Régressions par machines à vecteurs supports pour la prédiction de séries chaotiques

Herwig WENDT, Patrick FLANDRIN, Patrice ABRY

Laboratoire de Physique, UMR 5672, CNRS, Ecole Normale Supérieure de Lyon,
46, allée d'Italie, 69364, Lyon cedex 7, France.
prénom.nom@ens-lyon.fr

Résumé – Nous nous intéressons à la prédiction de séries temporelles chaotiques à partir de régressions réalisées à l'aide de machines à vecteurs supports (SVM). Après avoir rappelé le principe de ces régressions par SVM, nous détaillons le schéma de prédiction. Nous illustrons ses performances par une mise en oeuvre, d'une part, sur des données synthétiques produites par le système dynamique de Hénon et, d'autre part, sur des données expérimentales issues de la base de données de Santa Fe communément utilisées comme référence dans les problèmes de prédictions de séries temporelles. Nous comparons positivement nos résultats à ceux proposés antérieurement dans la littérature.

Abstract – We consider the problem of chaotic time series prediction by means of support vector machines (SVM) for regression estimation. After a short review of the principles of SVM for regression estimation, we detail the prediction procedure and illustrate its performance on the synthetic time series produced by the Hénon map, and on the real world time series of Santa Fe Data Set A, often considered as a reference in benchmarking time series predictors. A comparison of our results with others reported in literature demonstrates the excellent performance of the approach.

1 Motivation

Du fait de leur propriété de *sensibilité aux conditions initiales*, les séries temporelles produites par des systèmes dynamiques chaotiques présentent une difficulté particulière pour la prédiction de leurs valeurs futures. Il est notoirement connu que les méthodes usuelles de prédiction, modélisation linéaire ARMA, par exemple, ont des performances médiocres lorsqu'elles sont appliquées aux séries chaotiques (voir, par exemple, [1, 15]).

Diverses approches non linéaires, reposant principalement sur les *réseaux de neurones* (voir, par exemple, [11, 13, 14]) ou des statistiques d'ordres supérieurs [6], ont ensuite été proposées dans la littérature. Plus récemment, la technique des *machines à vecteurs supports* ("support vector machines", SVM) a également été envisagée [7, 8]. C'est à cette dernière approche que nous nous intéressons ici. Nous utilisons un schéma de prédiction par régressions réalisées par SVM. Sur des séries synthétiques issus de systèmes dynamiques chaotiques ainsi que sur des séries expérimentales, nous qualifions les performances de notre outil et les comparons à celles citées dans la littérature.

2 Machines à vecteurs supports

Les machines à vecteurs supports constituent des estimateurs non linéaires universels de fonctions, dont les fondements reposent sur la Théorie Statistique de l'Apprentissage. La formulation du problème d'estimation, introduite dans [3], peut recevoir une solution efficace qui laisse peu de paramètres libres à choisir. Les SVM travaillent à partir de classes de fonctions hypothèses $\mathcal{H}_{\text{SVM}}$ consistant en hyperplans $(\mathbf{w}, b)$ d'un espace $\mathcal{F}$ de *caractéristiques*. Celui-ci est implicitement défini, à partir de l'espace original, par une transformation non linéaire, construite via un noyau $K(\cdot,\cdot)$ (*astuce du noyau*).

2.1 Théorie Statistique de l'Apprentissage

Dans le cadre de la Théorie Statistique de l'Apprentissage [12], l'objectif est d'*estimer* une fonction $y = f(\mathbf{x})$ à partir d'un nombre limité d'échantillons S=$\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ $\in \mathbb{R}^d \times \mathbb{Y}$, et d'hypothèses faites sur les propriétés de $\mathcal{H}$. La classification, avec $y_i \in \mathbb{Y} = \{-1, 1\}$, et la régression, avec $y_i \in \mathbb{Y} \subseteq \mathbb{R}$ constituent deux exemples typiques de tels problèmes.

