4th WSEAS Int. Conf. on NON-LINEAR ANALYSIS, NON-LINEAR SYSTEMS and CHAOS, Sofia, Bulgaria, October 27-29, 2005 (pp60-67)

# Prediction and Dynamical Reconstruction of non-stationary data with Delay-Coordinates Embedding and Support Vector Machine Regression*

STERGIOS PAPADIMITRIOU and KONSTANTINOS TERZIDIS,

Department of Information Management, Technological Education Institute of Kavala, 65404 Kavala,Greece

**Abstract**

The paper presents a new effective approach for the construction of local Support Vector Machine (SVM) regression models for the prediction of non-stationary data. We illustrate that an analysis in the framework of dynamical systems theory can provide critically useful parameters for the effective training of the SVM predictors.

A correlation dimension parameter is approximated and is used in order to obtain an appropriate dimensionality for the input space of the predictive SVM. The presented prediction framework can be utilized both for continuous signals and for the case where the observable variable is a discrete symbol, a circumstance very common in data mining problems. Using the information extracted from the correlation dimension computation, local Support Vector Machine models are trained and they are used only for local predictions.

We apply this methodology to the difficult problem of evaluating the predictability of DNA sequences. The results support the importance of the estimation of the proper dimensionality of the embedding space by means of the correlation dimension. Additionally, they demonstrate the effectiveness of the presented SVM based prediction approach that is formulated under a dynamical systems reconstruction framework.

**keywords:** Support Vector Machine, Dynamics Reconstruction, Data Mining, Symbolic Sequence, Signal Prediction, Non-Stationarity, Correlation Dimension

*This work was partially supported from a European Union funded EPEAK II project "Arximidis", code 04-3-001/5, performed at the Technological Educational Institute of Kavalas, Dept. of Information Management, Greece.

# 1 Introduction

Observable temporal (e.g. stock market exchange rates) or spatial sequence data (e.g. DNA sequence) are often the result of complex and insufficiently understood interdependencies. Therefore, prediction models make use of incomplete information, while other factors not included in the models act as noise. In addition, most physical processes are *nonstationary*, meaning that the data distribution is changing over time. For the DNA example, the nonstationarity corresponds to location dependent dynamical rules of sequence composition. For non-stationary domains, a single model built on a certain data segment and used for all the subsequent predictions is generally inadequate. Forecasting these processes requires *on-line learning*, where a given model is used for a limited time and a new model is constructed whenever a change of the underlying data distribution is detected.

The dynamic reconstruction problem concerns the approximation of the unknown function that describes the state evolution of an unknown system [3, 10, 11, 12]. Usually, the state variables and the equations that describe the evolution of the system are unknown, only the measurements of a single output variable are available. However, by using the Takens theorem [9], we can describe the present state of the system by embedding the system output values into a set of lag-vectors. The embedding has the purpose of creating a pseudo-state space, called the *reconstruction space*. The dynamics of the original system that created the time series can be reconstructed within the *reconstruction space*.

The Takens embedding theorem offers the potentiality to treat the problem of dynamic reconstruction as a problem of approximating an unknown multidimensional function from only a finite number of available noisy input-output examples. However, we should note that there is a distinction between the dynamic reconstruction and the prediction problems. The ability to solve the prediction problem does not always imply the potentiality to capture the dynamics of the underlying system. Dynamic reconstruction aims at modeling the attractor state-space dynamics while for the prediction problem we are concerned only with the short term prediction task.

Therefore, the dynamic reconstruction problem should not be considered as a function approximation problem but instead of as a system approximation problem. The reconstructed system should be as close as possible to the original one in terms of the invariants of its dynamics [3]. The capability of the Support Vector Machine for dynamic reconstruction of chaotic systems described with differential equations (e.g. the Lorenz system) has been addressed by Mattera and Haykin [5]. These authors have developed a mathematically robust framework for dynamic reconstruction of chaotic systems and have supported that the SVM method is superior to a Regularized Radial Basis Function design (although they proved that both methods exhibit common theoretical roots).

We concern the dynamic reconstruction problem for actual measurements derived from unknown dynamical systems. Although these systems are nonstationary, in the sense that their dynamical rules of evolution do not remain constant but change over time, we can consider them as constant over a small time frame. Therefore, we can apply the tools of the dynamical systems theory and the Support Vector reconstruction of the dynamics with a "moving frame" like approach.

A related approach is the one proposed in [24, 25]. These authors apply similar techniques for phase space reconstruction in order to identify temporal patterns that are characteristic and predictive of significant events in a complex time series. In [24] an optimization method based on genetic algorithms is defined in order to detect interesting patterns at the reconstructed state space while in [25] the approach is extended to evaluate also the degree with which a candidate temporal pattern is characteristic and predictive of a significant event. However, these approaches are purely unsupervised, while the presented one proceeds in the supervised framework of dynamical systems identification.

