2009 International Symposium on Nonlinear Theory and its Applications
NOLTA'09, Sapporo, Japan, October 18-21, 2009

Hypothesis Testing for Feature Patterns Using Collective Synchronization in a Network of Non-Symmetrically Coupled Phase Oscillators

Takaya Miyano† and Takako Tsutsui‡

†Department of Micro System Technology, Ritsumeikan University,
1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, JAPAN
‡Department of Health and Social Services, National Institute of Public Health,
2-3-6 Minami, Wako, Saitama 351-0197, JAPAN
Email: tmiyano@se.ritsumei.ac.jp, tsutsui@niph.go.jp

Abstract—We have devised a method for hypothesis testing for the major features of multivariate data on the basis of collective synchronization in a network of non-symmetrically coupled phase oscillators subject to a variant of Kuramoto's dynamics. We show through numerical experiments that the non-symmetrical coupling allows testing whether given test vectors match the major features.

1. Introduction

Finding patterns that represent the major features of data is one of the most interesting applications of large-scale databases growing as pivotal information infrastructures of the society. Such social needs have accelerated the development of mathematical methods for data mining [1]–[4]. Recently, we have devised a method for feature extraction from multivariate data on the basis of collective synchronization in a network of coupled phase oscillators [5]–[8]. This method was termed data synchronization. In data synchronization, we use a network of coupled phase oscillators subject to a variant of Kuramoto's model [9]–[10]. The phase oscillators carry multivariate data in their natural frequencies and update their rhythms through nonlinear couplings between phase oscillators. Consequently, partial synchronizations of the oscillators are achieved, whose common frequencies are interpreted as the general features of the data set.

An advantage of data synchronization is that it requires no prior information about the feature patterns to be extracted, unlike the self-organizing map (SOM) algorithm, devised by Kohonen, as a popular method for feature extraction [1, 2]. Recently, data synchronization has been shown to be equivalent to SOM near synchronous states where the equations governing data synchronization can be linearized, in the sense that the linearized equations become equivalent to the competitive learning rule for SOM [6, 8]. It can be said that the reference vectors are spontaneously generated during the nonlinear regime of the dynamics in data synchronization. This fact enables us to apply data synchronization to databases of large-scale to which SOM cannot be applied because of lack of prior knowledge for the reference vectors. However, a question was raised to our previous study [8] in that SOM associated with statistical analysis such as principal component analysis for determining appropriate reference vectors, for instance, as has been shown in [11], may suffice. This question is the motivation of this study.

In this paper, we show that prior knowledge as well as hypothesis on the general features of data can be handled and tested by a network of non-symmetrically coupled phase oscillators. Test vectors are encoded into the natural frequencies of particular phase oscillators that are termed "stubborn oscillators". The stubborn oscillators do not change their rhythms due to a null coupling constant with other oscillators, whereas non-stubborn oscillators are forced to synchronize with the stubborn oscillators through a large coupling constant. If the stubborn oscillators recruit many oscillators into major synchronous groups, then the corresponding test vectors may be interpreted as capturing the general features of the data. This paper is organized as follows. In section 2, we describe the theory of our method. In section 3, we show numerical experiments for hypothesis testing of feature patterns for numerical data with three degrees of freedom. The low dimensionality of the data is taken for visual convenience of results and is not due to the inapplicability of our method to data with high degrees of freedom. In sections 4, we discuss results and make concluding remarks.

2. Theory

We describe the theory of data synchronization in a network of non-symmetrically coupled phase oscillators on the basis of our previous model [5, 6]. Suppose that we are given N sample vectors $\{\boldsymbol{x}_i = (x_{i1}, \cdots, x_{iD})\}_{i=1}^N$ and S test vectors $\{\boldsymbol{y}_j = (y_{j1}, \cdots, y_{jD})\}_{j=1}^S$, both $\{\boldsymbol{x}_i\}_{i=1}^N$ and $\{\boldsymbol{y}_j\}_{j=1}^S$ with D

degrees of freedom. The test vectors $\{\boldsymbol{y}_j\}_{j=1}^{S}$ represent a hypothesis for the general features of the sample data $\{\boldsymbol{x}_i\}_{i=1}^{N}$, which may be acquired using a data preprocessor. We assign $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ to the natural frequencies of the phase oscillators subject to the following equations with $n = 1, \cdots, D$:

$$\begin{aligned}\dot{\theta}_{in} &= x_{in} + \frac{1}{N_i}\left[\sum_{k=1}^{N} K_1 H\left(\tilde{d}_1(i,k)\right)\sin\left(\theta_{kn} - \theta_{in}\right)\right. \\ &\quad \left. + \sum_{j=1}^{S} K_2 H\left(\tilde{d}_2(i,j)\right)\sin\left(\phi_{jn} - \theta_{in}\right)\right], \quad (1) \\ \dot{\phi}_{jn} &= y_{jn}\ , \quad (2)\end{aligned}$$

