\varphi))^2 \frac{1-\cos nc}{1-\cos c} \tag{3}$$

where $M_c(n)$ and $E(\varphi)$ are defined as

$$M_c(n) = \frac{1}{N-n}\sum_{i=1}^{N-n}[(p(i+n)-p(i))^2 + (q(i+n)-q(i))^2] \tag{4}$$

$$M_c(nE(\varphi) = \lim_{L\to\infty}\frac{1}{L}\sum_{j=1}^{L}x(j) \tag{5}$$

According to [19], in order to yield good results, *n* takes the value *n=10/N*. The asymptotic growth rate K_c of mean square displacement can be calculated as

$$K_c = \lim_{n\to\infty} \lg M(n) / \lg n \tag{6}$$

If the dynamics is regular then the mean square displacement is a bounded function in time where Kc≈0, whereas if the dynamics is chaotic then the mean square displacement scales linearly with time where K_c≈1.

In order to explain how to accomplish the quantitative analysis of retweet process, we take a retweet process for example, whose ID is 3617517265784502. The retweets daily time series of tweet 3617517265784502 are shown in Fig. 2. In Fig. 2, Kc is shown as a function of c and K_c≈1 which represent chaotic characteristic of the retweeting time series.

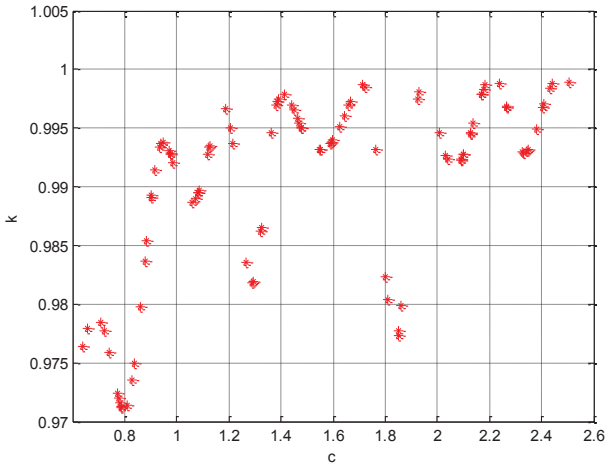

Fig. 3. Plot of Kc versus c for the logistic map calculated using 0-1 test. We use here N=288 data points, and 100 equally spaced values for c. Kc≈1 corresponds to chaotic dynamics.

In order to analyze whether retweet time series exhibits chaotic characteristic, we conduct 0-1 test on our dataset and obtain the asymptotic growth rate Kc of the 300 root tweets shown in Fig. 4. Through Fig. 4, we figure out that asymptotic growth rate which is larger than 0.9 accounts for 71.67 %. And the asymptotic growth rates of 90.33% of root tweets are larger than 0.7. Considering 0-1 test is a binary test, so we can conclude that asymptotic growth rate approximately equals to 1, which show significant chaotic feature of retweet time series. It can be seen that retweet time series possess chaotic features, providing a basis for performing short-term forecast of retweet count with the help of chaos theory.

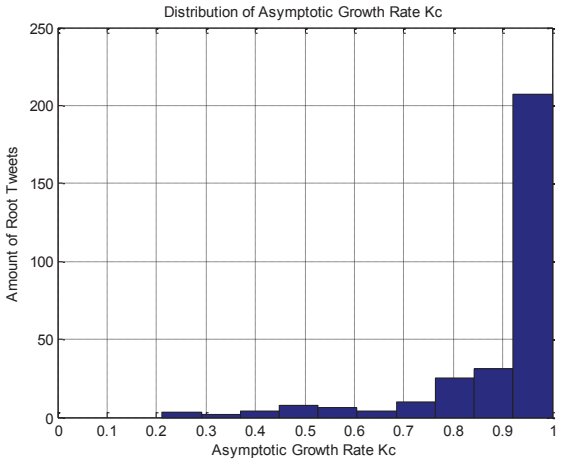

Fig. 4. Plot of distribution of asymptotic growth rate. The asymptotic growth rates of 90.33% of root tweets are larger than 0.7. The asymptotic growth rate which is larger than 0.9 accounts for 71.67%.

