2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Robust Prediction in Nearly Periodic Time Series using Motifs

Woon Huei Chai, Hongliang Guo, and Shen-Shyang Ho

***Abstract*— In this paper, we consider the prediction task for a process with nearly periodic property, i.e., patterns occur with some regularities but no exact periodicity. We propose an inference approach based on probabilistic Markov framework utilizing motif-driven transition probabilities for sequential prediction. In particular, a Markov-based weighting framework utilizing fully the information from recent historical data and sequential pattern regularities is developed for nearly periodic time series prediction. Preliminary experimental results show that our prediction approach is competitive against the moving average and multi-layer perceptron neural network approaches on synthetic data. Moreover, our proposed method is shown to be empirically robust on time-series with missing data and noise. We also demonstrate the usefulness of our proposed approach on a real-world vehicle parking lot availability prediction task.**

I. INTRODUCTION

Nearly periodic time series includes almost and quasi-periodic time series which have gradually changing periods. Almost periodic time series have fixed expectation of fundamental frequency along with time. They appear in musical signal [15], the strength of the global oceanic tide-raising forces [16], communication signals [9], [10], and financial data [27]. The fundamental period of a quasi-periodic time series is almost constant in short period but it changes gradually in long period. Such time series appear in astronomical near-infrared and x-ray oscillations [5], financial and economical indices [3], [25], seismic activities [14], and wind intensity measurements [7]. A seasonal time series could either have characteristics of an almost periodic or a quasi-periodic time series. They have additional drifting terms in their series. Seasonal time series are found in consumer behavior data [22] and financial data [4], [31]. Short term behaviors, whose underlying variables change frequently and has long term unpredictability, exist in some almost periodic highly dynamic systems and phenomena. For example, the dynamic behavior of flame induces by radiative heat loss [11]. Traditional predictive models whose accuracies depend on the historical training data size are well designed for forecasting time series. These models use either local recent information, such as linear regression [26], moving average [13] and exponential weighted moving average [8] or use periodic information, such as Fourier analysis, wavelet transformation [28], and Chirplet transformation [20]. However, they are not suitable for nearly periodic time series which has near-term (local) rapid behavior changes and recent periodic information. Consequently, using only either periodic information or local change information is not sufficient for robust prediction.

In this paper, we present a novel motif-driven approach to estimate transition probabilities for high order Markov chain for robust prediction in a nearly periodic time series such that its period T_i is a random variable with a finite expected value, i.e., $E(T_i) = T$ such that $T \in R^+$ with some sequential regularities. The proposed prediction model assumes that a time period from the current time to the prediction time contains patterns or regularities for prediction. Preliminary experimental results on synthetic data and real world data show that the performance of our proposed model is competitive against the moving average and multi-layer perceptron neural network prediction models. Moreover, our proposed method is shown to be empirically robust on time-series with missing data and noise.

Woon Huei Chai, Hongliang Guo and Shen-shyang Ho are with the School of Computer Engineering, Nanyang Technological University, Singapore (email: {chai0093, guohl, ssho}@ntu.edu.sg).

II. BACKGROUND

A. Markov Model

Markov chains exist in many dynamic systems including nearly periodic multi-class time series with noise [23]. The time series could have constant transition probabilities, stochastic transition probabilities or a combination of both. In this paper, a Markov model is represented by $(\{V_c, D\}, V_f, Q)$, where V_c is the current state at time c, D is the set of historical states in the past, V_f is the forecasted state at time $f = c + \Delta t$, Q is the transition matrix from the corresponding current state V_c and historical states D to the random future state V_f. $P_1 = P(V_f|V_c, D)$ is the transition probability in transition matrix.

B. Time Series Range Motif

A *time series range motif* in a given time series, T, with length m is a set of all time series with length much shorter than m, which share high similarity with each other and located at significant different position in T [19]. Formally, a time series motif and range motif [21] are defined as follows

Definition II.1. The *time series motif* in a time series, $\mathscr{T}$, is the most similar pair, $\mathscr{T}_i$ and $\mathscr{T}_j$, of time series among the other pairs in the time series $\mathscr{T}$ such that $\forall a,\ b,\ i,\ j,\ dist(\mathscr{T}_i, \mathscr{T}_j) \leq dist(\mathscr{T}_a, \mathscr{T}_b),\ i \neq j, a \neq b$ while $\mathscr{T}_i$ and $\mathscr{T}_j$ are not time series with slightly shift from each other as well as $\mathscr{T}_a$ and $\mathscr{T}_b$.

Definition II.2. The *range motif* set $\mathscr{T}^r = \{\mathscr{T}_1, \mathscr{T}_2, ...\}$ containing all time series which have distance less than $2r$ between any two time series such that $\forall \mathscr{T}_i,\ \mathscr{T}_j \in \mathscr{T}^r$, $dist(\mathscr{T}_i, \mathscr{T}_j) \le 2r$, $i \ne j$ and $\forall \mathscr{T}_k \in \mathscr{T} - \mathscr{T}^r$, $dist(\mathscr{T}_i, \mathscr{T}_k) > 2r$.

