The Typhoon Tracks Analysis using Tri-plots and Markov chain

John Chien-Han Tseng[1], Hsin-Kuo Pao[2], Christos Faloutsos[3]
[1]Central Weather Bureau, Taipei, Taiwan
[2]Department of computer science and Information Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan
[3]School of Computer Science, Carnegie Mellon University, PA, USA

Abstract

Based on the fractal dimension, the tri-plots can classify two large and not equal sizes of the time series datasets. The tri-plots measure three function values which include two self-plots and one cross-plot. The self-plot affords the character of one individual dataset. The cross-plot describes the relation between two datasets. Originally, the tri-plots just can get the relation in two datasets, but we can use tri-plots many times for multi-datasets. The time series data like typhoon trajectories, we are interested in the differences among the different annual events, e.g. ENSO and La Niña. In here, we propose the tri-plots method to analyze and classify different annual typhoon trajectories.

On the other hand, the Markov chain is used to deal with the time series data in data mining filed. Markov chain establishes the probability relation between two consecutive time steps and estimate one model for one trajectory. Basically, every trajectory has own probability model. We can repeat this process until all datasets finished computing. In implementation, we combine several trajectories to be one trajectory in order to corresponding the physical meaning and saving executing time. After all trajectories of all datasets finished estimating their own model, the dissimilarity matrix can be given by comparing all trajectory models pairs, that is, the dissimilarity matrix describes the relations between the trajectories. So, we use the Markov chain to be another alternative method for different annual events trajectories classification problems.

After the calculation of the tri-plots, the ENSO and La Niña years typhoon tracks can be separated by the classifier, the smooth support vector machines（SSVM）, which can get the training error about 0.023~0.268 and the testing error about 0.271~0.334. For Markov chain with the threshold of the pace （$\Delta\lambda \leq 10$ ）, the SSVM classifier can get the training error around 0.031~0.173 and the testing around 0.181~0.287. Moreover, the tri-plots or Markov chain concentrate the information of all events to one distribution figure that presents the dissimilarity of these typhoon trajectories or depicts which years should be probably regarded as one group. We believe that they can be very helpful for realizing ENSO and La Niña atmospheric circulation and for establishing typhoon databases. In other words, we think tri-plots and Markov chain can be use to find the intrinsic patterns in other traditional weather data.

Key word: fractal dimension, tri-plots, self-plot, cross-plot, Markov chain, smooth support vector machines

1. Introduction

It is an important and meaningful work of the typhoon tracks or trajectories classifications which can be used to diagnose the atmospheric circulations and establish typhoon databases. Some researches（Lee et al., 2007; Camargo et al., 2007a, b）focus on the different shapes or the clusters of the typhoon trajectories. That means they divide the typhoon trajectories into several groups; for example, the east-westward movement typhoons, or the recurved movement having more north-southward component movement typhoons, or maybe the typhoons close to the continent, or the typhoons trapped in some areas, etc. In brief, they describe the typhoon tracks themselves. The method of Lee et al.（2007） slices the typhoon tracks to different segmentations by different moving directions, and uses the minimum description length（MDL）to cluster the typhoon tracks. The method of Camago et al.（2007a, b）just follows Gaffney and Smyth（1999）trajectory clustering with mixtures regression models, and arranges the typhoon tracks into 7 clusters. The common point of these researches regards the every single typhoon track as one sequence, but we think we probably could check typhoons by collecting a period of time of typhoon cases to be one sequence. This sequence can represent the ensemble view of the period of time-space structure. When we compare these different time-space structures, the meaningful classification information also can be extracted.

In atmosphere, one of the most significant annual circulation variation events is ENSO（El Niño southern oscillation）or La Niña, the anti-ENSO. When the El Niño or ENSO event happens, the anomaly of equatorial sea surface temperature area will move to the central Pacific Ocean and that definitely causes the typhoons generation area eastward and then make the shapes of the typhoon tracks be changed. The purpose of this study is

to classify the typhoon trajectories in a period of time will be belonged to El Niño or La Niña events.