Le meilleur estimateur $\hat{f}(\mathbf{x}) = h^*(\mathbf{x}) \in \mathcal{H}$ de f est celui qui minimise le risque $R[h]$, mesuré via l'espérance mathématique d'une fonction de coût $L(y, h(\mathbf{x}))$:

$$R[h] = \int L(y, h(\mathbf{x}))dP(\mathbf{x}, y). \qquad (1)$$

Cependant, la distribution jointe $P(\mathbf{x}, y)$ est inconnue a priori et généralement inaccessible, il faut donc approximer (1). La Théorie Statistique de l'Apprentissage fournit alors des résultats sous forme d'inégalités,

$$R[h] \leq \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in S} L(y_i, h(\mathbf{x}_i)) + Q(n, h, \delta), \qquad (2)$$

où $Q(n, h, \delta)$ consiste en un terme de confiance, fonction du nombre n d'observations, de la capacité de la sous-classe $\mathcal{H}_k \subset \mathcal{H}$ qui contient l'hypothèse h, et de la probabilité $1 - \delta$ avec laquelle l'inégalité (2) est valide. La procédure de minimisation du risque structurel ("structural risk minimisation", SRM) procède par partage de $\mathcal{H}$ en sous-classes emboîtées $\mathcal{H}_k$: $\mathcal{H}_1 \subset$

$\mathcal{H}_2 \subset \cdots \subset \mathcal{H}_M$ de capacités croissantes, ce qui permet de chercher l'hypothèse $h^* \in \mathcal{H}$ qui minimise (2).

Dans le cadre des SVM, l'hyperplan $(\mathbf{w}^*, b^*)$ optimal de $\mathcal{H}_{\text{SVM}}$ est ensuite obtenu via la résolution de :

$$h^*(\mathbf{w}^*, b^*, \mathbf{x}) = \\ = \arg\min_{(\mathbf{w},b)} C \sum_{(\mathbf{x}_i,y_i)\in S} L(y_i, h(\mathbf{w}, b, \mathbf{x}_i)) + \frac{1}{2}\|\mathbf{w}\|^2, \quad (3)$$

où C est une constante à choisir.

2.2 Régressions par SVM

Pour la classification, essentiellement considérée dans la littérature SVM, on utilise la fonction de coût $L_{0/1}$. Ici, nous nous intéressons à la régression. Dans ce cas, des fonctions de coût linéaire ou quadratique L_ϵ sont utilisées, ne prenant en compte que les déviations $|y_i - h(\mathbf{x}_i)| > \epsilon$ supérieures à ϵ, où ϵ devient un autre paramètre à fixer. Que ce soit avec $L_{0/1}$ ou avec L_ϵ, l'équation (3) peut être réécrite comme un problème dual de Lagrange [4]. Cette reformulation, qui constitue la base de l'algorithme des SVM, se ramène à un problème de programmation quadratique, de solution unique et pour lequel des méthodes de résolution efficaces existent.

Pour le problème de la régression, la solution de l'équation (3) prend la forme :

$$h^*(\mathbf{x}) = \hat{f}(\mathbf{x}) = \sum_{(\mathbf{x}_i,y_i)\in S} \alpha_i^* K(\mathbf{x}_i, \mathbf{x}) + b^*. \quad (4)$$

Les α_i^* sont les multiplicateurs de Lagrange impliqués dans la solution du problème dual de Lagrange de (3). Quand $\epsilon > 0$, un nombre important des α_i^* sont égaux à zéro, et (4) fournit alors une représentation creuse. Les $\mathbf{x}_i$ correspondant aux $\alpha_i^* \neq 0$ sont appelés les *vecteurs de support*.

3 Prédiction de signaux chaotiques

3.1 Séries chaotiques

Nous nous intéressons maintenant à des séries produites par des systèmes dynamiques chaotiques, dont le caractère non linéaire et la sensibilité aux conditions initiales rendent particulièrement difficile la prédiction, même à court terme [1, 7]. Sous-jacent au système dynamique, existe un modèle non linéaire $y(k) = f(y(k-\kappa_1), y(k-\kappa_2), \cdots, y(k-\kappa_m))$, $\kappa_i \in \mathbb{N}^+$, que nous tentons d'approcher par SVM pour effectuer une prévision du futur de la série temporelle observée $\{y(k)\}_{k=1}^N \in \mathbb{R}$.