where $\tilde{d}_1(i,k) = \mid \boldsymbol{x}_i - \boldsymbol{x}_k \mid$ and $\tilde{d}_2(i,j) = \mid \boldsymbol{x}_i - \boldsymbol{y}_j \mid$. The phase oscillators carrying $\boldsymbol{y}_j$ are the stubborn oscillators. K_1 and K_2, $K_1 < K_2$, are a small positive coupling constant between non-stubborn phase oscillators and a large positive coupling constant between a non-stubborn oscillator and a stubborn oscillator, respectively. N_i is the number of neighboring vectors to $\boldsymbol{x}_i$. θ_{in} and ϕ_{jn} are the nth entries of the phase vectors $\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{iD})$ and $\boldsymbol{\phi}_j = (\phi_{j1}, \cdots, \phi_{jD})$ corresponding to $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, respectively. Their initial values are given as random numbers. $\dot{\boldsymbol{\theta}}_i$ represents the updated value of $\boldsymbol{x}_i$ at each instant in the time evolution. The partitioning function H restricts the range of interactions between phase oscillators: $H\left(\tilde{d}\right) = 1$ if $\tilde{d} \leq \tilde{d}_0$ and $H\left(\tilde{d}\right) = 0$ otherwise, where $\tilde{d} = \tilde{d}_1(i,k)$ or $\tilde{d}_2(i,j)$. $\tilde{d}_0 = \alpha \mid \boldsymbol{x}_i \mid$ with a positive constant α, which determines N_i neighboring vectors with which the phase vector $\boldsymbol{\theta}_i$ can interact. The parameter α represents tolerance to discriminate neighbors from non-neighbors having features distinct from those of $\boldsymbol{x}_i$.

In this way, each non-stubborn oscillator conveys the original and updated data via its natural and adaptive rhythms, respectively. In contrast, the stubborn oscillators do not change their rhythms, as is apparent from Eq. (2), but forces the non-stubborn oscillators to synchronize with them through the coupling constant K_2.

Under appropriate settings of K_1, K_2 and α, partial synchronous groups of phase oscillators will be generated [5, 6]. The common frequency vector of each synchronous group is interpreted as the general feature of the synchronous group. In particular, the test vector $\boldsymbol{y}_j$ can be said to match a major feature of the sample data, if the jth stubborn oscillator recruits many oscillators to form a synchronous group. Thus, data synchronization with non-symmetrically coupled phase oscillators achieves hypothesis testing for the test vectors $\{\boldsymbol{y}_j\}_{j=1}^{S}$ as candidates of the major features of the sample data.

3. Numerical Experiments

We conducted numerical experiments for data clustering using multivariate data with three degrees of freedom ($D = 3$). In these experiments, we supposed three groups of multivariate data to each of which sixty data vectors should belong, given as $\boldsymbol{x}_i = (1+\epsilon, \epsilon, \epsilon)$, $(\epsilon, 1+\epsilon, \epsilon)$ and $(\epsilon, \epsilon, 1+\epsilon)$ with Gaussian random numbers ϵ of mean 0 and variance 0.1. These groups were represented by the template vectors $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$, respectively. We randomly chosen three test vectors, one vector from each of the three groups. These test vectors were labeled as classes 1, 2 and 3 and were supposed to represent prior knowledge for the general features of the data set. The test vectors were encoded into the natural frequencies of stubborn oscillators.

In the first experiment, we used a network of symmetrically coupled phase oscillators without the stubborn oscillators. The data vectors excluding the test vectors, i.e., 177 data vectors were encoded into the natural frequencies of non-stubborn oscillators. This experiment represents a situation where no prior knowledge is available for the general features of the data set. The coupling constants were set to $K_1 = 0.5$ and $K_2 = 0$. The setting of K_1 was made taking the variance in ϵ into consideration. Results are shown in Fig. 1. Data synchronization generated the correct feature of each group, i.e., $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$, to which all member vectors were entrained.

In the second experiment, we used a network of non-symmetrically coupled phase oscillators with the stubborn oscillators. The multivariate data including the test vectors, i.e., 180 data vectors were encoded into the natural frequencies of the oscillators. The coupling constants were set to $K_1 = 0.5$ and $K_2 = 5.0$. This experiment represents a situation where appropriate prior knowledge is given for the general features of the data set. Results are shown in Fig. 2. As expected, the stubborn oscillators seem to have recruited all members of each group. Thus, the hypothesis that the test vectors of classes 1, 2 and 3 match the major features of the data set can be accepted on the basis of the size of each synchronous group.