We approximate a correlation dimension parameter in order to obtain an appropriate dimension for the reconstruction space. This measure has been widely explored for the construction of embedding spaces for systems evolved according to continuous dynamical rules for which continuous measurement variables are available [13, 10, 11]. We attempt to adapt some concepts to the case where the observable variable is a discrete symbol, a circumstance very common in data mining problems. We use as a particular example of a symbolic sequence the DNA sequence.

The next step of the presented prediction approach, is to utilize the information extracted from the correlation dimension computation, in order to train local Support Vector Machine regression models that they are used only for local predictions. The input space dimensionality of these models is determined by the correlation dimension.

We consider an application to the difficult problem of evaluating the predictability of DNA sequences. The results support the importance of the estimation of the proper dimensionality of the embedding space by means of the correlation dimension. Also, they demonstrate the effectiveness of the presented Support Vector Regression (SVR) prediction approach that is formulated under a dynamical systems reconstruction framework.

Section 2 reviews the concept of the correlation dimension that is applied in order to estimate a proper dimensionality for the embedding space. Section 3 discusses the delay coordinates embedding technique and deals with the subtleties of its application at the dynamical systems reconstruction framework. Section 4 considers the regression Support Vector Machine and its application at the context of nonstationary signal prediction and dynamical systems reconstruction. Thereafter, Section 5 applies the methods to the prediction of DNA sequence. Finally, the conclusions are presented along with directions for future work.

# 2 Correlation Dimension

This section attempts to formulate the *intrinsic dimensionality* of a dataset in order to construct a proper preprocessing framework for prediction. The computation of the correlation dimension al-

lows the detection of the proper embedding dimensionality. We note beforehand that we utilize also the concept of the embedding dimensionality for symbolic attributes in a manner that resembles the formulation for continuous dynamical systems.

The *Embedding Dimension* $E$, of a dataset is defined as the number of attributes of the dataset. By the other hand, the Intrinsic Dimensionality $D$ is the dimensionality of the object represented by the point set, regardless of the *space where it is embedded.* For example, the intrinsic dimensionality of Euclidean objects equals their Euclidean dimension, regardless of the dimension of the space where they are embedded in. Thus, lines, circles and standard curves have $D = 1$; planes, squares and surfaces have $D = 2$; Euclidean volumes have $D = 3$, and so on.

The embedding dimensionality of a dataset can conceal the actual distribution of the data. For example, the DNA series exhibits nonuniform distributions and correlations between attributes (nucleotides in this case). However, the correlated nucleotides and the specific type of correlation are usually not known. The intrinsic dimensionality $D$ can indicate the *existence of correlations* and can give an estimate of the number of attributes that are actually required to characterize a point set. At the context of DNA analysis it offers a novel way to view the process. Indeed, the intrinsic dimensionality gives a *lower bound* of the number of nucleotides that we have to retain in order to keep the essential characteristics of the current state of the DNA sequence. The intrinsic dimensionality of a dataset can be estimated by the correlation dimension metric that is presented below.

Given a finite data set $P$, in order to compute the correlation dimension. we analyze it in a range of scales $(r_{low}, r_{up})$. We cover its data points with a grid having cells of size $r$. We denote by $C_{P,i}(r)$ the count of points of $P$ that fall inside the grid cell $i$ of size $r$. The *Correlation Dimension* $D_2$ is defined as [10, 11]:

$$D_2 = \frac{\partial \log(\sum_i C_{P,i}^2(r))}{\partial \log(r)} = c,\ r_{low} < r < r_{up} \qquad (1)$$

Intuitively, $D_2$ quantifies effectively how "dense" a set $P$ is, since irrespectively of the embedding space dimensionality, the numerator depends only on the density of the set $P$ at the space where it lives (which owns a dimensionality equal to $P$'s intrinsic dimensionality). The Correlation Dimension $D_2$ is of significant importance, since it accounts for the probability of finding one or more points whose distance from a given point in the set is at most $r$, i.e. it is related closely to the correlation concept.