V. CHAOS-SVM PREDICTION MODEL

In this section, we firstly reconstruct phase space of retweet time series based on calculating time delay and embedding dimension. Based on the phase space, we conduct least square support vector machine prediction for retweet time series.

A. Phase Space Reconstruction

The *phase space reconstruction (PSR)* method can help us indirectly detect attractors in real-world dynamical systems using time series data on a single variable. Therefore, we can build more accurate information models of such systems.

To reconstruct phase space of retweet time series, we need to determine appropriate time delays and embedding dimensions. There are several methods to determine the time delays. Considering a nonlinear correlation relationship, we choose the mutual information method to find the time delay. For discrete random variables $\{x_n\}$ and $\{x_{n+\tau}\}$, the mutual information between them can be calculated using Eq. (7):

$$I(x_n, x_{n+\tau}) = \sum_{n=1}^{N} P(x_n, x_{n+\tau}) \log_2 \frac{P(x_n, x_{n+\tau})}{P(x_n)P(x_{n+\tau})} \tag{7}$$

Mutual information is a function of time delay τ, represented by I(τ). Time delay τ is the duration when mutual information I(τ) first reaches a local minimum value. Taking retweet time series of root tweet, whose ID is 3617517265784502 as an example shown in Fig. 5. When time delay equals to 4, the mutual information reaches the first local minimum value represented by a red circle. So the time delay is 4.

After determining the time delay, we further identify the embedding dimension. Considering the actual situation, we choose correlation integral method to determine embedding dimension.

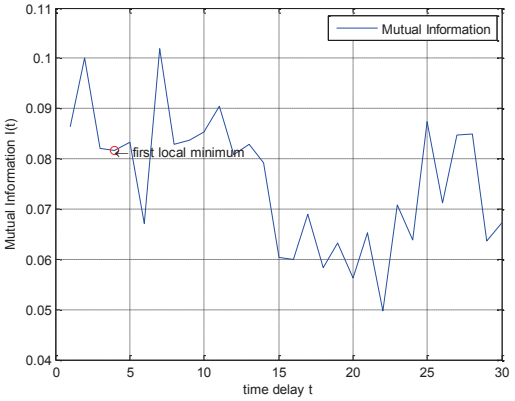

Fig. 5. Mutual Information of the Tweet: 3617517265784502

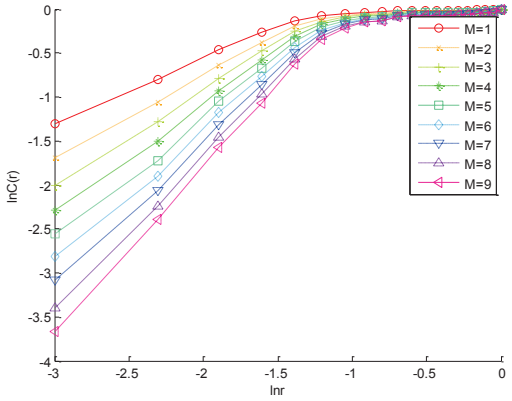

Fig. 6. The Correlation Integral of Tweet: 3617517265784502

We define the distance between two points x_i and x_j as $\|x_i-x_j\|$. For a given value r, we define C(r) as proportion of point pairs (x_i, x_j) among all possible pairs, where $\|x_i-x_j\| \le r$, as shown in Eq. (8).

$$C(r) = \frac{1}{N^2} \sum_{\substack{i,j=1,2,\ldots,N \\ (i \neq j)}} \theta(r - \|x_i - x_j\|) \tag{8}$$

where θ is defined as follows.