Intuitively, a nearly periodic time series contains range motifs at almost regular time intervals. Towards this end, we utilize the concept of range motifs to characterize local variations and regularities in time series to estimate transition probabilities for a Markov-based prediction model [6], [29]. Let D_c^k and D_f^k represent the two sets of states in the neighborhoods of V_c and V_f at the k^{th} previous period as follows:

$$\begin{aligned} D_c^k &= \{V_{c-k\hat{T}-w}, V_{c-k\hat{T}-w+1}, ..., V_{c-k\hat{T}}, ..., V_{c-k\hat{T}+w-1}, \\ &\quad V_{c-k\hat{T}+w}\} \\ D_f^k &= \{V_{f-k\hat{T}-w}, V_{f-k\hat{T}-w+1}, ..., V_{f-k\hat{T}}, \\ &\quad ..., V_{f-k\hat{T}+w-1}, V_{f-k\hat{T}+w}\} \end{aligned}$$

where V_t is the state at time t, and w controls the set size such that each set consists of $(2w+1)$ states. The two sets are assumed to be range motifs.

In Figure 1, we demonstrate the above definitions on the current state, V_c, the future state of interest V_f, time interval between V_c and V_f, $\Delta t = f - c$, and the neighborhood sets D_c^k and D_f^k where $k = 1$ and 2 using a sinusoidal function. Information from $l = 2$ historical periods are used to estimate the unknown state V_f.

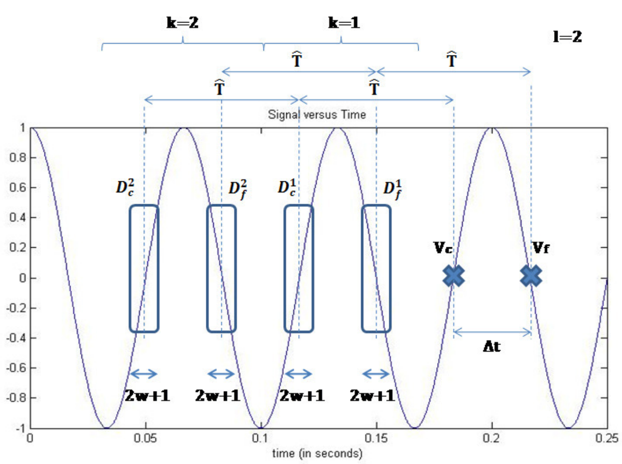

Fig. 1. A simple example to illustrate our definitions when $l = 2$ historical periods are used in the predictive model for estimating the future state of interest, V_f, in a sinusoidal function

The fundamental period of time series can be either estimated from the frequency domain (e.g., fast Fourier transform) or the time domain (e.g., autocorrelation). The estimated fundamental period

$$\hat{T} = \frac{1}{2\pi\hat{f}} \tag{1}$$

where $\hat{f}$ is the estimated fundamental frequency is obtained from the peak in the fast Fourier transform of the time series. It can also be estimated from the interval between peaks of autocorrelation.

III. Methodology

In our model, the data generating process of the time series is assumed to follow a periodic-driven homogeneous Markov chain. The model is a N-step Markov chain model which predicts the state at Δt, $N * t'$, ahead where N can be any positive integer which represents the N^{th} state ahead from current state, V_c, and t' is the data sampling granularity. Markov chain model predicts future state of interest with large N without predicting the unknown states between the known steps and state of interest. Hence, it avoids the significant accumulated errors from other one-step predictive models (e.g, autoregression and ARIMA). In order to solve the lack of information because of local change, the prediction depends on a set of states

$$D = D_c^1 \cup D_f^1 \cup ... \cup D_c^k \cup D_f^k ... \cup D_c^l \cup D_f^l$$

which is a union set of the range motif D_c^1, D_f^1, ..., D_c^l and D_f^l.

Our model estimates transition probability in two ways for the two cases: (1). $V_c = a$; $a \in D_c^k$; (2). $V_c = a$; $a \notin D_c^k$, as follows.

Case 1: $V_c = a$; $a \in D_c^k$. For the case when current state V_c appears in range motifs D_c^k, the transition probability

$$\begin{aligned} P(V_f = v | V_c, D) &= \frac{1}{Z}[\varphi(v, V_c, D_c^1, D_f^1) \\ &\quad + ... \\ &\quad + \varphi(v, V_c, D_c^k, D_f^k) \\ &\quad + ... \\ &\quad + \varphi(v, V_c, D_c^l, D_f^l)] \end{aligned} \tag{2}$$

where v is any possible future state of interest and φ is the potential function containing the information of characteristics of each range motif pair, D_c^k and D_f^k. To form the prediction model, the information of three characteristics: (1) periodicity of current, $c - k\hat{T}$; (2) forecasted states, $f - k\hat{T}$, and (3) the time interval of pair of states from each range motif pair, D_c^k and D_f^k are used to estimate the transition probability. Each transition probability composes of a summation of multiple potential functions, $\varphi(v, V_c, D_c^k, D_f^k)$, from information in each range motif pairs, D_c^k and D_f^k divided by a normalized denominator, Z.

The function φ is defined as

$$\varphi(v,V_c,D_c^k,D_f^k) = \sum_{a=-w}^{w}\sum_{b=-w}^{w} \delta(V_{c+a}=V_c)\,\delta(V_{f+b}=v)\,W(a,b) \tag{3}$$

where $\delta(expression)$ is an indicator function which has only two values, 1 or 0, corresponding to the "True" and "False" of the expression in the bracket of the function,

$$W(a,b)=\rho_1(a)\rho_2(b)\rho_3(b-a) \tag{4}$$

and functions ρ_m, $m=1,2,3$ are bi-linear functions

$$\rho_m(x)=\begin{cases} \frac{w+1-|x|}{w^2+2w+1} & |x|\leq w \\ 0 & otherwise. \end{cases} \tag{5}$$

where x determines the value of the ρ_m function and w is model dependent controlling the gradients of the two lines in the function. The highest value of the function is at $x=0$. The higher the absolute value of x, the lower value of ρ_m. The sum of the ρ_m for all x is equal to 1 as it acts as a probability in the weight function W. Figure 2 shows an example of the bi-linear function ρ_m with $w=1$.