The *Sum of Squared Occupancy* $S_2(r)$ measures sums of occupancies over all grid cells. Let a point set $P$ in an $E$-dimensional space and an $E$-grid with cell side $r$. We denote as usual by $C_{P,i}$ the count of points which fall inside the grid cell $i$ (count of occupancy). Then the Sum of Squared Occupancy $S_2(r)$ is defined as:

$$S_2(r) = \sum_i C_{P,i}^2 \qquad (2)$$

Also, given a point set $P$, the *Correlation Integral* $C(r)$ of $P$ is defined as:

$$C(r) = \frac{\sum C_{non-ordered}(r)}{N \cdot (N-1)/2} \qquad (3)$$

where $C_{non-ordered}(r)$ denotes the count of non-ordered pairs $< p_i, p_j >, i \neq j, p_i, p_j \in P$ within a distance $r$, i.e. $dist(p_i, p_j) \leq r$, and $N \cdot (N-1)/2$ is the number of *unique* pairs $< p_i, p_j >$ in $P$. The pairs are non-ordered in the sense that we count $< p_i, p_j >$ and $< p_j, p_i >$ only once. The correlation integral $C(r)$ expresses the fraction of pairwise distances smaller than $r$.

The Schuster Lemma [21] demonstrates that the correlation integral $C(r)$ is proportional to the sum of squared occupancies $S_2(r)$, i.e.

$$C(r) = c \cdot S_2(r) \qquad (4)$$

with $c$ a constant of proportionality.

Therefore, we can reformulate the definition of the correlation dimension $D_2$ of equation (1) by using instead of the term $\sum_i C_{P,i}^2$ (that is the sum of squared occupancies $S_2(r)$, according to equation (2), the correlation integral $C(r)$. Clearly, the new definition of $D_2$ as

$$D_2 = \frac{\partial \log(C(r))}{\partial \log(r)} = \text{constant}, \quad r_{low} < r < r_{up} \qquad (5)$$

is equivalent to equation (1).

In practice, for a data set that exhibits self-similarity in a range of scales $(r_{low}, r_{up})$, the plot in log-log scale of $C(r)$ versus $r$ will be close to a line in that range. The slope of the best fitted line in that range will approximate the value of $D_2$.

We should note at this point that since the $D_2$ measure accounts for the correlations that exist among the dimensions of the dataset, it also represents the degree of freedom of each attribute in the dataset. Therefore, $D_2$ is a suitable measure for the characterization of the intrinsic dimensionality. We will use the $D_2$ measure computed according to equation (5) in order to estimate the proper dimensionality of the embedding space at the section that follows. In turn, the effectiveness of the prediction support vector regression formulation depends critically on the correct estimation of the embedding dimensionality, as the presented results illustrate.

# 3 Delay Coordinate Embedding

The mathematical background of the prediction formulation is based on the celebrated method of delay coordinates embedding [9, 10, 11]. We assume the existence of an unknown complex underlying process that determines the composition of the observed pattern. For example, at the case of DNA prediction, this process determines the selection of the next nucleotide in a manner that depends on the cell dynamics of the particular organism and the local composition of the previously constructed DNA chain. We cannot observe directly these processes, but we observe only their outcome, e.g. the next nucleotide added to the chain.

This method allows the reconstruction of the dynamics from the observed variable. Let assume that the dynamics of the signal

can be described by a deterministic flow of $N$ generally coupled, nonlinear ordinary differential equations (ODEs),

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}),\ \mathbf{x} = [x_1(t), x_2(t), \ldots, x_D(t)] \quad (6)$$

where $\mathbf{F} = (f_1, f_2, \ldots, f_D)$ are unknown functions of the coordinates $x(t)$.

Let $S(t)$ be a scalar observable depending on the system state $\mathbf{x}(t)$ and obtained by a measurement process, which implements an observation equation $S(t) = h(\mathbf{x}(t))$. Specifically, in many cases the scalar $s(t)$ is the sampled value of the signal value at time $t$.

Given a time series of a scalar observable,

$$[S(t_1), S(t_2), \ldots, S(t_i), \ldots], \quad \Delta t_i = t_i - t_{i-1} \quad (7)$$

recorded at successive intervals, $i = 1, 2, \ldots, N$, a vector $\mathbf{X}_R(t)$, whose coordinates are time-delayed copies of $S(t)$ can be constructed as follows:

$$\mathbf{X}_R(t_i) = [S(t_i), S(t_i - \tau), \ldots, S(t_i - (m-1)\cdot\tau)] \quad (8)$$

where $\tau$ is the delay time and $m$ is the dimensionality of the embedding space. The mapping from the original state space $\mathbf{x} = [x_1(t), x_2(t), \ldots, x_D(t)]$ to the space consisting of the time delayed observables of a single variable is called the *delay coordinate map*.

The number of state variables that describe the system's dynamics (i.e. $x_1, x_2, \ldots, x_D$) in Equation 6 determines the *dimensionality* of the attractor. For dynamics corresponding to an attractor of state-space dimension $D$, a *necessary* condition for determining a $\mathbf{X}_R(t_i)$ able to reconstruct the original system dynamics is $m \geq D$. With a sufficiently large embedding dimension $m$, $\mathbf{X}_R(t_i)$ *unambiguously* describes the state of the system at time $t_i$, thus there exists an equation for points on the attractor, which is of the form

$$\mathbf{X}_R(t+p) = F^*(\mathbf{X}_R(t)) \quad (9)$$

The function $F^*$ of equation 9 allows to predict future values of the time series $\mathbf{X}_R(t)$ given past values, with $p$ being the prediction horizon.