Les systèmes dynamiques sont souvent efficacement étudiés via leur espace des phases. Le célèbre théorème de Takens (voir, par exemple, [1]) assure une reconstruction non ambiguë de cet espace par la méthode des *vecteurs de retard* : $\mathbf{x}(k) = [y(k), y(k-\tau), \cdots, y(k-(m-1)\tau)] \in \mathbb{R}^m$, où m et τ constituent respectivement la dimension de plongement et le retard. A condition que ces paramètres soient convenablement choisis, la topologie de l'espace des phases réel est respectée, on peut alors envisager d'estimer la fonction $f : \mathbf{x}(k) \rightarrow y(k+1)$ à partir des observations y [1].
Nous sélectionnons τ par un argument d'information mutuelle (voir, par exemple, [1]). Le paramètre m est soit choisi à partir d'arguments théoriques, soit traité comme un paramètre libre du schéma de prédiction.

3.2 Schéma de prédiction

A partir de N observations, $\{y(k)\}_{k=1}^N \in \mathbb{R}$, sont construits $n = N - (m-1)\tau - 1$ échantillons $S = \{(\mathbf{x}(k), y(k+1))\}_{k=(m-1)\tau+1}^{N-1} \in \mathbb{R}^m \times \mathbb{R}$. Le schéma de prédiction à un pas $\hat{y}(k+1) = \hat{f}(\mathbf{x}(k))$ par SVM est alors de la forme (4),

$$\hat{y}(k+1) = \sum_{i=(m-1)\tau+1}^{N-1} \alpha_i^* K(\mathbf{x}(i), \mathbf{x}(k)) + b^*, k \geq N, \quad (5)$$

dont les α_i^* sont obtenus par résolution du problème dual de Lagrange de (3).

Nous utilisons la boîte à outils SVM standard SVMlight [5] (svmlight.joachims.org), sans apporter de modification spécifique pour la prédiction de séries temporelles. Les paramètres libres, C, ϵ et la taille du noyau (fixé gaussien), éventuellement la dimension de plongement m, sont sélectionnés à partir d'une recherche exhaustive dans l'espace des paramètres pour optimiser les performances de la prédiction sur l'ensemble de *validation*. Les N observations disponibles sont donc partagées entre deux ensembles d'entraînement et de validation de tailles respectives N_e et N_V. Les valeurs pour lesquelles l'erreur de prédiction à un pas sur l'ensemble de validation est minimale sont retenues pour la prédiction finale.

Une fois les paramètres fixés, nous réalisons la prédiction en utilisant la totalité des N observations disponibles. Les prédictions à plusieurs pas, i.e., pour les valeurs $(k \geq N+1)$, sont réalisées par itération de la prédiction à un pas (5), en utilisant les vecteurs estimés $\hat{\mathbf{x}}(k)$ aux itérations précédentes et non les observations elles-mêmes.

4 Expérimentations

4.1 Séries temporelles

Nous appliquons cette approche à la prédiction, d'une part, d'une série issue du système dynamique de Hénon, en régime chaotique (cf. figure 1 en haut à gauche) :

$$y(k+1) = 1 - 1.4y(k)^2 + 0.3y(k-1), \quad (6)$$

avec $a = 1.4, b = 0.3$ et, d'autre part, aux données expérimentales réelles constituées par les fluctuations chaotiques de l'intensité d'un laser dans l'infrarouge, *Santa Fe Data Set A* [15]. Cette dernière série constitue une référence souvent utilisée pour étalonner les performances des méthodes de prédictions.

Pour la série **Hénon**, $N = 500$, $N_e = 450$, $N_V = 50$, l'équation (6) indique que les observations $y(k+1)$ sont complètement déterminées par leurs deux prédécesseurs immédiats, $y(k)$ et $y(k-1)$. Les paramètres de plongement sont donc fixés a priori ($\tau = 1$ et $m = 2$).