Figure 3 presents the comparison of extracted feature vectors between the two experiments, visualizing how the member vectors of each group is attracted to the test vectors. The member vectors of each group have been entrained to the test vectors due to the non-symmetric coupling, which gives rise to a certain degree of bias in data synchronization. To assess the bias, we estimated the mean diversity of frequency vectors over synchronized clusters [5], denoted by σ defined as

$$\sigma = \frac{1}{N+S}\sum_{i=1}^{N+S}\sigma_i$$

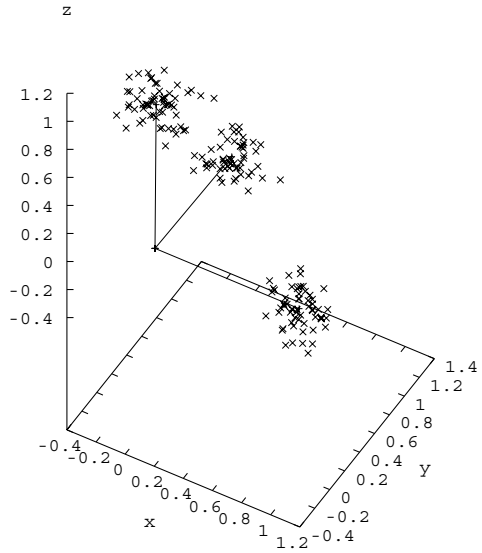

Figure 1: Data clustering and feature extraction for three-dimensional data vectors using symmetrically coupled phase oscillators. The data are shown by × and three extracted feature vectors by + and solid lines.

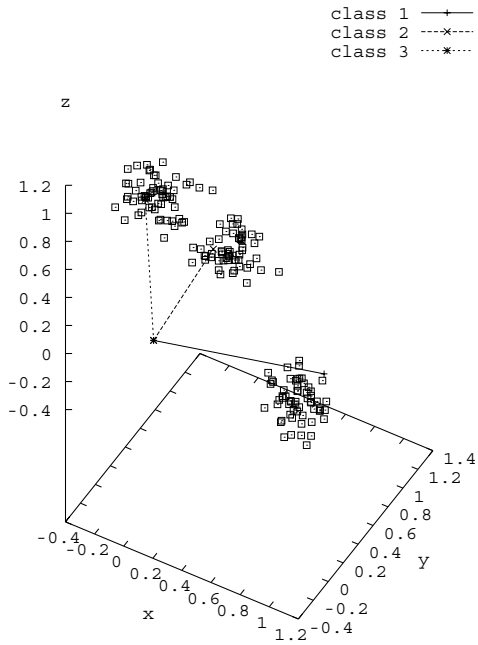

Figure 2: Data clustering and feature extraction for three-dimensional data vectors using non-symmetrically coupled phase oscillators with the stubborn oscillators of classes 1, 2 and 3. The data are shown by squares and three extracted feature vectors by + and solid, dashed and dotted lines.

$$= \frac{1}{N+S} \sum_{i=1}^{N+S} \left(\frac{1}{N_i} \sum_{k=1}^{N+S} H\left(\tilde{d}_{i,k}\right) \frac{d_{i,k}}{\tilde{d}_0} \right) , \quad (3)$$

where $d_{i,k}$ is the distance between neighboring frequency vectors at each instant during the synchronization process. If $N_i = 0$, σ_i is defined to be zero. As perfect synchronization is achieved, $\sigma \to 0$. Estimates of σ for the two experiments are shown in Fig. 4. A much higher degree of synchrony is achieved by data synchronization without the stubborn oscillators. The difference in σ between the two experiments expresses the bias of the test vectors to the true features.

4. Discussion and Conclusion

The present numerical experiments support the validity of our method. When a test vector matches one of the general features of the data set, it will recruit many data vectors into a synchronous cluster. If the test vector matches neither of the general features, it will generate no synchronous cluster or a tiny synchronous cluster. The size of the synchronous cluster can be a measure for evaluating the degree of generality of the test vector. The reliability of hypothesis testing for the test vectors may be related to the mean diversity σ, as shown in Fig. 4.