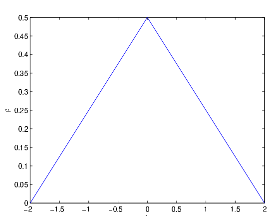

Fig. 2. bi-linear weighting function where w=1

Each potential function $\varphi(v,V_c,D_c^k,D_f^k)$ includes the information from range motifs, D_c^k and D_f^k. These information are from a pair of two states equal to V_c and v in D_c^k and D_f^k, respectively. Each data point is assigned a weight from W, the product of the values of ρ_1, ρ_2, and ρ_3. The sum of all weighted pair data is the value of the potential function $\varphi(v,V_c,D_c^k,D_f^k)$ in transition probability, $P(v|V_c,D)$. Figure 3 is an example which shows the weight function, W, of all pairs of state in historical periods, $k=1$, with $w=1$.

The values of ρ_1, ρ_2 and ρ_3 are related to the three characteristics of range motifs used in the prediction. These relationships are described in the following observations in nearly periodic time series.

Observation III.1. *Due to the near periodicity at current state V_c, any state, $V_{c-k\hat{T}+a}$, in motif D_c^k near to the V_c tends to repeat at V_c.*

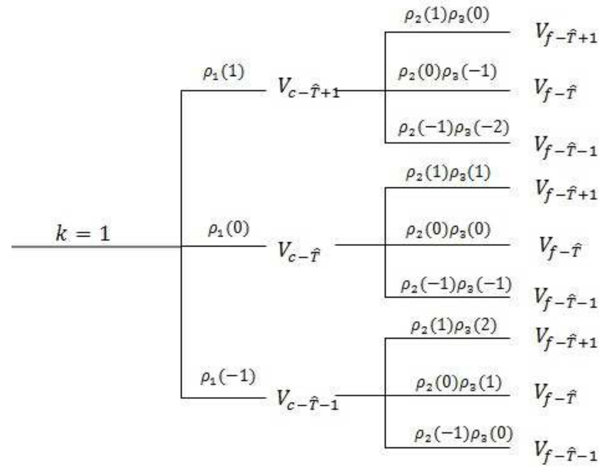

Fig. 3. The weight functions, W, for all pairs of state in historical periods, $k=1$, with $w=1$

The likelihood for the repetition of the state, $V_{c-k\hat{T}+a}$, at V_c, causes the weight of information of pair of states which the data in D_c^k is near to the time, $c-k\hat{T}$, to be higher. It is reflected in the bi-linear function, ρ_1, in Eq. (5) where a is the time difference between the time of $V_{c-k\hat{T}+a}$ and $c-k\hat{T}$.

Observation III.2. *Due to the near periodicity at future state V_f, the nearer the state, $V_{f-k\hat{T}+b}$, in motif D_f^k to V_f, the higher the likelihood the state repeats itself at, V_f.*

Similar to the bi-linear function, ρ_1, the likelihood of the occurrence of $V_{f-k\hat{T}+b}$ at V_f, is reflected in the value of ρ_2 which is the same bi-linear function as ρ_1 in Eq. (5) where b is the time difference between the time of $V_{f-k\hat{T}+b}$ and $f-k\hat{T}$. Hence, the nearer the state in D_f^k in the pair of states to the $f-k\hat{T}$, the higher the value of ρ_2 in the weight function of the information.

Observation III.3. *Due to near periodicity in the time series, any pair of states with a time interval in previous periods is likely to repeat themselves in current and future periods with the same time interval.*

There is a likelihood of repetition of previous pair of states in the current and future periods. Therefore, the pair of states which are Δt apart tend to repeat themselves at V_c and V_f. The likelihood of this repetition is reflected in a bi-linear function, ρ_3, in Eq. (4) where $b-a$ is the difference between the time interval of $V_{c-k\hat{T}+a}$ and $V_{f-k\hat{T}+b}$ and the Δt, the time interval between V_c and V_f. Hence, the nearer the time interval of previous pair of states to the Δt, the higher the tendency of the repetition of the pair of states in current and future periods as well as the higher the value of ρ_3 in the weight function of the pair of states.

Through the estimation of all W and φ, we are able to compute all numerators in the transition probabilities function, $P(v|V_c,D)$, in Eq. (2).

Without loss of generality, we assume the state values range from 1 to m. The normalized term, Z, in Eq. (2), is computed as follows.

$$Z = \sum_{v=1}^{m}\sum_{k=1}^{l}\sum_{a=-w}^{w}\sum_{b=-w}^{w}\delta(V_{c-k\hat{T}+a}=V_c) \delta(V_{f-k\hat{T}+b}=v)\rho(a)\rho(b)\rho(b-a). \quad (6)$$

It is the sum of all numerators of all possible future state, v, with respect to the current state, V_c in the transition probability.