The tri-plots（Traina et al., 2001） are one kind of data mining tools to finding the intrinsic patterns between two large and multidimensional datasets. We think the tri-plots are suitable for our research. The tri-plots are composed of two self-plots and one cross-plot. Through comparing two datasets after tri-plots, the two self-plots return the intrinsic characters of the two individual datasets and the cross-plot returns the relation character between two datasets. Taking all El Niño and La Niña events data and comparing two datasets in turn, we can depict the distribution by self-plots and one kind of the distance measurements by cross-plots. The distances from cross-plots can be inputted into ISOMAP（isometric feature mapping）. After that, the SSVM（smooth support vector machines）is used to classify the pattern from ISOMAP.

At the same time, Markov chain properties （Lin, 2009; Bishop, 2006）also can help us to extract the hidden pattern of the typhoon trajectories. The two consecutive time steps of the typhoon coordinates can be described by the probability. In here, the one sequence（a period of time typhoon tracks）, now is controlled or described by the probability model, then we measure all the dissimilarity distances between any two models. Similarly, the dissimilarity distance we established can be the distances of the ISOMAP method and the SSVM is used for the following classification problem. At the same time, Risi（2004）also use Markov chain to analyze hurricane tracks, but this report focuses on the predictions of the tracks.

2. Data

The typhoon trajectories data are from Japan Meteorological Agency（JMA）. The data recorded the western Pacific ocean which covers the west of the longitude 180E to the east of the 100E typhoon centers movements（positions）and other observation variables, e.g., pressure, wind speed, etc. The data features in this study include longitude, latitude, minimum pressure, and average wind speed of the typhoon center. Currently, the JMA data recorded from 1951 to 2009 typhoon trajectories. The time resolution is about 3~6 hours. Furthermore, the high level Meteorological data are from National Center for Environment Prediction（NCEP）reanalysis-2 data. We use higher level winds data in this study. The ENSO years and the LaNiña years are based from NOAA's definition by Nino 3.4 index. In order to use the NCEP reanalysis-2 data, we just focus all the time events after 1980. The ENSO years are 7 events and the labels are set to be 1. The LaNiña years are 5 events and the labels are -1. Also, we have 10 neutral events.

3. Performance of Tri-plots

Traina et al. (2001) proposed the tri-plots based on fractal dimension to calculate the distribution of distances between two datasets points. We think that the typhoon trajectories can be classified by tri-plots analysis. The tri-plots include three kinds plot analyses: cross-plot (two datasets), self-plot (one dataset), and self-plot (the other dataset). Assuming there are two datasets *A* and *B*, and the cross-plot function is defined as

$$Cross_{A,B}(r) = \log(\sum_i C_{A,i}\ C_{B,i}) \quad ,$$

where $C_{A,i}$（$C_{B,i}$）is the number of points from set *A*（*B*）in the *i*-th cell, and *r* is the distance of the pairs of points. Hence, the cross-plot function is proportional to the count of *A*-*B* pairs within distance *r*, and the cross-plot is the figure of the cross-plot function versus log(*r*). The self-plot function is defined as

$$Self_A(r) = \log(\frac{\sum_i C_{A,i} \cdot (C_{A,i} - 1)}{2})$$

and the self-plot is the figure of the self-plot function versus log(*r*). If *A* is self similar, then the self-plot of *A* is like straight line and the slope is just the intrinsic dimension which is correlated with fractal dimension（Belussi and Faloutsos, 1995）.

The features we used in this study are longitude, latitude, minimum pressure, u(300 hPa), v(300 hPa), and topography effect. The slope and intercept based on self-plots function of tri-plots could be regarded as the point in space. We could plot the distribution of all the annual ENSO and La Niña events together. The results are shown in Figure 1. Basically, the ENSO points(red) and the La Niña points(blue) are separated into two groups. Under this kind of features selection, the cross-plots function are calculated and regarded as the distance of the ISOMAP. We choose the intrinsic dimensionality equal to 5 after ISOMAP, and then the reconstructed structure are classified by SSVM. The classification results（Table 1）are: training error around 0.023~0.268 and the testing error about 0.271~0.334.