$$\theta(r - \|x_i - x_j\|) = \begin{cases} 1 & r - \|x_i - x_j\| \ge 0 \\ 0 & r - \|x_i - x_j\| < 0 \end{cases} \tag{9}$$

It is obvious that C(r) is a cumulative distribution function representing for the probability that the distance between a pairs of points less than r. It is also a characterization of the aggregation degree concerning to any random points. If r is large, then C(r) is almost 1. In this case, the internal features of process cannot be illustrated. However, if r is small, occasional noises may be presented. So, the selection of r is critical to the correlation dimension. Basically, we choose the value of r such that $0 \le C(r) \le 1$. As the number of points tends to be infinity, and the distance between them tends to zero, for small values of r, the correlation integral will take the form as shown in Eq. (10).

$$C(r) \sim r^m \tag{10}$$

Then the correlation dimension m can be obtained in Eq. (11):

$$m = \lim_{r \to 0} \frac{\lg C(r)}{\lg r} \tag{11}$$

Based on time delay obtained above, we increase embedding dimension until correlation integral stay stable to identify the appropriate embedding dimension. Also taking retweet time series of root tweet: 3617517265784502 as an example. The embedding dimension is about 3 shown in Fig. 6, where correlation integral becomes stable when dimension increases.

After obtaining time delay τ and embedding dimension *m*, we can reconstruct $x=\{x_n\}$ into m-dimension vectors $X=\{X_i\}$, i=1, 2, ..., M, M= N-(m-1)τ, which can be shown as following:

$$\begin{aligned} X_1 &= [x(1), x(1+\tau), \ldots, x(1+(m-1)\tau)] \\ X_2 &= [x(2), x(2+\tau), \ldots, x(2+(m-1)\tau)] \\ &\ldots \\ X_M &= [x(N), x(N+\tau), \ldots, x(N+(m-1)\tau)] \end{aligned} \tag{12}$$

B. *Chaos-LSSVM Prediction Model*

Through the phase space reconstruction, we map original time series $x=\{x_n\}$ into high dimensional space $X=\{X_i\}$. Consider the given training set of M data points $\{X_i, Y_i\}$, i=1, 2,…, M with input data $X_i \epsilon R^m$ and output $Y_i=X_{i+1} \epsilon R^m$. The retweet time series prediction problem takes the form:

$$f(X_i) = w \bullet \varphi(X_i) + b \tag{13}$$

where φ(x) denotes the high dimensional feature space which is nonlinearly mapped from the input space. This leads retweet prediction problem to determine *w* and *b*. In LS-SVM for function estimation following optimization problem is formulated:

$$\begin{cases} \min \dfrac{1}{2}\|w\|^2 + \dfrac{C}{2}\sum_{i=1}^{N} e_i^2 \\ s.t.\ Y_i = w^T \varphi(X_i) + b + e_i \end{cases} \tag{14}$$

The solution is obtained after constructing the Lagrangian:

$$L(w, b, \alpha_i, e_i) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N} \alpha_i [w^T \varphi(X_i) + b + e_i - Y_i] \tag{15}$$

with Lagrange multipliers α_i. After optimizing Eq. (15) and eliminating e_i, *w*, the solution is given by the following set of linear equations:

$$\begin{bmatrix} 0 & I^T \\ I & K + C^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{16}$$

Where $Y=[Y_1, Y_2, \ldots, Y_M]$, $I=[1, 1, \ldots, 1]$, $\alpha=[\alpha_1, \ldots, \alpha_M]$ and $K(X_i, X_k)=\varphi(X_i)^T \varphi(X_k)$, i, k=1,2,…M. Then resulting LS-SVM model for regression can be expressed as follows:

$$f(X)=\sum_{i=1}^{M}\alpha_i K(X_i,X)+b \tag{17}$$

The values obtained by Equation (17) are the predicted retweet count of every time interval. In conclusion, the steps of chaos LS-SVM prediction model are as follows:

Step 1. According to retweet time series, applying mutual information method discussed in A to find out the time delay of each time series.