In the limit of perfect synchronization, the mean diversity $\sigma \to 0$. Accordingly, the deviation of estimated σ from this limit may be used to evaluate the degree of bias of the test vectors to the true feature vectors. As the deviation becomes large, the test vectors become less reliable in representing the general features of the data set. Let us consider a situation such that given two sets of test vectors, we are required to judge which set is more appropriate to represent the general features of the data set. In such a situation, we can make a judge by comparing estimates of σ. Although this is a rough idea to assess the statistical reliability of the test vectors, possible use of σ may be an advantage of data synchronization in that it is difficult, if not impossible, for the SOM algorithm to evaluate the statistical reliability of the outcome of reference vectors.

Establishing an algorithm for assessing the reliability of hypothesis testing on the basis of σ as a function of the coupling constants K_1 and K_2 and applying the present method to real-world data are issues of interest and worth investigating in future studies.

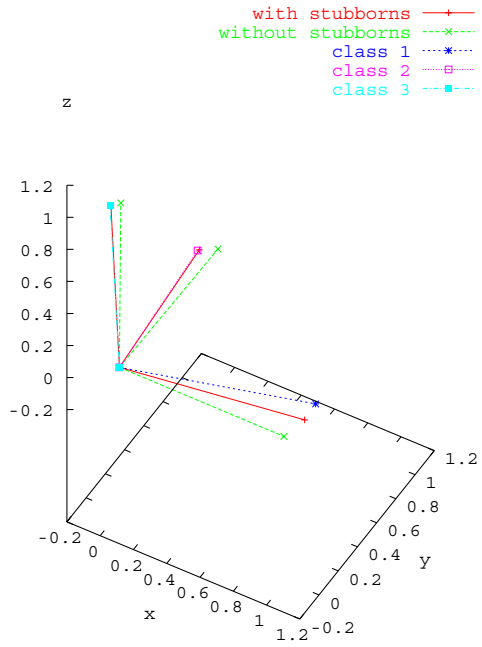

Figure 3: Comparison of extracted feature vectors between the two experiments. Feature vectors extracted using the stubborn oscillators are indicated by red lines and +, those without the stubborn oscillators by green lines and ×. The test vectors are labeled as classes 1, 2 and 3.

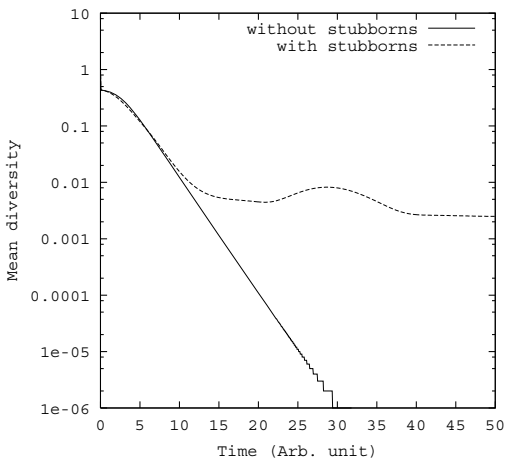

Figure 4: Mean diversity as a function of dimensionless time during the synchronization process. Solid and dashed traces indicate estimated mean diversity for data synchronizations without the stubborn oscillators and with the stubborn oscillators, respectively.

References

[1] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol.78, pp.1464–1480, 1990.

[2] T. Kohonen, *Self-Organizing Maps* (3rd Ed.), Springer-Verlag, Berlin HeidelBerg, 2001.

[3] J. Makhoul, S. Roucos and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol.73, p.1551–1588, 1985.

[4] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, Inc., New York, 1998.

[5] T. Miyano and T. Tsutsui, "Data synchronization in a network of coupled phase oscillators," *Phys. Rev. Lett.*, vol.98, pp.024102-1–024102-4, 2007.

[6] T. Miyano and T. Tsutsui, "Collective synchronization as a method of learning and generalization from sparse data," *Phys. Rev.* E, vol.77, pp.026112-1–026112-11, 2008.

[7] T. Miyano and T. Tsutsui, "Finding major patterns of aging process by data synchronization," *IEICE Trans. Fundamentals*, vol.E91-A, no.9, pp.2514–2519, 2008.

[8] T. Miyano and T. Tsutsui, "Link of data synchronization to self-organizing map algorithm," *IEICE Trans. Fundamentals*, vol.E92-A, no.1, pp.263–269, 2009.

[9] Y. Kuramoto, *Chemical Oscillations, Waves, and Turbulence*, Springer-Verlag, Berlin HeidelBerg, 1984.

[10] J. A. Acebrón, L. L. Bonilla, C. J. P. Vincente, F. Rotort, and R. Spigler, "The Kuramoto model: A simple paradigm for synchronization phenomena," *Rev. Mod. Phys.*, vol.77, pp.137–185, 2005.

[11] S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi and H. Mori, "Analysis of codon usage diversity for bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome," *Gene*, vol.276, pp.89–99, 2003.