Case 2: $V_c = a;\ a \notin D_c^k$. In some cases, there are some particular states V_c that do not occur in any of its range motif in previous periods from the current state, D_c^k. As a result, there is no pair of states in range motifs between D_c^k and D_f^k to estimate the transition probability, $P(V_f = v|V_c,D)$ in Eq. (2) for that particular V_c. In order to estimate the transition probability, we apply the transition distribution model in Eq. (7) to obtain

$$P(V_f = v|V_c,D) = \frac{1}{Z}\sum_{\overline{V}_c\in D_c}\rho(\overline{V}_c - V_c)P_{case1}(v|\overline{V}_c,D) \quad (7)$$

where $P_{case1}(v|\overline{V}_c,D)$ is the transition probability from Eq. (2), and $\overline{V}_c$ is a pseudo current state belonging to

$$D_c = D_c^1 \cup D_c^2 \cup ... \cup D_c^l.$$

The transition probability is the sum of weighted transition probabilities estimated from Eq. (2) by replacing the V_c in the equation by a pseudo current state, $\overline{V}_c \in D_c$. The weighting function, $\rho(\overline{V}_c - V_c)$, of each transition probability with pseudo current state depends on a bi-linear function ρ in Eq. (5) with $w = max(m - V_c, V_c - 1)$. w determines the highest value between the difference of the maximal state and the current state and the difference of the current state and the minimal state so that the bi-linear function, ρ, is positive for all the states. It depends on the similarity of the transition probability $P(v|\overline{V}_c,D)$ of the pseudo current state with that for the current state, V_c.

The normalized term

$$Z = \sum_{v=1}^{m}\sum_{\overline{V}_c\in D_c}\frac{1}{Z_1(\overline{V}_c)}\rho(\overline{V}_c - V_c)P_{case1}(v|\overline{V}_c,D)$$

is the sum of all weighted transition probability in Eq. (7) with all possible (m) future states of interest.

Prediction: The estimated future state of interest

$$\hat{V}_f = E(V_f) = \sum_{v=1}^{m} vP(v|D) \quad (8)$$

is the expectation of the possible future states.

IV. Algorithm

Algorithm 1 shows the procedure to predict the future state of interest. In Line 1, the algorithm computes the transition probability t_3, assuming the time series belongs to case 1 with Algorithm 2. It returns a variable c to show which case of the time series belongs to and a non empty transition probability if the time series belongs to case 1. Line 2 checks whether the data belongs to case 1 or case 2 (see Section III) from the value of c return from in Line 1. If the data belongs to case 1, the algorithm jumps to Line 27 to evaluate the estimated state of interest, $\hat{V}_f$, with the transition probability from Line 1. If the data belongs to case 2, Line 3 to 26 computes the unnormalized transition probabilities, t_2, for the prediction. In Line 16, t_1 stores the unnormalized transition probabilities of case 1 with assigned the pseudo current states with states other than actual state into the Eq. (2). These unnormalized transition probabilities of the pseudo current states are used for the estimation of the unnormalized transition probabilities, t_2 for the data of case 2. Line 3 to 13 computes the bi-linear function, ρ_b, as the ρ function in Eq. (7). Line 14 to 22 computes the t_2 and normalized terms, Z, of all transition probabilities as in Eq. (7 and 8). Line 23 to 25 determines the normalized transition probabilities, t_3. Line 27 to 29 performs the prediction of future state of interest by computing the estimated future state as in Eq. (8).

Algorithm 2 computes the transition probabilities, t_3, for the time series of case 1 and c, a variable with values, 1 or 2, indicates the case the time series belongs to. Line 2 initiates variable $c = 2$. From Line 2 to 4, it computes the function ρ_a (the function ρ in Eq. (5)) whose values depend on the value of input variable w. The computed values are used as ρ_1, ρ_2 and ρ_3, the weight function in Eq. (4) in Line 10. Line 7 checks whether V_c exists in D_c^k and Line 8 assigns $c = 1$ if it exists. Line 10 computes the unnormalized transition probability, t_2, as in Eq. (2) for time series of case 1. Line 15 to 17 calculates the normalized term, Z, of the transition probabilities as in Eq. (6). Line 18 to 20 determines the transition probabilities, t_3 of the time series of case 1. Line 21 returns the variable c and transition probability, t_3, of the procedure.

V. Experimental Results

In this section, we compare the prediction performance of our proposed methods with moving average and multi-layer perceptron network models. The experiments are performed on periodic and nearly periodic synthetic data under the condition of missing data or noisy data and real datasets on parking availability.

Simple moving average [13] is a statistical technique which takes a series of unweighted averages of continuous data. It is a low-pass filter to smooth continuous

Input: V_t, historical time series; c, time of current state; m, total number of states; $\hat{T}$, estimated fundamental period; w, the constant control the information size; l, the periods of historical states to be used and N, the number of state ahead from current state to be forecasted
Output: $\hat{V}_f$, forecasted future state

Global Variables: t_1, t_2, t_3, unnormalized transition probability vectors conditioned to the current class for data of case 1, the final unnormalized and normalized transition probabilities for the prediction; n_1, n_2, numerator of transition probability for case 1 and case 2, respectively; Z_1, Z_2, normalized term of transition probability for case 1 and case 2, respectively; ρ_a, bi-linear function for either $\rho 1$, $\rho 2$ or $\rho 3$ in Weight function, Eq. (4); ρ_b, bi-linear function ρ in Eq. (7)