Pour la série **Santa Fe Data Set A**, $N = 1000$ (les 1000 premières observations de la série) , $N_e = 900$, $N_V = 100$, l'information mutuelle indique $\tau = 1$. La recherche exhaustive, effectuée pour m à partir d'un intervalle de valeurs choisi par un argument géométrique ("false nearest neighbours" ; voir, par exemple, [1]) sélectionne $m = 18$.

4.2 Résultats

Nous présentons les résultats obtenus dans les tableaux 1 et 2 et les comparons à ceux obtenus par d'autres auteurs. Les

performances des prédictions à un pas et itérées sont exprimées en erreur quadratique moyenne normalisée (NMSE) :

$$\text{NMSE} = \frac{\sum_{k\in\mathcal{T}}(y(k)-\hat{y}(k))^2}{\sum_{k\in\mathcal{T}}(y(k)-\bar{y}_{\mathcal{T}})^2} \approx \frac{1}{\hat{\sigma}^2_{\mathcal{T}} N}\sum_{k\in\mathcal{T}}(y(k)-\hat{y}(k))^2,$$

où $\mathcal{T}$ dénote l'ensemble d'épreuve de N points, et $\bar{y}_{\mathcal{T}}$ et $\hat{\sigma}^2_{\mathcal{T}}$ dénotent respectivement la moyenne et la variance empiriques de $\mathcal{T}$ [15].

La série de Hénon. Les résultats pour la série de Hénon sont obtenus à partir des prédictions itérées sur 50 échantillons, moyennées sur 50 ensembles d'épreuve différents. Nous constatons que, quoique l'ensemble d'entraînement utilisé dans notre prédiction SVM soit beaucoup plus court que ceux mis en oeuvre pour obtenir les meilleurs résultats disponibles à notre connaissance dans la littérature, les performances que nous obtenons sont supérieures de presque deux ordres de grandeur (cf. tableau 1).

TAB. 1 – Comparaison des performances de prédicteurs sur la série de Hénon. La longueur de l'ensemble d'entraînement est donnée en points. Les résultats sont obtenus par moyenne des prédictions sur 50 ensembles d'épreuve différents de 50 échantillons.

Référence	Méthode	N	Préd. un pas
	SVM Gauss	500	$\mathbf{7.7 \cdot 10^{-7}}$
[9]	RBF Net	4000	$7.4 \cdot 10^{-5}$
[13]	Neural Net	5000	$3.4 \cdot 10^{-5}$

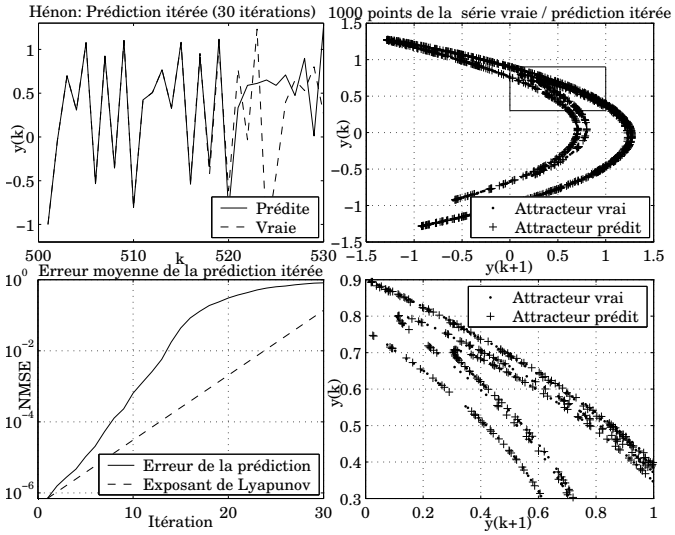

FIG. 1 – **Série de Hénon.** En haut, à gauche, séries de Hénon réelle et prédite, en bas, à gauche, évolution de l'erreur quadratique moyenne en fonction de l'horizon de prédiction et comparaison à la croissance de l'erreur initiale déterminée par l'exposant de Lyapunov global théorique. A droite, en haut, attracteurs de Hénon, reconstitués à partir de 1000 échantillons de la série observée et de la série prédite par itération successive, et agrandissement de la région des attracteurs matérialisée par un rectangle (en bas).