4. Performance of Markov chain

We assume the typhoon trajectories can be represented as $s = (x_1, x_2, ..., x_t, ..., x_T)$, where *x* is the coordinate value of typhoon center composed of longitude and latitude; the subscripts 1, 2, …, *T* mean the different time steps of the typhoon positions. Then, we define the pace vector $x_{t+1} - x_t$, and the Euclidean pace size $\lambda = \|x_{t+1} - x_t\|$. So the information of pace change is $\Delta\lambda_t = \lambda_{t+1} - \lambda_t$. The angle θ_t is between $x_{t+1} - x_t$ and x-axis. The information of angle change is $\Delta\theta_t = \theta_{t+1} - \theta_t$. Let $m(\sigma_\lambda, \sigma_\theta)$ denote the model of trajectory sequence, and the transition parameters are σ_λ, σ_θ which represent the standard deviation of pace changes and angle changes.

According to Markov chain theory, the probability between two consecutive time steps x_t and x_{t+1} can be described as

$$P(\lambda_{t+1} | \lambda_t) \sim N(\lambda_t, \sigma_\lambda^2)$$
$$= \frac{1}{\sqrt{2\pi}\sigma_\lambda} \exp(-\frac{(\Delta\lambda_t - \Delta\lambda_{mean})^2}{2\sigma_\lambda^2})$$
$$P(\theta_{t+1} | \theta_t) \sim N(\theta_t, \sigma_\theta^2)$$
$$= \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp(-\frac{(\Delta\theta_t - \Delta\theta_{mean})^2}{2\sigma_\theta^2})$$

The log-likelihood $\ell(s\,;m)$ associated with trajectory *s* and model *m* is written as

$$\begin{aligned}\ell(s\,;m) &= \log L(s\,;m)\\ &= \log(P(x_1)\prod_{t=1}^{T} P(x_{t+1}|x_t))\\ &= \log P(x_1) + \sum_{t=1}^{T}\log(P(x_{t+1}|x_t))\end{aligned}$$

The code length of the trajectory *s* now can be the form of negative log-likelihood as

$$c(s|m) = -\ell(s\,;m) = -\log L(s\,;m).$$

Hence, the dissimilarity value *d* is

$$d(s_1,s_2) = \frac{1}{2}\left(\frac{c(s_1|m_2)}{c(s_2|m_2)} + \frac{c(s_2|m_1)}{c(s_1|m_1)}\right)$$

In this equation, it estimates the effect of one trajectory and the other model. The dissimilarity matrix *D* can be given after calculating all the trajectory pairs $d(s_i,s_j)$. In this study, in order to having representative typhoon trajectories and saving computing time, we combine several typhoon tracks to be one trajectory. There are two models described by dissimilarity matrix $D_{\Delta\lambda}$ and $D_{\Delta\theta}$ models, which can combine to one target dissimilarity as follows

$$D_{target} = \alpha D_{\Delta\lambda} + (1-\alpha)D_{\Delta\theta}\,,\quad 0 \le \alpha \le 1.$$

It is one kind of trade-off between $D_{\Delta\lambda}$ and $D_{\Delta\theta}$. In chapter 4, we will see the different α how to affect the results.

Because of collecting a period time typhoon trajectories as one sequence, the value of pace between the ending point of one trajectory and the beginning point of another trajectory is larger than the pace of two consecutive time steps. We can choose the threshold with $\Delta\lambda = 10$ and ignore all the values of $\Delta\lambda > 10$. Again, we can calculate the Markov chain dissimilarity matrix and SSVM with the intrinsic dimensionality 5 from ISOMAP, then we get the results in Figure 2 and Table 2. In Figure 2, we show the 3-D structure from ISOMAP. The La Niña 9596 event located around left corner with blue inverse triangle is the most different case from the ENSO 8688 or ENSO 8283, and this result is consistent with the tri-plots（Figure 1）. Moreover, the blue marks area around $0 \le x \le 4,\ 0 \le y \le 2,\ -1 \le z \le 2$ in Figure 2 are from 8889, 9800 and 0708 which somehow show the trajectories close to Asian continent. Based on 10-fold cross validation, the training error 0.031~0.173 and the testing around 0.181~0.287.