Takens [9] proved that there is an upper bound $m_{upper}$

$$m_{upper} \leq 2 \cdot D_c + 1$$

for the dimension $m$, such that a continuous function $F^*$ of the form of equation 9 that reconstructs the system dynamics can be found within this bound. The parameter $D_c$ is the dimensionality of the underlying state space of the dynamical system. Although $D_c$ is unknown we estimate it by means of the correlation dimension, that gives a rough estimate of the independent state variables of the system. Therefore, the condition $m \geq 2 \cdot D_c + 1$ is a *sufficient* but not a necessary condition for dynamic reconstruction.

More accurately, the Takens theorem states that, under reasonable conditions on the dynamics $\mathbf{F}$ of the system and the observation function $h$, the delay coordinate map from a $D_c$-dimensional smooth compact manifold to $R^{2\cdot D_c+1}$ is a diffeomorphism [1] on that manifold where $D_c$ is the capacity dimension of the attractor of the dynamical system.

Clearly, if the dimension $m$ of the embedding space is too small, the orbit constitutes a projection, which will tend to "fill" completely the available $m$-dimensional state space. On the contrary, if $m$ increases beyond a critical integer value $m_E$, called the *embedding dimension*, some of the main properties of the dynamics are expected to remain unchanged and thus the correlation dimension of the reconstructed attractor remains constant as we increase further the dimensionality of the embedding (i.e. the parameter $m$).

The procedure for finding a suitable $m$ is called *embedding*. The best dynamical reconstruction results are obtained when we use the embedding dimension $m_E$ as the dimensionality of the delay coordinate space. With smaller dimensionality we loss the attractor's structure. A utilization of greater dimensionality increases significantly the inaccuracies from the noisy and also perplexes the design of the prediction neural network, since as is well known neural network performance usually degrades fastly with the increasing input space dimensionality. Furthermore, usually for applications that involve mining of spatio-temporal data, the dynamics of the signal seem very nonstationary and can be modeled only locally by the *on-line* construction of local dynamical systems, valid only in the neighborhood of the region on which they were trained. Thus, in our case a long embedding vector (i.e. parameter $m$), is improper and is not expected to produce correct results even in the absence of noise.

The delay-embedding theorem implies that the evolution of the points $\mathbf{X}_R(n) \rightarrow \mathbf{X}_R(n+1)$ in the reconstruction space follows that of the unknown dynamics $\mathbf{x}(n) \rightarrow \mathbf{x}(n+1)$ in the original space. This implies a powerful result: many important properties of the unobservable state vector $\mathbf{x}(n)$ are reproduced without ambiguity in the reconstruction space defined by $\mathbf{X}_R(n)$ that evolves according to the reconstructed dynamics described by Equation 9. The developed Support Vector Machine (SVM) regression method aims to estimate the prediction function 9 on the basis of time-delay coordinates according to 8. From the above discussion it becomes evident that the correct estimation of the embedding dimension $m_E$ parameter and of the embedding delay $\tau$ from the time-series data is critical for successful prediction.

# 4 Support Vector Prediction

The dynamical reconstruction problem is the problem of approximating the unknown function $F^*$ of equation 9. In practice, usually

[1] The mapping $\mathbf{f} : U \rightarrow V$ is said to be a *diffeomorphism* of U onto V if it satisfies the following three conditions:

1. $\mathbf{f}(U) = V$.
2. The mapping $\mathbf{f} : U \rightarrow V$ is a one-to-one (i.e., invertible).
3. Each component of the inverse mapping $\mathbf{f}^{-1} : V \rightarrow U$ is continuously differentiable with respect to its argument.

only a few number of samples of the time series can be assumed to be known. Unfortunately, for many data mining applications, we are limited to a very small number of samples. The nonlinear reconstruction problem, can be stated as the problem of getting an approximation $\hat{F}$ of the function $F^*$ when only noisy values assumed by $F^*$ in a small number of points are available. Therefore, the Statistical Learning Theory and the accompaning implementation framework of Support Vector Machines (SVMs) [8] find a natural application for this type of dynamical reconstruction problem. Some other approaches to the dynamical system reconstruction utilize regularized Radial Basis Function networks [3, 15, 15]. However, the Support Vector Machine is based on the robust theory of Empirical Risk Minimization developed by Vapnik [8] and allows more disciplined model selection. Comparative experiments that we have performed with Radial Basis Function networks proved a superiority in favor of the Support Vector Machine.

