Time Series Analysis

Implementation of Chaotic Analysis on Retweet Time Series

Retweet

- Three main actions on social network:
    - Post
    - Retweet
    - Mention
- Retweet is most powerful mechanism to diffuse information
- Purpose of retweeting a post is to copy the entire post and retweet to one’s followers.

Retweet Time Series

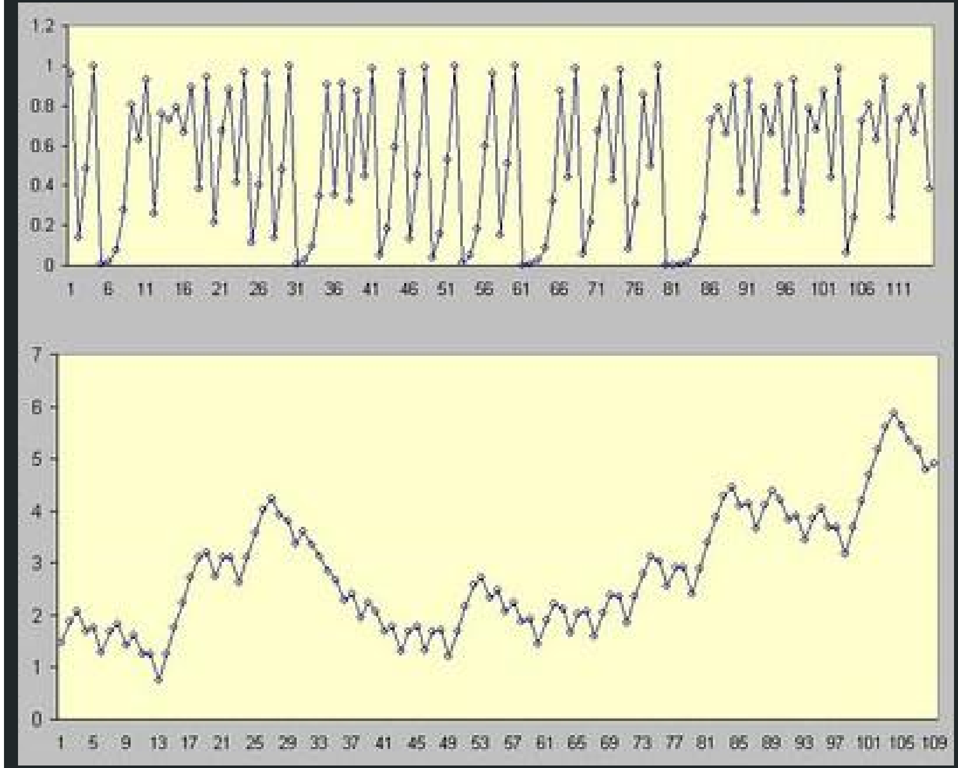
- 300 different tweets with highest retweet counts are chosen from the dataset. They are called as root tweets.
- For each root tweet we have a separate time series. Thus in total we have 300 time series.
- The retweet count in each time interval "**δ**" forms the retweet time series $x=\{x_1,x_2,x_3 \ldots x_N\}$ having length "24*60/**δ**"

Objectives

- Investigate whether the retweet time series is chaotic
- Conduct a phase space reconstruction on the time series
- Implement Least Squares Support Vector Machine (LS-SVM) on phase space.
- Predict retweet number using only information of early retweet time series.

Chaos Analysis

- ➔ X(t+1)=r*X(t)*(1-X(t))
  - ➔ Image on top - X(0)=1
  - ➔ Image on bottom - X(0)=2