La figure 1 propose une prédiction typique sur la série de Hénon. Nous observons que la prédiction s'écarte de la vraie série après environ 18 itérations (en haut à gauche). C'est aussi la valeur pour laquelle l'erreur de prédiction est proche de 1. L'évolution de cette erreur de prédiction est comparée à la croissance de l'erreur initiale déterminée par l'exposant de Lyapunov global théorique ($\lambda_{\text{Hénon}} \approx 0.42$) (en bas à gauche). Néanmoins, la prédiction, au delà de l'échantillon 18, continue d'explorer l'espace des phases de façon identique à la série réelle. En effet, la colonne de droite compare les attracteurs reconstruits à partir de 1000 points de la vraie série de Hénon (en haut), et d'une prédiction itérée de 1000 points (en bas). Le prédicteur a complètement saisi les dynamiques du système puisque les deux attracteurs ne peuvent pas être distingués, contrairement aux résultats présentés dans [2], où l'attracteur présente des déformations visibles.

Santa Fe Data Set A. Les prédictions sont obtenues pour des ensembles d'épreuve proposés dans la littérature pour Data Set A (5 ensembles de 100 observations, correspondant aux observations 1001 à 1100, 2181 à 2280, 3871 à 3970, 4001 à 4100 et 5181 à 5280). Dans le tableau 2, nous observons que la performance du prédicteur à un pas obtenue par la régression à l'aide de SVM, évaluée sur l'ensemble d'épreuve 1, est comparable à l'état de l'art et que la qualité de la prédiction itérée dépend fortement de l'ensemble d'épreuve considéré. Bien que la performance de la prédiction itérée obtenue par SVM soit dégradée d'un ordre de grandeur pour les ensembles 2 et 3, elle reste supérieure d'un ordre de grandeur comparée au meilleur résultat de l'ensemble 4, et comparable à l'état de l'art pour les ensembles 1 et 5.

Nous soulignons que notre prédicteur est entraîné exclusivement sur les 1000 premières observations, et ce pour tous les ensembles d'épreuves. De plus, nous travaillons directement à partir de la série temporelle, sans aucune technique additionnelle du type de celles considérées dans [10].

La figure 2 illustre le Data Set A et le premier ensemble d'épreuve (en haut, à gauche) et les prédictions itérées, sur 100 points, pour les cinq ensembles d'épreuve considérés. Nous remarquons que la prédiction par SVM est capable de prédire la brutale diminution de l'intensité du laser pour deux des trois cas (ensemble 1 et 5), bien qu'il existe seulement un exemple d'un tel effondrement dans l'ensemble d'entraînement. Par contre, pour l'ensemble 2, le prédicteur échoue à prédire la diminution de l'intensité et la prédiction n'est précise que pour les 50 premières itérations. Nous observons un comportement similaire pour l'ensemble 3 : bien que la prédiction soit de haute précision pendant les 60 premières itérations, une diminution de l'intensité qui n'existe pas est prédite après 80 itérations. Nous constatons aussi que la prédiction itérée et la vraie série ne peuvent pas être distinguées pour l'ensemble 4.

5 Conclusions et perspectives

Quoique la régression par SVM pour la prédiction de séries temporelles soit appliquée ici de façon très directe, les performances sont excellentes pour les séries considérées. Nous envisageons donc diverses modifications, susceptibles d'en améliorer les performances. D'une part, l'usage de noyaux plus flexibles que l'usuel noyau gaussien ouvre une direction prometteuse. D'autre part, des stratégies locales, déjà envisagées dans la littérature, sont en cours d'investigation et peuvent fa-

TAB. 2 – Data Set A : Comparaison des performances des prédicteurs sur les ensembles d'épreuve (observations 1001-1100, 2181-2280, 3871-3970, 4001-4100, 5181-5280). Le résultat pour la prédiction à un pas est présenté pour les observations 1001 à 1100. Les autres résultats sont obtenus par prédictions itérées.