Step 2. Applying correlation integral method discussed in A to figure out embedding dimension of each root tweet.

Step 3. Based on time delay and embedding dimension of time series, we conduct phase space reconstruction.

Step 4. Applying chaos LS-SVM to perform prediction task of each retweet time series.

Following the above steps, we can accomplish prediction of retweet count.

VI. Experiments and Evaluation

To evaluate Chaos LS-SVM prediction model, we conduct experiments on Sina Weibo dataset. Each root tweet has a time series P={t_1, t_2, …} composed by the posted time of each retweet. We sum the retweet count in each time interval δ and obtain retweet time series x={x_1, x_2, …, x_N}, which is our research objective. Based on these data, we conduct Chaos LS-SVM prediction model in retweet count prediction. In order to evaluate effectiveness of chaos LS-SVM prediction model, we choose the mean absolute percentage error (MAPE) as the evaluating indicator. MAPE is an accuracy measurement for prediction model, which usually expresses accuracy as a percentage. Assume that y(t) (t=1,2,…,n) is actual retweet count and y'(t) (t=1,2,…,n) is the predicted retweet count. Mean absolute percentage error (MAPE) is shown as follows.

$$MAPE=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{y'(t)-y(t)}{y(t)}\right| \tag{18}$$

Take root tweet: 3433159230421868 as an example, the observed retweet count of root tweet: 3433159230421868 is shown in blue color in the top figure of Fig. 7. The predicted retweet count obtained by chaos LS-SVM is shown in red color in the top figure of Fig. 7. From the Fig. 7, we can conclude that the predicted values are strongly consistent with the observed values. The prediction absolute percentage errors between the predicted and actual values are shown in the bottom figure of Fig. 7, which are all below 3 and the MAPE of prediction is 11.2%, showing high prediction accuracy.

We conduct the same experiments on 300 root tweets. No matter what the topic is and when the root tweet is posted, the mean absolute percentage error (MAPE) is all around 30% shown in Fig. 8 and the mean MAPE of all the root tweets is 26.22%, which illustrates that the chaos LS-SVM prediction model has higher accuracy in retweet prediction.

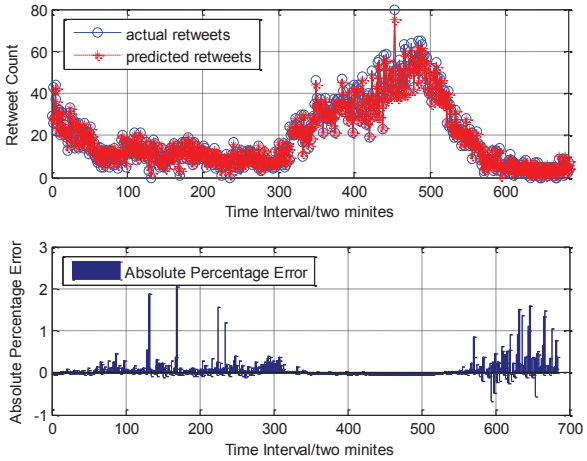

Fig. 7. The top figure is a plot of actual retweet count (blue circle) and predicted retweet count (red asterisk) versus time interval. The bottom figure is a plot of absolute percentage error versus time interval.

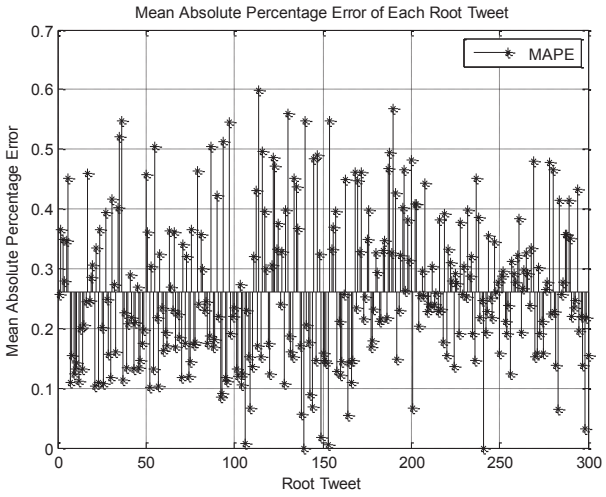

Fig. 8. Plot of Mean Absolute Percentage Errors between Actual and Predicted Retweet Count of 300 Root Tweets