5. Conclusions and discussions

In this study, we proposed the quantitative and objective tools, tri-plots and Markov chain model, to distinguish the differences between the different yearly typhoon tracks events. Just based on statistical learning, we have about 70% accuracy to classify the typhoon tracks belonged to El Niño or La Niña events. In tri-plots, we afford a global view to see different annual events by the self-plots distribution, and the result is consistent with the report of World Meteorological Organization（WMO, 2009）, that is, no two El Niño events are identical. Moreover, the self-plots distribution affords the possible clue to do the next clustering work of El Niño or La Niña events. Besides this, the tri-plots experiments tell us what kinds of features we probably should have in the classification problems or maybe we should have in typhoon databases. Until now, the world typhoon databases just store the low level features of the typhoons. Adding the high level winds or the consideration of topographical effect should be included in typhoon databases. Meanwhile, the probability estimations of the Markov chain got better classification results just by considering the normal distribution probability between two consecutive time steps. However, we think we should consider more in the future. In this study, we had used Chi-square distribution to estimate the relation between two steps but we cannot get the better results（not shown）. We think Markov chain still has more potential to do the following classification researches.

Finally, either tri-plots or different Markov chain models, can be extended to other traditional Meteorological data analysis. Because the quantity of the Meteorological is very large; for example, the space resolution of NCEP reanalysis data is about 144x73 in horizontal; the 17 layers in vertical for one specific feature and the time resolution is about 4 times in one day. So, when we extend our methods to analyze the annual events, we need to modify our current algorithm. First, we need alternative method to solve the eigenproblem in ISOMAP（solving large and sparse matrix）, and we need more efficiency box-counting data structure in tri-plots.

6. References

Belussi, A. and C. Faloutsos, 1995: Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In *the 21-th conference of Very Large Data Bases*, 299-310.

Bishop, C. M, 2006: *Pattern recognition and machine learning*. Springer press.

Camargo, S, J., A. W. Robertson, S. J. Gaffney, and P. Smyth, 2007a: Cluster analysis of typhoon tracks. Part I: general properties. *Journal of Climate*, **20**, 3635-3653.

Camargo, S, J., A. W. Robertson, S. J. Gaffney, and P. Smyth, 2007b: Cluster analysis of typhoon tracks.

Part II: large circulation and ENSO. *Journal of Climate*, **20**, 3654-3676.
Chan, J. C. L. and W. M. Gray, 1982: Tropical cyclone movement and surrounding flow relationships. *Monthly Weather Review*, **100**, 1354-1374 .
Emanuel, K, 2005: *Divine wind: the history and science of hurricanes*. Oxford university press.
Gaffney, S. and P. Smyth. Trajectory clustering with mixtures of regression models. In *the International Conference on Knowledge Discovery and Data Mining*, pages 63-72, 1999.
Harr, P. A. and R. L. Elsberry, 1991: Tropical cyclone track characteristics as a function of large-scale circulation anomalies. *Monthly Weather Review*, **119**, 1448-1468.
Hsieh, W. *Machine learning methods in the environmental sciences*. Cambridge university press, Cambridge, 2009.
Lee, J.-G., J. Han, and K.-Y. Whang, 2007: Trajectory clustering: A partition-and-group framework. *International Conference on Management of Data*, 593-640.
Lee, Yuh-Jye and O. L. Mangasarian,2001: SSVM: A smooth support vector machine for classification. *Comput. Optim. Appl*., **20(1)**, 5-22.
Lin, H, 2009: *Trajectory based on behavior analysis for verification and recognition*. National Taiwan University of Science and Technology, master thesis, Taipei.
Miller J., P. B. Weichman, and M. C. Cross, 1992: Statistical mechanics, Euler's equation, and Jupiter's Red Spot. *Physics Review*, **A45**, 2328–2359.
Risi, C, 2004: *Statistical synthesis of tropical cyclone tracks in a risk evaluation perspective*. Massachusetts Institute of Technology, internship report, Cambridge.
Strogatz, S. H, 2001: *Nonlinear dynamics and chaos: with applications to Physics, Biology, Chemistry, and engineering*. Westview press.
Tenebaum, J. B., V. d. Silva, and J. C. Langford, 2000: A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319-2323.
Traina, A., C. Traina, S. Papadimitriou, and C. Faloutsos, 2001: Tri-Plots: Scalable tools for multi dimensional data. In *the International Conference on Knowledge Discovery and Data Mining*, 184-193.
Vlachos, M. D. Gunopulos, and G. Kollios, 2002: Discovering similar multidimensional trajectories. In *the International Conference on Data Engineering*, 673-684.
Wang, B., H. Rui and J. C. L. Chan, 2002: How strong ENSO events affect tropical activity over the western North Pacific. *Journal of Climate*, **15**, 1643-1658.
Webster, P. J., G. J. Holland, J. A. Curry, and H.-R. Chang, 2005: Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, **309**, 1844-1846.
World Meteorological Organization, 2009: *El Niño/La Niña update*.