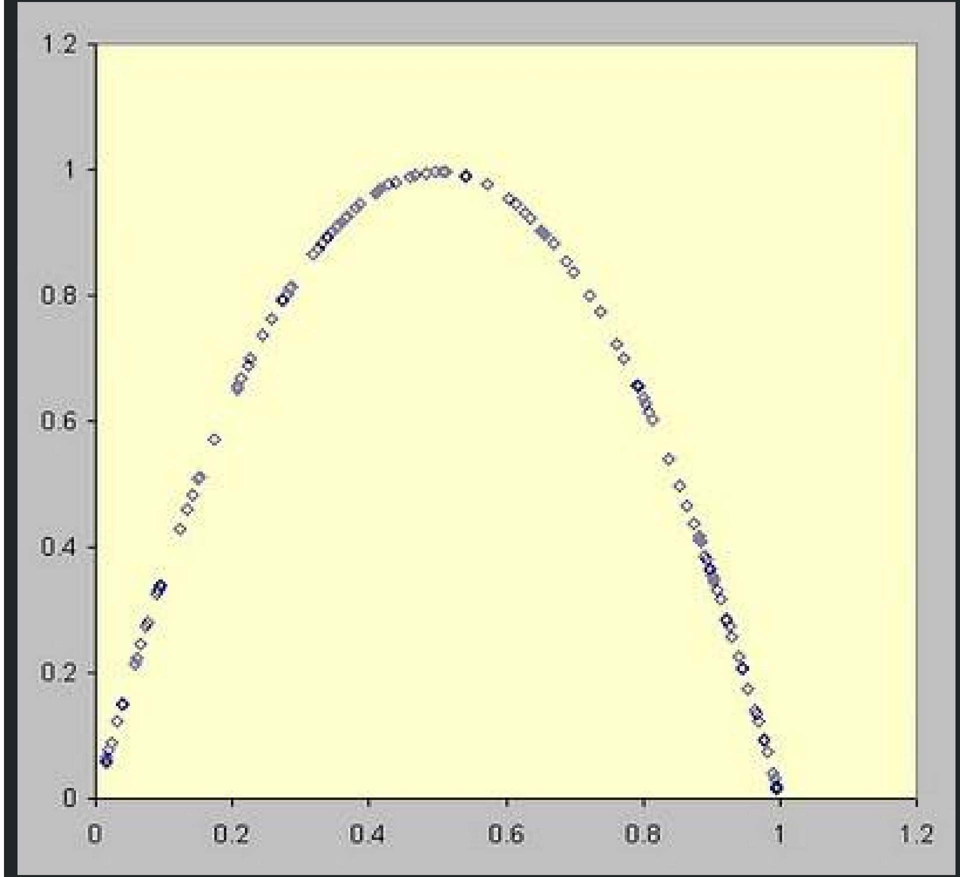


Chaos Analysis

- → Nonlinear Dynamic Behaviour
- → Seemingly Random but Deterministic
- → Heavily dependent on initial conditions

Chaos Analysis

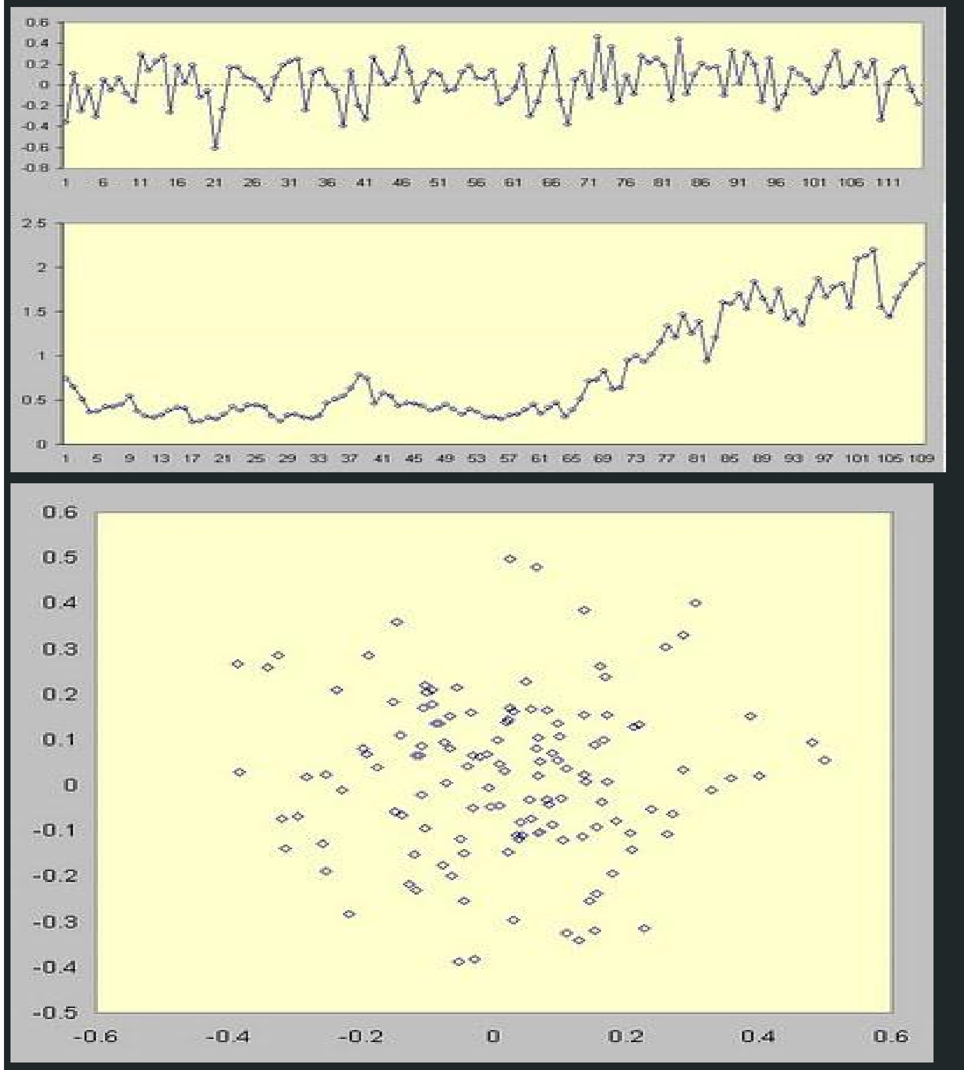
- How to forecast such series?
- Phase state plot to capture the sequence.
- The phase state plot is shown in the figure
- Using it one can predict the values of the series in future