Procedure Prediction

$[e, t_3]$ = ComputeTP(V_t, $\hat{V}_c$, c, w, l, $\hat{T}$, N)
if $e = 2$ **then**
  $L_1 = m - V_c$
  $L_2 = V_c - 1$
  if $L_1 > L_2$ **then**
    for $r_2 = -L_1$ to L_1 **do**
      $\rho_b = \frac{L_1 + 1 - |r_2|}{L_1^2 + 2L_1 + 1}$
    end for
  else
    for $r_2 = -L_2$ to L_2 **do**
      $\rho_b = \frac{L_2 + 1 - |r_2|}{L_2^2 + 2L_2 + 1}$
    end for
  end if
  for $\hat{V}_c = 1$ to m **do**
    if $\hat{V}_c \neq V_c$ **then**
      $[x, t_1]$ = ComputeTP(V_t, $\hat{V}_c$, c, w, l, $\hat{T}$, N)
      for $d_3 = 1$ to m **do**
        $t_2(d_3) = t_2(d_3) + \rho_b(\hat{V}_c - V_c) t_1(d_3)$
        $Z = Z + \rho_b(\hat{V}_c - V_c) t_1(d_3)$
      end for
    end if
  end for
  for $d_4 = 1$ to m **do**
    $t_3(d_4) = \frac{1}{Z} t_3(d_4)$
  end for
end if
for $d_5 = 1$ to m **do**
  $\hat{V}_f = \hat{V}_f + d_5 t_3(d_5)$
end for

Algorithm 1: Prediction

Procedure ComputeTP(V_t, V_c, c, w, l, $\hat{T}$, N)

$e = 2$
for $r_1 = -w$ to w **do**
  $\rho_a = \frac{w + 1 - |r_1|}{w^2 + 2w + 1}$
end for
for $k = 1$ to l **do**
  for $a = -w$ to w **do**
    if $V_{c - k\hat{T} + a} = V_c$ **then**
      $e = 1$
      for $b = -w$ to w **do**
        $t_2(V_{f - k\hat{T} + b}) = t_2(V_{f - k\hat{T} + b}) + \rho_a(a)\rho_a(b)\rho_a(b - a)$
      end for
    end if
  end for
end for
for $d_1 = 1$ to m **do**
  $Z = Z + t_2(d_1)$
end for
for $d_2 = 1$ to m **do**
  $t_3(d_2) = \frac{1}{Z} t_2(d_2)$
end for
return e, t_3

Algorithm 2: Compute Transition Probability

data. Moving average of n consecutive data point is defined as follow,

$$SMA_n(t) = \frac{V_{t - \frac{n-1}{2}} + V_{t - \frac{n-1}{2} + 1} + \ldots + V_{t + \frac{n-1}{2}}}{n} \quad (9)$$

where $V_{t - \frac{n-1}{2}}$, $V_{t - \frac{n-1}{2} + 1}$, ... , $V_{t + \frac{n-1}{2}}$ are n subsequence data which midpoint is V_t In our experiments, moving average predictive models take the mean of l periods of $SMA_n(f - k\hat{T})$ as the predicted states, $\hat{V}_f$ as follows.

$$\hat{V}_f = \frac{1}{l}\left(\sum_{k=1}^{l} SMA_n(f - k\hat{T})\right) \quad (10)$$

where l is the total periods of historical data to train the predictive model.

A multi-layer perceptron network is a feedforward neural network with multiple hidden layers. In our experiments, multi-layer perceptron networks with three hidden layers of 10 nodes are used for the prediction. The input values are the n subsequence data $V_{f - k\hat{T} - \frac{n-1}{2}}$, $V_{f - k\hat{T} - \frac{n-1}{2} + 1}$, ... , $V_{f - k\hat{T} + \frac{n-1}{2}}$ of different periods of k and the output value is the predicted state, V_f. Levenberg-Marquardt backpropagation [12] is used to train the weights of the networks.

Mean squared error is used as the performance evaluation criterion. It measures the amount of difference between the actual and the predicted states of interest as follows.

$$\varepsilon_{mse} = \frac{1}{t_f - t_i + 1} \sum_{f = t_i}^{t_f} \left(\hat{V}_f - V_f\right)^2 \quad (11)$$

where t_i and t_f are the initial and the final time stamps of the testing time series.

A. Synthetic Data Generation

We generate synthetic data with either random noise or missing data. These time series are either periodic or nearly periodic. All synthetic data are five-day time series consisting of two states: 0 and 1. The (expected) period of all data is 24 hours and the data is generated with sampling rate of 5 minutes. The generation of the synthetic data is similar to the synthetic data generation in [18].

1) Periodic Synthetic Data Generation: A one-day time series is generated from the Bernoulli distribution. For simplicity, the success probability is set to 0.5. The state is set to 1 corresponding to the random variable with success outcome and otherwise it is set to 0. The one-day time series is replicated 5 times to form a five-day periodic time series.

The random noise is induced in the time series based on the Bernoulli distribution. Each data point in the five-day time series corresponds to a random variable. The random variable with a (new) success outcome causes its corresponding data point to be flipped to the opposite state.

To generate synthetic data with missing data, the process is similar to the noise generation above. The random variables with success outcome cause the data points of the five-day periodic time series to be unknown (or missing data).