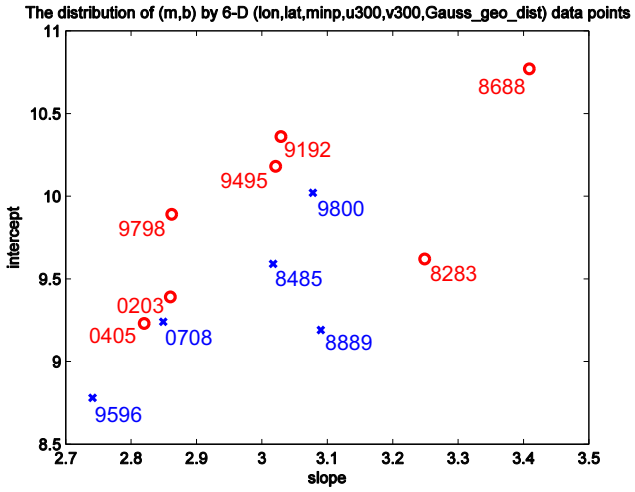

Figure 1: The distribution based on the slope and intercept of the tri-plots. The longitude, latitude, minimum pressure, u, v and topography effect are used.

Table 2: Error table of the SSVM based on cross-plots and ISOMAP

k	5	6	7	8	9	10
SSVM						
Training error	0.268	0.032	0.116	0.025	0.023	0.027
Testing error	0.334	0.271	0.358	0.289	0.320	0.327

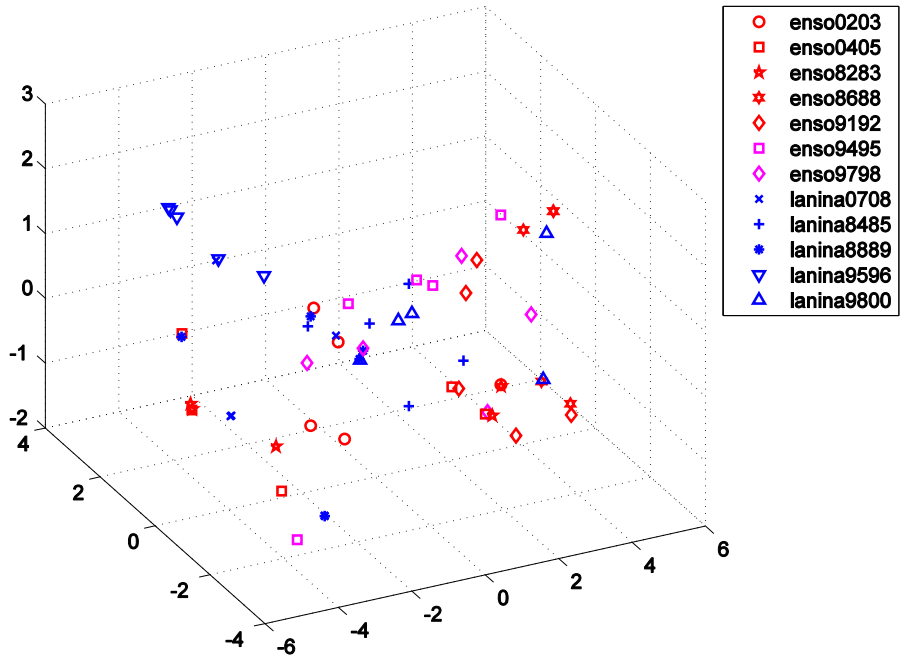

Figure 3: The 3-D structure from ISOMAP with k=4 based on threshold $\Delta\lambda \leq 10$ Markov Chain experiments with $\alpha = 0.8$ in dissimilarity matrix calculation.

Table 4: Error table of the SSVM based on Markov chain with thresho$\Delta\lambda \leq 10$ and ISOMAP(k=4)

α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SSVM											
Training error	0.173	0.052	0.135	0.048	0.161	0.168	0.154	0.037	0.031	0.040	0.045
Testing error	0.277	0.253	0.248	0.223	0.287	0.287	0.280	0.196	0.181	0.197	0.248