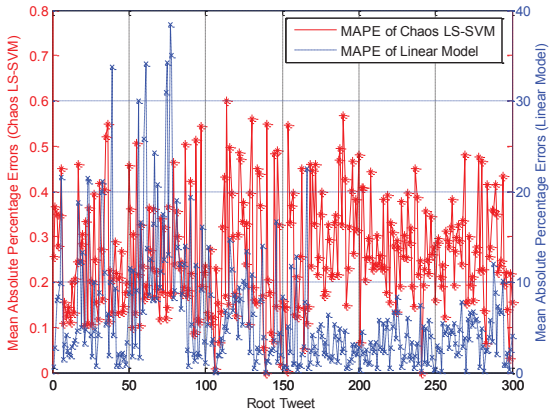

Fig. 9. Plot of Mean Absolute Percentage versus root tweet of 300 Root Tweets. The lines in red represent the MAPE of chaos LS-SVM prediction model. The lines in blue represent the MAPE of linear model. The total mean value of MAPEs of chaos LS-SVM is 26.22%. The total mean value of MAPEs of classical linear prediction model is 603.37%.

In order to compare the chaos LS-SVM model with linear prediction model, we apply both linear least square prediction method and chaos LS-SVM prediction model on the same Sina Weibo data set. Linear least square prediction model is a model of the form:

$$y_i'(t) = \alpha + \beta x_i \quad t \quad (19)$$

Where $y_i'(t)$ is independent random variable and α, β are the parameters to be solved. And can obtain the least square estimator of α and β can be written in the form:

$$\alpha = \frac{\sum_{t=1}^{n} y_i(t) - \beta \sum_{t=1}^{n} x_i(t)}{n}$$
$$\beta = \frac{n\sum_{t=1}^{n} x_i(t) y_i(t) - \sum_{t=1}^{n} x_i(t) \sum_{t=1}^{n} y_i(t)}{n\sum_{t=1}^{n} x_i^2(t) - (\sum_{t=1}^{n} x_i(t))^2} \quad (20)$$

By applying the above two models, we calculate and compare the mean absolute percentage error (MAPE) of the above two models. MAPE of both models are shown in Fig. 9. MAPE of the linear model in blue color are all higher than theirs of the chaos LS-SVM model in red color. Through statistics, we obtain that the average MAPEs of the linear prediction model and the proposed nonlinear prediction model are 603.37% and 26.22%, respectively. The comparisons demonstrate that chaos LS-SVM prediction model with higher performance is more rational to be applied for retweet prediction.

Besides, we also compare chaos LS-SVM prediction model with a Bayesian model and strawman model [17] by using MAPE. Bayesian model is a probabilistic model for evolution of the retweets and can form predictions using observation on the retweet time series and local network or "graph" structure of the retweeters. Through this method [17], for an observation of 10%, the error of the strawman model is very high (MAPE=80%) and Bayesian model is much better (MAPE=29%). The models have a MAPE of less than 40% when at least 10% of the total numbers of retweets are observed. In our chaos LS-SVM model, we only use the retweet time series in the first 24 hours after the root tweet being posted. The data we need is much less than Bayesian model. And the MAPE of chaos LS-SVM model is only 26.22%, which demonstrates the chaos LS-SVM model has better performance compared with Bayesian model and strawman model. After comparison, we conclude that chaos LS-SVM prediction model has large advantages in retweet prediction.