2) Nearly Periodic Synthetic Data Generation: The initial step of the generation of nearly periodic synthetic data is the same as the first step in the generation of periodic synthetic data. A ten-day periodic time-series is generated in the initial step. The second step is to remove and add additional data points into the original time series. Two sets of random variables of nine days were generated with Bernoulli distribution. The generated random variables in one of the two sets determines which corresponding data point in the original time series to be removed and the other set determines which data point in the original time series to be added a data point after it. Each data point of each set corresponds to a data point of the first nine days of the ten-day periodic time series. The removing and adding of a data point will only occur when one data point from the two sets of a particular data point in the original time series is success. In our synthetic data the success probability of both removing and adding data is 0.5. The third step is to add random noise and creating missing data. The procedures are the same as those for the periodic time series. The final step is to truncate the last five days data points from the time series to form a five-day nearly periodic time series.

B. Vehicle Parking Lot Availability Dataset

Fifteen set of twenty-nine consecutive days parking lot availability dataset are acquired from Singapore Land Transport Authority [1] in 2013. The data are the number of parking lot available in 15 commercial parking garages in Singapore. The data were recorded with a sampling rates of 5 minutes. In the experiments, we assume the (expected) periodicity of the time series is 7 days. We transform the parking lot availability time series into a time series containing two states: zero and one. Zero represents parking availability of less than 5 percent and one represents otherwise.

C. Results Comparison

First, we discuss the prediction results using the synthetic data. In the performance evaluation, Δt is the time interval between the current time, c, and prediction time, f(see Section II-A). The data from the second day to fourth day are the training data and the data of the fifth day is the testing data.

In each experiment, we generate five random time series with the same parameter setting. The performance of each prediction model is the mean of its performance on the five time series. In our experiments, we compare the results of our proposed method with $w=0, l=1$ and $w=1, l=1$(PM(0,1) and PM(1,1)) moving average models with $w=0, l=1$ and $w=1, l=1$(MA(0,1) and MA(1,1)) and neural network models with $n=1, l=1$ and $n=3, l=1$(NN(1,1) and NN(3,1)). Δt is one hour in all experiments. All missing data are ignored in the prediction of proposed method and moving average models unless all data to be used in constructing the model are missing. If the situation happens, the model will generate a random guess of predicted state between state 0 and 1 by 1 trial of binomial distribution with the success probability of 0.5. For neural network models, all missing data are replaced by binomial random variable with 1 trial and success probability of 0.5 too.

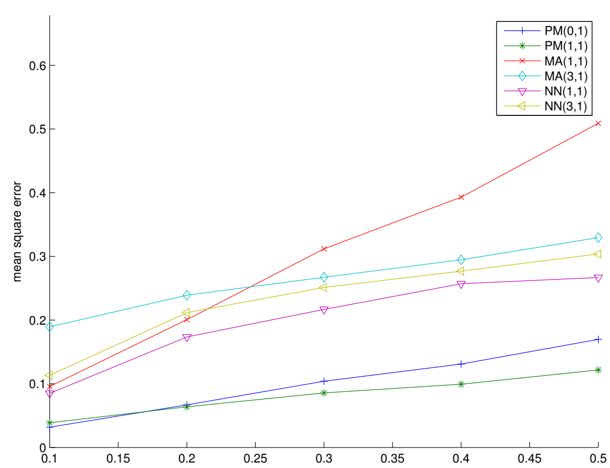

Fig. 4. Comparison of mean squared errors in periodic synthetic data with various noise probabilities.

[1] http://www.mytransport.sg/content/mytransport/home.html

Fig. 4 shows the comparison results of all models in predicting periodic time series with noise probabilities ranging from 0.1 to 0.5. We observe that the best predictive models are PM(0,1) and PM(1,1). They have mean squared errors less than 0.17. Other than time series with noise probabilities less than 0.20, the performance of PM(1,1) is better than PM(0,1). The two feedforward neural network models have significantly higher mean squared errors than the two proposed methods. Their mean squared errors are more than two times of PM(1,1). MA(3,1) has higher errors than all four models mentioned previously. The mean squared errors of MA(1,1) is nearly equal to the noise probabilities. It implies that the erroneous probabilities of the predictive state depend on noise probabilities. It is equal to the noise probabilities as the predictive model only depends on only one data in previous period.

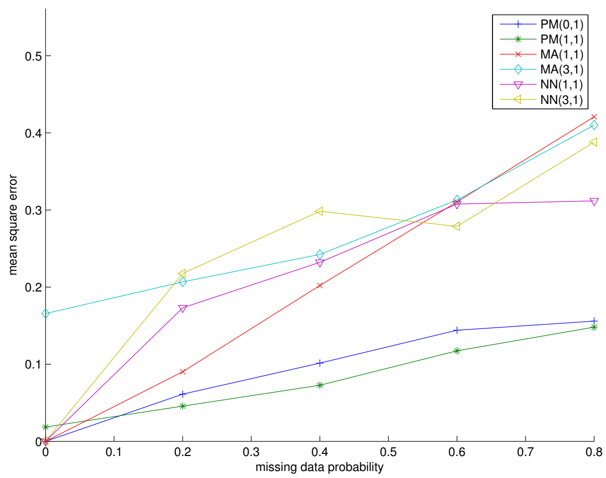

Fig. 5. Comparison of mean squared errors in periodic synthetic data with various missing data probabilities.

Fig. 5 shows the comparison results on periodic time series with missing data probabilities ranging from 0 to 0.8. Similarly, we observe the two best predictive models are PM(0,1) and PM(1,1). Other than time series with noise probabilities less than 0.10, the performance of PM(1,1) is better than PM(0,1). PM(0,1), MA(0,1) and NN(3,1) have zero or nearly zero error rate in the prediction of data with 0 missing data probability. However their mean squared errors increase significantly with the increase of missing data probabilities except PM(0,1). The performance of neural network models change slightly down trend and moving average models have poorer performance in all predictions. The errors of MA(1,1) are similar to its performance in prediction with noise corrupted data. It is linearly dependent on the missing data probability.