The training set for the predictive SVM is constructed according to the delay coordinates method presented in the previous section. However, in order to be able to apply this important theory, *reliable estimates* of the embedding dimension $m_E$ and of the embedding delay should be obtained. We discuss some basic issues arising in this rather complicated estimation problem.

*Embedding Dimension*

The sufficient condition $m \geq 2 \cdot D_c + 1$ makes it possible to undo the intersections of an attractor's orbit with itself. These intersections can arise from projections of that orbit to lower dimensions. Frequently the embedding dimension $m_E$ is less than $2 \cdot D_c + 1$. Fortunately, there exist methods that estimate the embedding dimension directly from the observable data. The method of *false nearest neighbors* is one reliable choice for this task [12]. This algorithm systematically for all the data points and their neighbors, starting with dimension $d = 1$, then $d = 2$ and so on, examines when their apparent neighbors stop being "unprojected" by the addition of more elements to the reconstruction vector $\mathbf{S}(n)$. The value of $d$ that halts the unprojection is a reliable estimate of the embedding dimension. Another method is to estimate the correlation dimension $D_c$ of the attractor and to obtain the embedding dimensionality as $2 \cdot D_c + 1$. We adopted the later method since as we noted the correlation dimension concept allows also to gain insight to the data's intrinsic dimensionality.

*Embedding Delay*

The time $\tau$ is the time lag between successive values of the scalar observable, i.e. it corresponds to the sampling period that we use in order to extract state vectors from the scalar observable. The proper prescription for choosing $\tau$ is to recognize that the normalized embedding delay $\tau$ should be large enough for $\mathbf{S}(n)$ and $\mathbf{S}(n-\tau)$ to be in some extent independent of each other but not too much independent in order to retain the geometry of the attractor. This independence constraint allows to use them as coordinates of the reconstruction space.

The main idea for Support Vector Prediction is to map the data $\mathbf{x}$ into a high-dimensional feature space $\mathcal{F}$ via a nonlinear mapping $\Phi$, and to perform linear regression in this space, i.e. [7, 8, 16, 17, 19]

$$f(\mathbf{x}) = (\mathbf{w} \cdot \Phi(\mathbf{x})) + \mathbf{b} \tag{10}$$

with $\Phi$:$\Re^n \rightarrow \mathcal{F}$, $\mathbf{w} \in \mathcal{F}$ and $\mathbf{b}$ is a threshold.

Thus, *linear* regression in a high dimensional feature space corresponds to *nonlinear* regression in the *low* dimensional input space $\Re^n$. The dot product of Equation 10 between $\mathbf{w}$ and $\Phi(\mathbf{x})$, $\mathbf{w} \cdot \Phi(\mathbf{x})$, would have to be computed in the high dimensional feature space $\mathcal{F}$. The direct computation of this inner product is usually computationally intractable for most applications. However, it can be computed efficiently in $\Re^n$ with the inner-product space *kernel* function mapping trick. This avoids the need to compute inner products in the high dimensional feature space $\mathbf{F}$. Since $\Phi$ is fixed, we determine $\mathbf{w}$ from the data by minimizing the sum of the empirical risk $R_{emp}(f)$ and a complexity term $\|\mathbf{w}\|^2$. The later term enforces *flatness* in feature space. Therefore, the regularized risk functional $R_{reg}$ becomes:

$$R_{reg}(f) = R_{emp}(f) + \lambda\|\mathbf{w}\|^2 = \sum_{i=1}^{N} L(f(\mathbf{x}_i) - y_i) + \lambda\|\mathbf{w}\|^2 \tag{11}$$

where $N$ denotes the sample size $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, $L(.)$, is a loss function, and $\lambda$ is a regularization constant [1, 3, 8]. A wide range of loss functions exists for which equation 11 can be minimized by solving a quadratic programming problem. This problem obtains a global unique optimal solution and therefore the case for trapping at local minima is avoided [8]. The vector $\mathbf{w}$ can be written in terms of data points as:

$$\mathbf{w} = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)\Phi(\mathbf{x}_i) \tag{12}$$

with $\alpha_i, \alpha_i^*$ being the solution of the quadratic programming problem mentioned earlier.

Taking Equations 10 and 12 into account, we can rewrite the solution to the whole problem in terms of dot products in the low dimensional input space:

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)(\Phi(\mathbf{x}_i)\Phi(\mathbf{x})) + b \\ &= \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) k(\mathbf{x}_i, \mathbf{x}) + b \end{aligned} \tag{13}$$

In equation 13 we introduced a kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$.

