Recurrence Plots

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

Abstract—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

Index Terms—component, formatting, style, styling, insert

I. Mutual Information

In order to speed up the identification of malware, we need to compress some of the information, we found a good method named recurrence plots(RP), which can make good use of the dynamic characteristics of the RP for major information extraction. To achieve this goal we need require the knowledge of two parameters that are the delay parameter τ and the embedding parameter D. therefor there are two the methods to estimate these parameters, that are average mutual information(AMI) function and the False nearest neighbors function. For pattern of the malware contains the time series $x(t) = x(t_0 + k\tau)$. we can find the time delay τ to get other sample data. These data are just different in their initial states. That is we can use these time series construct a new time series $y(t)$ of D-dimensional points from the original series $x(t)$ as follows:

$$y(t) = (x(r), x(t+\tau), \ldots, x(t+(D-1)\tau)) \tag{1}$$

Where, two parameters τ and t are integers used to index the sample data. We can also use average mutual information function to calculate the mutual information of series, the formula is

$$I(x(t), x(t+\tau)) = \sum_{i,j} p_{ij}(\tau) \log \frac{p_{ij}(\tau)}{p_i p_j} \tag{2}$$

where the p_i and p_j denote the associated probability of the whole system which consists of the $x(t)$ and $x(t+\tau)$, respectively. $p_{ij}(\tau)$ is the joint probability of the associated probabilities for the $x(t)$ and $x(t+\tau)$, we must select the size of bins to calculate the mutual information of the sequence data, and ensure it is more accurate. Then we find local minimum value of the average mutual information with a certain range time lag. if the value is relatively large, then we set a appropriate threshold, the parameter corresponding to this values is the better parameter τ. But sometimes the τ is too large, thus it would not be the real dimension of the sequence series. So we consider the point of AMI where the series changes significantly. We find several significant points and choose the minimum. The minimum value is what we want, We think this value is the real dimensions of the sequence data. This gives a different auto mutual information function, we get optimal value.

Identify applicable funding agency here. If none, delete this.

II. False Nearest Neighbors

In the nonlinear dynamics, utilizing the time delay from scalar and multiple time series to reconstruct the phase space. The diagram can estimated the character of the original system, such as Lyapunov exponents dimensions and prediction, etc. A time series can be reconstructed D-dimension point, the formula is as follow:

$$y(t) = (x(t), x(t+\tau), \ldots, x(t+(D-1)\tau)) \tag{3}$$

where both t and τ are integers used to index the sample. The parameter τ named time delay that can be estimated by mutual information mentioned before. In this paper [1] the authors show the method that can use a geometrical construction to provide a appropriate value as the embedding dimension. Then according the Takens' theorem the parameter D can be using a geometrical construction to determining embedding dimension for the phase space. If the distance of two neighbouring point of the time series changes appreciably, then these point are dubbed false neighbors. This means we need to embed the higher dimensions for the phase space. In general, for all point $\vec{y}(t)$ in the time series we look for rth nearest neighbor $\vec{y}^{(r)}(t)$ in a D-dimensional phase space. we can calculate the square of the Euclidean distance $R_D(t,r)$ of these two points. The formula is as follow:

$$R_D^2(t,r) = \left|\vec{y}^{(r)}(t) - \vec{y}(t)\right|^2 \tag{4}$$

If we substitute the Eq. (3) into Eq. (4) which can be written as

$$R_D^2(t,r) = \sum_{k=0}^{D-1} \left[x(t+k\tau) - x^{(r)}(t+k\tau)\right]^2 \tag{5}$$

Then we use the time delay embedding and go from D-dimensional to $(D+1)$-dimensional phase space. So in the $(D+1)$-dimensional phase space we can calculate the Euclidean distance between $\vec{y}(t)$ and the same rth nearest neighbor. we get

$$R^2_{D+1}(t,r) = R^2_D(t,r) + \left[x(t+D\tau) - x^{(r)}(t+D\tau)\right]. \quad (6)$$

How to judge the false nearest neighbor for two points? In the paper [] we know that we need to set a natural criterion R_{tol}. In their numerical work, they find that for $R_{tol} \geq 10$ the false neighbors are clearly identified. The criterion is by designating as follow:

$$\begin{aligned} &\left[\frac{R^2_{D+1}(t,r) - R^2_D(t,r)}{R^2_d(t,r)}\right]^{1/2} \\ =&\frac{\left|x(t+k\tau) - x^{(r)}(t+k\tau)\right|}{R_d(t+k\tau)} > R_{tol} \end{aligned} \quad (7)$$

The R_{tol} is some threshold. They set $R_D(t) \equiv R_d(t, r=1)$ and compare with the size of the attractor R_A.

$$\frac{R_{D+1}(t)}{R_A} > A_{tol} \quad (8)$$

where the R_A is

$$R_A = \sqrt{\frac{1}{T}\sum_{t=1}^{T}[x(t) - \bar{x}]}. \quad (9)$$

The $\bar{x}$ is the mean value of the $x(t)$

References

[1] Kennel, M. B., Brown, R., and Abarbanel, H. D. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. Phys. Rev. A 45:3403. doi: 10.1103/PhysRevA.45.3403