Fig. 6 shows the comparison results on nearly periodic time series with noise. Interestingly, performances of all models do not depend on the noise probabilities. All models have nearly constant mean squared errors

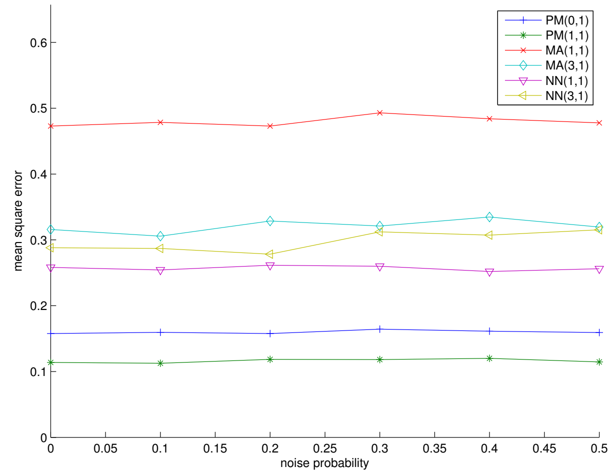

Fig. 6. Comparison of mean squared errors in nearly periodic synthetic data with various noise probabilities.

in all predictions. PM(1,1) has the best performance with mean squared errors less than 0.12, when PM(0,1) has the second best performance with mean squared errors around 0.16. MA(3,1) and the two neural network models have higher mean squared errors around 0.3. The model has the worst performance is MA(1,1) with mean squared errors slightly lower than 0.5 of the mean squared error of random guess.

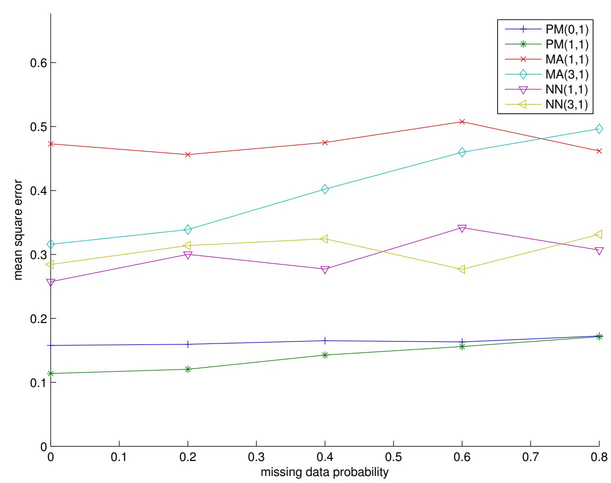

Fig. 7. Comparison of mean square errors in nearly periodic synthetic data with various missing data probabilities.

Fig. 7 shows the comparison results on nearly periodic time series with missing data. The best predictive models are PM(0,1) and PM(1,1). PM(1,1) has significantly lower mean square errors than PM(0,1). The gap of the differences between their mean squared errors approach zero with increase missing data probabilities. The performance of PM(0,1), NN(1,1), NN(3,1) and MA(1,1) do not change significantly with all missing data probabilities while the other models perform poorer with higher missing data probabilities. Again, MA(1,1) has mean square errors slightly lower than 0.5. MA(3,1) has mean square errors around 0.5 at missing data probability of 0.8.

For the vehicle parking lot availability prediction task, Δts are 5, 10, 15, and 20 minutes. From Table I, one observes that PM(1,1) has the best performance.

TABLE I
THE MEAN SQUARED ERRORS OF ALL MODELS IN PREDICTING THE STATES AFTER 5, 10, 15 AND 20 MINUTES FROM THE TIME PERFORMING THE PREDICTION

Δt (minutes)	proposed method ($w=0$, $l=1$)	proposed method ($w=1$, $l=1$)	Moving Average ($n=1$, $l=1$)	Moving Average ($n=3$, $l=1$)	Neural Network ($n=1$, $l=1$)	Neural Network ($n=3$, $l=1$)
5	0.0193	0.0187	0.0578	0.0558	0.0451	0.0465
10	0.0193	0.0188	0.0578	0.0558	0.0447	0.0452
15	0.0193	0.0188	0.0578	0.0558	0.0433	0.0453
20	0.0193	0.0188	0.0578	0.0558	0.0431	0.0461

VI. CONCLUSION AND FUTURE WORKS

In this paper, we propose an inference approach based on probabilistic Markov framework utilizing motif-driven transition probabilities for sequential prediction. In particular, a Markov-based weighting framework utilizing fully the information from recent historical data and sequential pattern regularities is developed for nearly periodic time series prediction. Preliminary experimental results show that our prediction approach is competitive against the moving average and multi-layer perceptron neural network approaches on synthetic data. Moreover, our proposed method is shown to be empirically robust on time-series with missing data and noise. We also demonstrate the usefulness of our proposed approach on a real-world vehicle parking lot availability prediction task.

In future, we will apply our proposed approach on data with multiple states. We will also use motifs from more than one time period to train our proposed method.

REFERENCES

[1] Alspector, J., Gupta, B., & Allen, R. B. (1990, January). Performance of a stochastic learning microchip. In Artificial neural networks (pp. 66-78). IEEE Press.