Effective computational methods exist for solving the above quadratic programming problems [4, 6]. We used the method of Joachims [4], implemented with the public available software package SVMLight. Therefore, a particular loss function does not seem to own a clear advantage, at the specific prediction applications. As a kernel type, we found that the Radial Basis Function kernel obtained the best results. The next section presents and discusses the DNA sequence prediction application.

# 5 Application

We present an application of the presented framework for the prediction of non-stationary data. The application deals with the prediction of DNA sequences. The prediction of the DNA sequences, concerns prediction at the symbolic domain and thus is quite different from the prediction of numeric time-series. DNA series from different species are expected to present quantitatively different predictabilities. Additionaly, different segments of the eukaryotic DNA (e.g. coding regions, regulatory regions, repetitive elements) are expected to impose their own limitations to prediction. The predictability outcomes with increasing prediction horizon fall gracefully. This fact seems to indicate long range dependencies at the dynamical process of DNA construction.

The methodology that we have used has the following basic steps:

- We use *annotated* DNA segments for well studied organisms (e.g.E-coli, S-cerevisia, mouse, Homo-sapiens) and we create a simple numeric representation (e.g. 0 for adenine, 1 for thymine, 2 for guanine, 3 for cytosine).

- We evaluate the Correlation Dimension of the DNA signal.

  In order to compute the Correlation Dimension of the DNA signal we read the encoded nucleotides creating a data matrix. The distance between the patterns is computed as the number of different nucleotides they have, normalized properly in order to fit the range of radius on which we evaluate the fractality of the signal. Using a dense sampling of radius we compute the Correlation Dimension according to Equation (1). Using the simple numeric representation we obtain a Correlation Dimension estimate of about 7.6. Figure 1 illustrates the computation. This implies that according to the nonlinear systems theory we expect to obtain a proper embedding for an embedding dimension of greater than $2 \cdot 7.6 + 1$, i.e. for a time delay vector of a dimension of 17. Indeed, we have verified that the prediction results are optimized for embedding vectors in the range of 16-18.

  Alternatively, we can utilize different more sophisticated representations of the DNA sequence [22] as the *Chaos Game Representation (CGR)* [23]. At the later representation, the positions $CGR_I$, of each nucleotide, $g_i$, of a sequence, $g$ of length $n_G$, is calculated by moving a pointer to half the distance between the previous position and the current binary representation (equation 14). The binary CGR vertices are assigned to the four nucleotides as $A = (0,0), C = (0,1), G = (1,1), T = (1,0)$. Then the computation proceeds as:

$$CGR_k = CGR_{k-1} + 0.5 \cdot (CGR_{k-1} - g_k), \quad k = 1, \ldots, n_G, CGR_0 = (0.5, 0.5) \qquad (14)$$

The Chaos Game Representation offers more compact encoding of the DNA sequence and yields a correlation dimension estimate of 2.6. Therefore, the theoretical optimal embedding dimension space is 6. We have verified that the bset results are obtained at about this theoretically expected value, i.e. for embedding vectors of dimension 6 or 7. Figure 2 illustrates the prediction results deteriorate significantly for large embedding vectors, due to the effects on neural training of the curse of dimensionality (since the dynamical systems theory we know that a larger embedding dimensionality from the one required does not alter the structure of the attractor [11, 10]).

- We specify the parameters of the local SVM model that will be trained in a local moving frame in order to predict the next nucleotide. The underlying dynamics that determine the composition of the DNA seem to exhibit long-range correlations. The predictability outcomes with increasing prediction horizon (i.e. by trying to predict two nucleotides ahead instead of one) fall gracefully. This fact seems to indicate long range dependencies at the dynamical process of DNA construction.

  We succeed relatively well to predict the next nucleotide, and for a few nucleotides ahead the results are still good (the ability to predict falls slowly). Predictability is completely lost only for a large number of nucleotides ahead (about 50-60).

  Other parameters that need specification concern the type of the SVM model. We consider the SVM with radial basis kernel as a particularly effective model.

# 6 Conclusions

We have applied a new SVM prediction framework for the analysis of non-stationary signals. We have demonstrated the importance of the correct evaluation of the embedding space dimensionality and we have used the correlation dimension measure as a means to estimate a proper embedding dimensionality. Therefore the analysis of the time-series data in the framework of dynamical systems theory can provide critically useful parameters for the effective training of SVM predictors.

We have presented the development of local Support Vector prediction models for the prediction of non-stationary time series. We demonstrate that the same methods can be used for another difficult problem that involves discrete symbols: the prediction of the DNA sequence. In this case, the results provide additional support for the importance of the estimation of the proper dimensionality of the embedding space by means of the correlation dimension.