Random Walk Plot

- ➔ It does not just appear random, but it is random as shown in image above

- ➔ Phase State Plot shows randomness in the image below



0-1 Test for Chaos

- Define two translation variables p(n), q(n)

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

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

- Obtain Mean Square Displacement M(n)

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

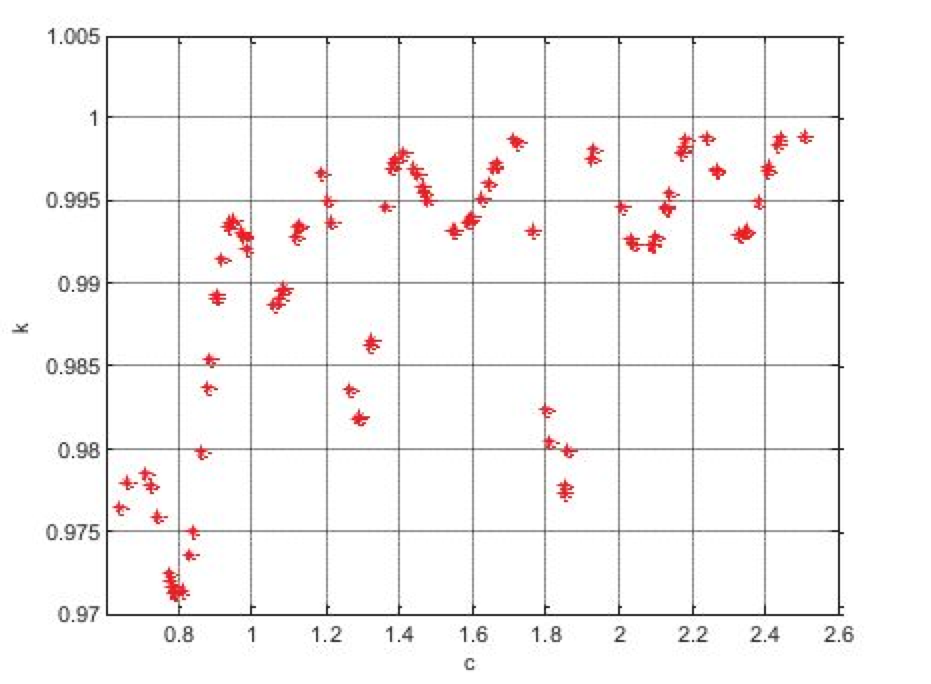
where $M_c(n)$ and $E(\varphi)$ are defined as

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

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

0-1 Test for Chaos

- ➔ Compute the Asymptotic Growth Rate from the plot
- ➔ If the time series is regular then $K_c = 0$
- ➔ If the time series is chaotic the $K_c = 1$
- ➔ Retweet series has K_c approximately equal to 1. Thus it's a chaotic process.