[2] Bloomfield, P. (2004). Fourier analysis of time series: an introduction. Wiley. com.

[3] Bodyanskiy, Y., & Popov, S. (2006). Neural network approach to forecasting of quasiperiodic financial time series. European Journal of Operational Research, 175(3), 1357-1366.

[4] Broszkiewicz-Suwaj, E., Makagon, A., Weron, R., & WyÅĆomaÅĎska, A. (2004). On detecting and modeling periodic correlation in financial data. Physica A: Statistical Mechanics and its Applications, 336(1), 196-205.

[5] Dolence, J. C., Gammie, C. F., Shiokawa, H., & Noble, S. C. (2012). Near-infrared and X-Ray Quasi-periodic Oscillations in Numerical Models of Sgr A*. The Astrophysical Journal Letters, 746(1), L10.

[6] Eiben, A. E., Aarts, E. H., & Van Hee, K. M. (1991). Global convergence of genetic algorithms: A Markov chain analysis. In Parallel problem solving from nature (pp. 3-12). Springer Berlin Heidelberg.

[7] Furuya, M., Hamano, Y., & Naito, I. (1996). Quasi-periodic wind signal as a possible excitation of Chandler wobble. Journal of geophysical research, 101(B11), 25537-25.

[8] Gardner, E. S. (1985). Exponential smoothing: The state of the art. Journal of forecasting, 4(1), 1-28.

[9] Gardner, W. A. (1994). Cyclostationarity in communications and signal processing. STATISTICAL SIGNAL PROCESSING INC YOUNTVILLE CA.

[10] Giacomelli, G., Giudici, M., Balle, S., & Tredicce, J. R. (2000). Experimental evidence of coherence resonance in an optical system. Physical review letters, 84(15), 3298.

[11] Gotoda, H., Ikawa, T., Maki, K., & Miyano, T. (2012). Short-term prediction of dynamical behavior of flame front instability induced by radiative heat loss. Chaos: An Interdisciplinary Journal of Nonlinear Science, 22(3), 033106-033106.

[12] Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. Neural Networks, IEEE Transactions on, 5(6), 989-993.

[13] Johnston, F. R., Boyland, J. E., Meadows, M., & Shale, E. (1999). Some properties of a simple moving average when applied to forecasting a time series. Journal of the Operational Research Society, 1267-1271.

[14] Kagan, Y. Y., & Jackson, D. D. (1991). Seismic gap hypothesis: Ten years after. Journal of Geophysical Research: Solid Earth (1978âĂŞ2012), 96(B13), 21419-21431.

[15] Keeler, J. (1972). Piecewise-periodic analysis of almost-periodic sounds and musical transients. Audio and Electroacoustics, IEEE Transactions on, 20(5), 338-344.

[16] Keeling, C. D., & Whorf, T. P. (2000). The 1,800-year oceanic tidal cycle: A possible cause of rapid climate change. Proceedings of the National Academy of Sciences, 97(8), 3814-3819.

[17] Kemeny, J. G. & Snell, J. L. (1976) FiniteMarkov Chains: Springer-Verlag, New York, 224 pp.

[18] Li, Z., Wang, J., & Han, J. (2012, August). Mining event periodicity from incomplete observations. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 444-452). ACM.

[19] Lonardi, J. L. E. K. S., & Patel, P. (2002). Finding motifs in time series. In Proc. of the 2nd Workshop on Temporal Data Mining (pp. 53-68).

[20] Mann, S., & Haykin, S. (1991, June). The chirplet transform: A generalization of Gabor's logon transform. In Vision Interface (Vol. 91, pp. 205-212).

[21] Mueen, A., Keogh, E. J., Zhu, Q., Cash, S., & Westover, M. B. (2009, April). Exact Discovery of Time Series Motifs. In SDM (pp. 473-484).

[22] Osborn, D. R., & Smith, J. P. (1989). The performance of periodic autoregressive models in forecasting seasonal UK consumption. Journal of Business & Economic Statistics, 7(1), 117-127.

[23] Rajagopalan, B., Lall, U., & Tarboton, D. G. (1996). Nonhomogeneous Markov model for daily precipitation. Journal of Hydrologic Engineering, 1(1), 33-40.

[24] Rosenblatt, F. (1962). Principles of neurodynamics.

[25] Sansone, A., & Garofalo, G. (2007). Asset price dynamics in a financial market with heterogeneous trading strategies and time delays. Physica A: Statistical Mechanics and its Applications, 382(1), 247-257.

[26] Seber, G. A., & Lee, A. J. (2012). Linear regression analysis (Vol. 936). John Wiley & Sons.

[27] Stamov, G. T., Alzabut, J. O., Atanasov, P., & Stamov, A. G. (2011). Almost periodic solutions for an impulsive delay model of price fluctuations in commodity markets. Nonlinear Analysis: Real World Applications, 12(6), 3170-3176.

[28] Strang, G. (1989). Wavelets and dilation equations: A brief introduction. SIAM review, 31(4), 614-627.

[29] Suzuki, J. (1995). A Markov chain analysis on simple genetic algorithms. Systems, Man and Cybernetics, IEEE Transactions on, 25(4), 655-659.

[30] Tesauro, G. (1989). Neurogammon wins computer Olympiad. Neural Computation, 1(3), 321-323.

[31] Winniford, M. (2003). Real Estate Investment Trusts and seasonal volatility: a periodic GARCH model. Duke University, Durham, NC. Spring.