Much future work remains in order to obtain better insights to more effective approaches for utilizing the dynamical systems theory at the development of better Support Vector prediction machinery. The utilization of the improved support vector algorithms [17, 16] and of alternative implementation approaches [19, 6] perhaps can improve further the results. Also, alternative machine learning approaches [18, 20] need to be developed for these problems, in order to be able to compare the strengths and the weakness of the support vector approach with them.

Figure 1: The Correlation Dimension Computation for the DNA signal obtains an estimate of about 7.6

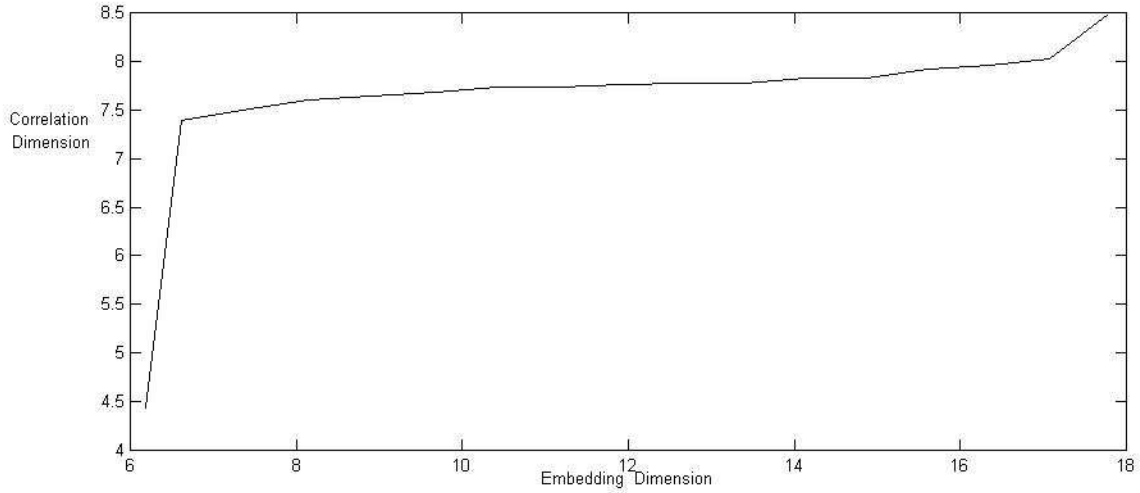
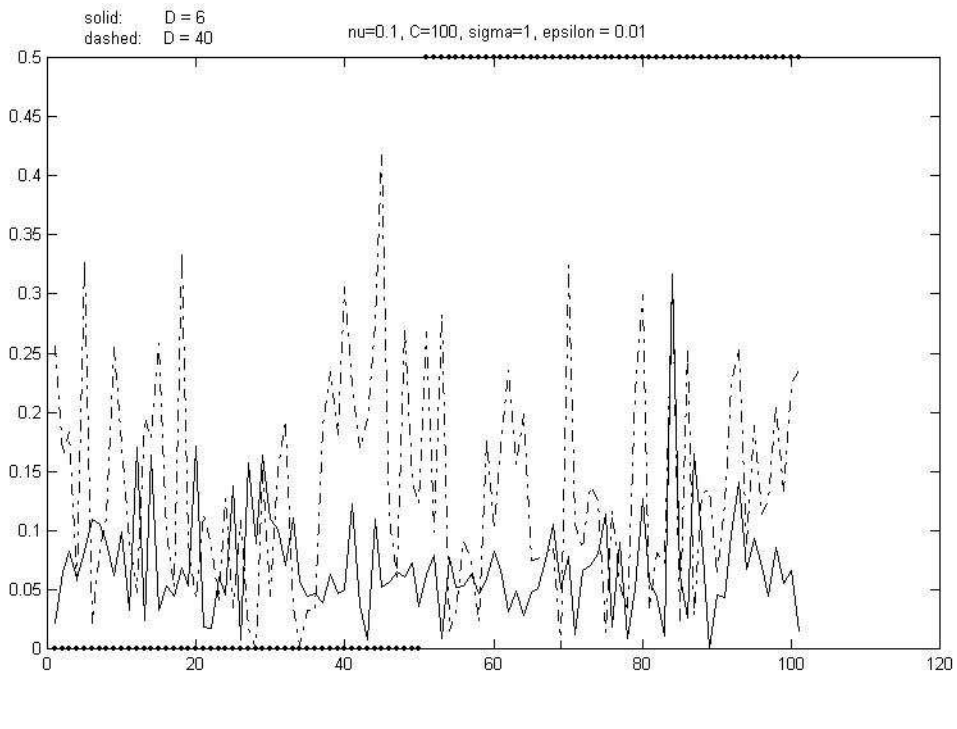



Figure 2: The prediction error is much smaller for embedding dimension D=6, i.e. for an embedding dimension value near the predicted value



# References


[1] P. Bartlett, J. S.-Taylor, "Generalization Performance of Support Vector Machines and Other Pattern Classifiers", Advances in Kernel Methods, Support Vector Learning, The MIT Press, 1999, pp. 43-55.