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

0-1 Test for Chaos

- ➔ Distribution of K_c versus number of root tweets.
- ➔ 90.33% of root tweets have K_c larger than 0.7
- ➔ 71.67% of root tweets have K_c larger than 0.9
- ➔ Hence retweet time series is chaotic in nature

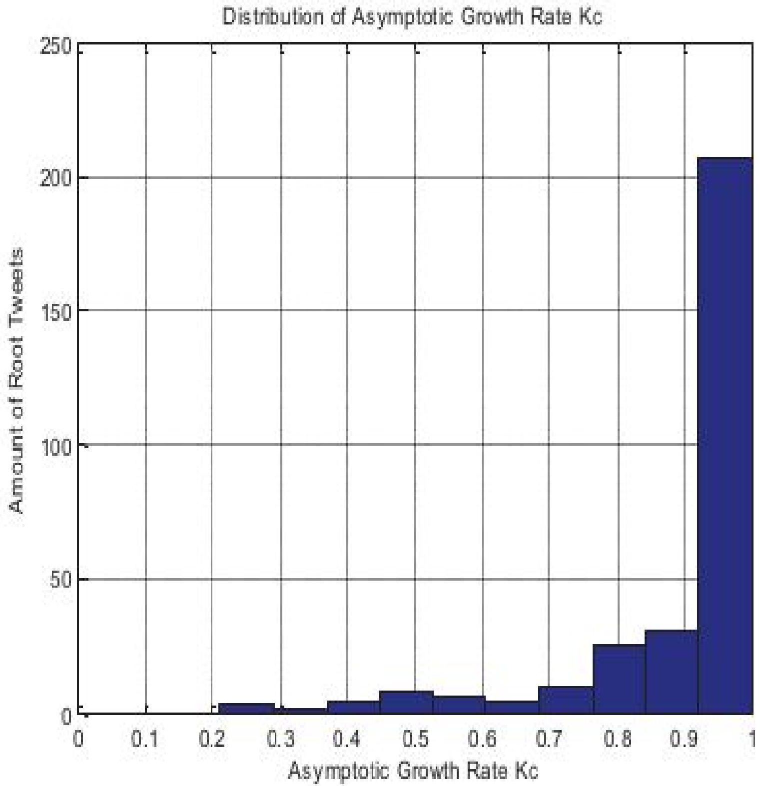


Fig. 4. Plot of distribution of asymptotic growth rate. The asymptotic growth rates of 90.33% of root tweets are larger than 0.7. The asymptotic growth rate which is larger than 0.9 accounts for 71.67%.

LS-SVM model

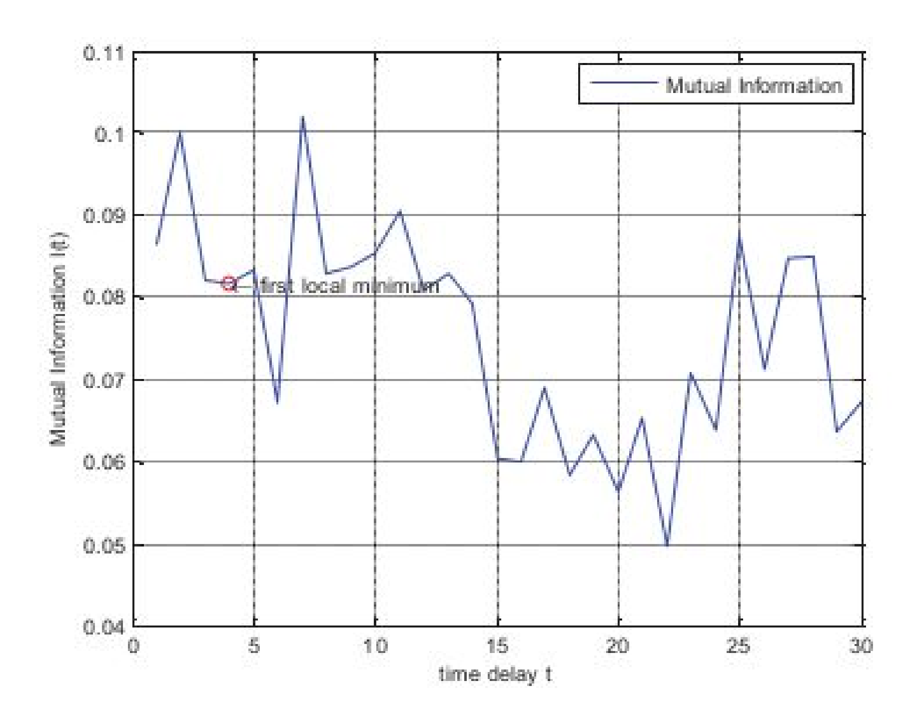
The steps of Chaos LS-SVM prediction model are as follows:

1. Apply mutual information method to find out time delay.
2. Apply correlation integral method to find out the embedding dimension of each root tweet.
3. Perform phase space reconstruction.
4. Apply chaos LS-SVM to perform prediction task and forecast

Time Delay

- Mutual information between two variables is calculated using the formula below

- Time delay "τ" is the duration when the mutual information I(τ) first reaches a local minimum value



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

Embedding Dimension (m)

➔ C(r) is the cumulative distribution function representing the probability that the distance between a pair of points is less than r.

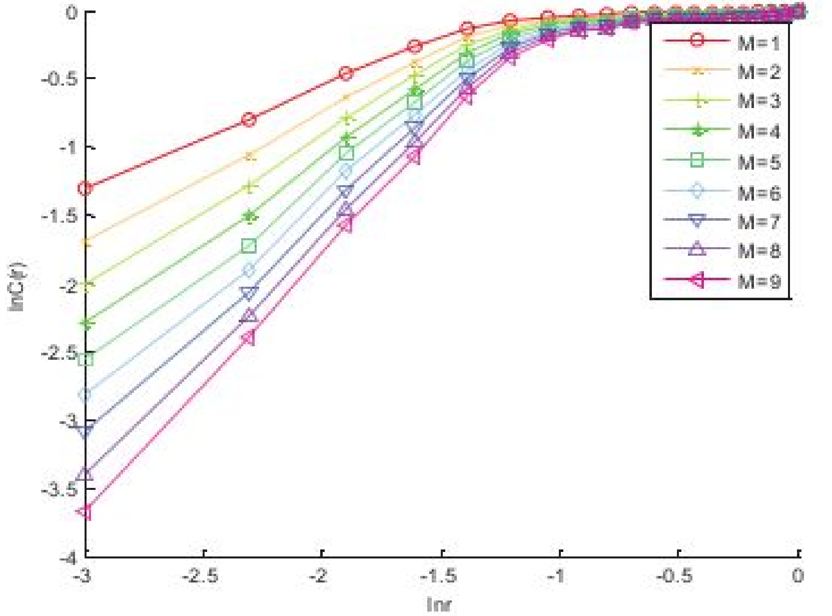


Fig. 6. The Correlation Integral of Tweet: 3617517265784502

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

where θ is defined as follows.

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

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

Constructing Phase Space

- After obtaining time delay "τ" and embedding dimension "m" we re-construct $x=\{x_n\}$ into M dimensional vectors $X=\{X_i\}$, $i=1,2,\ldots.M$, $M=N-(m-1)\tau$, where X is defined as follows:
- $X_1 = \{x_1, x_{(1+\tau)}, \ldots, x_{(1+(m-1)\tau)}\}$
- $X_2 = \{x_2, x_{(2+\tau)}, \ldots, x_{(2+(m-1)\tau)}\}$
- …………
- $X_M = \{x_M, x_{(M+\tau)}, \ldots, x_{(M+(m-1)\tau)}\}$

Chaos LS-SVM Prediction Model

- We map original time series $x=\{x_n\}$ into high dimensional space $X=\{X_i\}$.
- Training set has "M" data points $\{X_i, Y_i\}$; i=1,2…M; $Y_i=X_{i+1}$.
- LS-SVM takes the form:

$$f(X_i) = w \bullet \varphi(X_i) + b$$

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

Experiments and Evaluation

- Dataset : Sina Weibo(a twitter like website popular in china) dataset.

- Effectiveness measure used : MAPE (Mean Absolute Percentage Error).

MAPE(Mean Absolute Percentage Error)

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

where, n is number of retweet intervals considered,

y(t) is the actual retweet count,

y’(t) is the predicted retweet count.

Comparison of LS-SVM with other models.

MODEL	**MAPE**(accuracy)
LS-SVM	26.22%
Bayesian model	29%
Linear prediction model	603.37%

Results and Analysis

- We performed the 01-chaos test on about 397 time series, 81% of them were found to be chaotic.
- On the 326 chaotic time series we implemented the proposed prediction model and got an average MAPE of 27%.
- We also found out that the given model is highly unstable for retweet series with a sudden spike in the number of retweets.

THANK YOU!