VII. Conclusion and Future Work

In this paper, we have first shown that retweet process is indeed chaotic by 0-1 test for chaos. Furthermore, based on such chaotic feature, we obtain the time delay and embedding dimension and conduct phase space reconstruction on retweet time series. And then we propose a chaos LS-SVM method to accomplish retweet count prediction. On real Sina Weibo dataset, we order to demonstrate the effectiveness of chaos LS-SVM model in retweet prediction.

Acknowledgment

This work was supported by National Key Technology R&D Program of China under Grant No.2012BAH46B04.

References

[1] M. Shi, D. Wang, J. Zhang. An Improved Method of Detecting Chaotic Motion for Rotor-Bearing Systems. Journal of shanghai Jiaotong University (Science), 2013, V18(2): 229-236.

[2] Ashhan Albostan, Bihrat Önöz. Implementation of Chaotic Analysis on River Discharge Time Series. Journal of Energy and Power Engineering, 2015, 7, 81-92.

[3] J. Jiang, et al. Modeling of chaotic vibrations in symmetric vocal folds. J Acoust Soc Am, 2002, 110(4): 2120.

[4] W. Galuba, K. Aberer. Outtweeting the Twitterers – Predicting Information Cascades in Microblogs. Conference on Online Social Networks (WOSN), 2010.

[5] G. V. Steeg, R. Rhosh, K. Lerman. What Stops Social Epidemics?. In Proc. of Inter. AAAI Conf. on Weblogs and Social Media, 2011.

[6] Danah Boyd, Scott Golder, Gilad Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. HICSS, 2010.

[7] Zi Yang, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang, Zhong Su. Understanding Retweeting Behaviors in Social Networks. CIKM, 2010.

[8] Bakshy E., Hofman J. M., Mason W. A., Watts D. J. Everyone's an Influencer: Quantifying Influence on Twitter. In Proc. WSDM, 2010.

[9] Suh B., Hong L., Pirolli P., Chi E. H. Wanted to be Retweeted? Large Scale Analysis on Factors Impacting Retweet in Twitter Network. In IEEE International Conference on Social Computing, 177-184.

[10] Bandari R., Asur S., Huberman B. A. The Pulse of News in Social Media: Forecasting Popularity. In AAAI Conference on Weblogs and Social Media, 2012.

[11] Hong L., Dan O., Davison B. D. Predicting Popular Messages in Twitter. In WWW, 2011.

[12] D. Wang, H. Park, G. Xie. A Genealogy of Information Spreading on Microblogs: a Galton-Watson-based Explicative Model. 32nd IEEE INFOCOM, April 2013, Turin, Italy.

[13] F. Wang, H. Wang, K. Xu. Diffusive Logistic Model Towards Predicting Information Diffusion in Online Social Networks. In Proc. of ICDCS Workshop, 2012.

[14] Yang Yang, Jie Tang, Cane Wing-ki Leung, Yizhou Sun, Qicong Chen, Juanzi Li, Qiang Yang. RAIN: Social Role-Aware Information Diffusion. AAAI, 2014.

[15] J. Yang, J. Leskovec. Modeling Information Diffusion in Implicit Networks. In Proc. of IEEE Inter. Conf. on Data Mining, 2010.

[16] Szabo G., Huberman B. A. Predicting the Popularity of Online Content. Communications of the ACM, 2010, 8, 80-88.

[17] Tauhid Zaman, Emily B. Fox, Eric T. Bradlow. A Bayesian Approach for Predicting the Popularity of Tweets. Annals of Applied Statistics, September, 2014, 8, 3, 1583-1611.

[18] G. A. Gottald, I. Melbourne. On the implementation of the 0-1 test for chaos. SIAM Journal on Applied Dynamical Systems, 2009, 8: 129-145.

[19] G. A. Gottald, I. Melbourne. Testing for chaos in deterministic systems with noise. Physica D, 2005, 212: 100-11.