[2] C. Cortes, V. Vapnik, "Support vector networks", Machine Learning, vol. 20, pp 1-25, 1995

[3] Simon Haykin, Neural Networks, MacMillan College Publishing Company, Second Edition, 1999

[4] T. Joachims, "Making Large-Scale SVM Learning Practical", Advances in Kernel Methods - Support Vector Learning", Bernhard Scholkopf, Christopher J. C. Burges, and Alexander J. Smola (eds), MIT Press, Cambridge, USA, 1998

[5] D. Mattera, S. Haykin, "Support vector machines for dynamic reconstruction of a chaotic system", in Advances in Kernel Methods - Support Vector Learning, B. Scholkopf, J. Burges, A. J. Smola,, Eds. Cambridge, MA:MIT Press, 1999, pp. 211-242

[6] E. Osuna, R. Freund, F. Girosi, "An improved training algorithm for support vector machines", Neural Networks for Signal Processing VII, Proceedings of the 1997 IEEE Workshop pp. 276-285, Amelia Island, FL.

[7] B. Scholkopf, S. Mika, J. C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch and A. Smola, "Input Space Versus Feature Space in Kernel-Based Methods", IEEE Trans. On Neural Networks, vol. 10, no. 5, 1999.

[8] V. N. Vapnik., 1998, Statistical Learning Theory, New York, Wiley

[9] F. Takens, "Detecting strange attractors in turbulence", in Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1981, Vol. 898, pp.366-381

[10] Anastasios A. Tsonis, Chaos: From Theory to Applications, Plenum Press, 1992

[11] Edward Ott, Chaos in Dynamical Systems, Cambridge University Press, 1993

[12] Abarbanel, H.D.I., 1996, Analysis of Observed Chaotic Data, New York: Springer-Verlag

[13] T. Sauer, J. A. Yorke, and M. Casdagli, Embedology, J. Stat. Phys., 65:579-616, 1991

[14] S. Papadimitriou, A. Bezerianos, A. Bountis, "Radial Basis Function Networks as Chaotic Generators for Secure Communication Systems", International Journal On Bifurcation and Chaos, Vol. 9, No. 1,1999, p. 221-232

[15] A. Bezerianos, S. Papadimitriou, D. Alexopoulos, "Radial Basis Function Neural Networks for the Characterization of Heart Rate Variability Dynamics", Artificial Intelligence in Medicine, Vol. 15,1999, p. 215-234

[16] Bernhard Scholkopf, Alexander J. Smola, "Learning with Kernels: Support Vector Machines, Regularization and Beyond", MIT Press 2002

[17] B. Scholkopf, A. J. Smola, R. C. Williamson, P. L. Bartlett, "New support vector algorithms", Neural Computation:1207-1245, 2000

[18] Vasilios Petridis, Vasilis Kaburlasos, "Fuzzy Lattice Neural Network (FLNN): A Hybbrid Model for Learning", IEEE Transactions on Neural Networks, Vol. 9, No 5, Semptember 1998

[19] Chih-Chung Chang, Chih-Jen Lin, "Training $\nu$-Support Vector Classifiers: Theory and Algorithms", Neural Computation 13, 2119-2147 (2001)

[20] William Bialek, Ilya Nemenman, Naftali Tishby, "Predictability, Complexity, and Learning", Neural Computation 13, 2409-2463 (2001)

[21] H. G. Schuster, Deterministic Chaos, Physik-Verlag, Weinheim, Germany, 1984

[22] Joao-Aires-de-Sousa, Luisa Aires-de-Sousa, "Representation of DNA sequences with virtual potentials and their processing by (SEQREP) Kohonen self-organized maps", Bioinformatics, Vol. 19, no 1, pp. 30-36, 2003

[23] Jonas S. Almeida, Joao A. Carrico, Antonio Maretzek, Peter A. Noble, Madilyn Fletcher, "Analysis of genomic sequences by Chaos Game Representation", Bioinformatics, Vol. 17, no. 5, pp. 420-437, 2001

[24] R.J. Povinelli and X. Feng, "A New Temporal Pattern Identification Method for Characterization and Prediction of Complex Time Series Events", *IEEE Trans. Knowledge and Data Eng.*, vol. 15, no. 2, pp. 339-352, Mar./Apr. 2003

[25] Xin Feng, Hai Huang, "A Fuzzy-Set-Based Reconstructed Phase Space Method for Identification of Temporal Patterns in Complex Time Series", IEEE Trans. Knowledge and Data Eng., vol. 17, No. 5, May 2005