Réf.	Méthode	Préd. un pas	1001-1100
	SVM Gauss	$9.20 \cdot 10^{-3}$	$4.46 \cdot 10^{-2}$
[10]	Local Linear	-	$7.7 \cdot 10^{-2}$
[11]	Neural Net	-	$6.6 \cdot 10^{-2}$
[13]	Neural Net	$\mathbf{2.76 \cdot 10^{-3}}$	$\mathbf{1.94 \cdot 10^{-2}}$
[14]	Neural Net	$2.3 \cdot 10^{-2}$	$2.73 \cdot 10^{-2}$

Réf.	2181-2280	3871-3970	4001-4100	5181-5280
	$3.93 \cdot 10^{-1}$	$4.31 \cdot 10^{-1}$	$\mathbf{8.89 \cdot 10^{-4}}$	$4.81 \cdot 10^{-2}$
[10]	$1.74 \cdot 10^{-1}$	$1.83 \cdot 10^{-1}$	$6.0 \cdot 10^{-3}$	$1.11 \cdot 10^{-1}$
[11]	$\mathbf{6.1 \cdot 10^{-2}}$	$\mathbf{8.6 \cdot 10^{-2}}$	$4.79 \cdot 10^{-1}$	$\mathbf{3.8 \cdot 10^{-2}}$
[14]	$6.5 \cdot 10^{-2}$	$4.87 \cdot 10^{-1}$	$2.3 \cdot 10^{-2}$	$1.6 \cdot 10^{-1}$

cilement être incorporées dans notre schéma de prédiction. De plus, une alternative fiable à la recherche systématique pour l'optimisation des valeurs des paramètres libres, permettra de rendre la procédure de prédiction plus efficace et accessible, y compris pour des utilisateurs profanes.

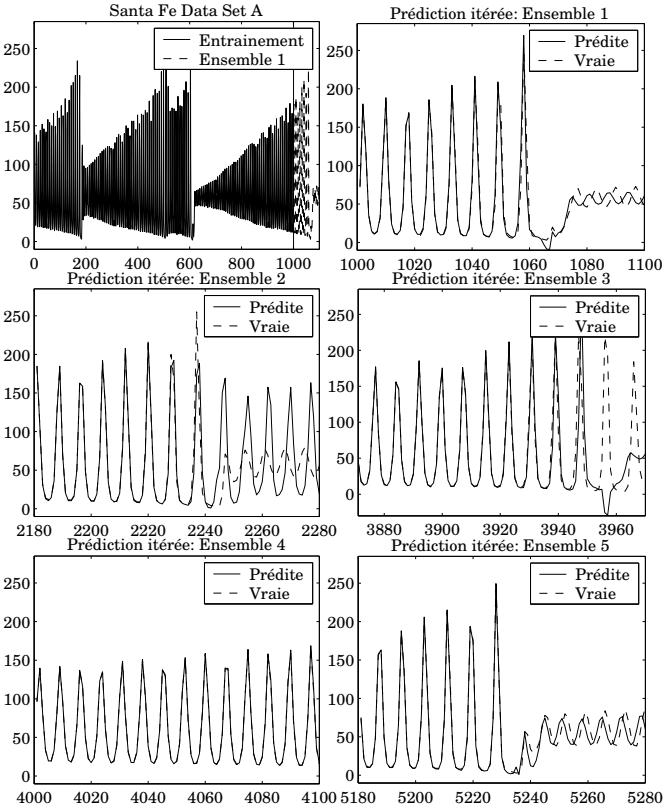

FIG. 2 – **Santa Fe Data Set A :** En haut : à gauche, ensemble d'entraînement et premier ensemble d'épreuve (observations 1001-1100) , à droite, premier ensemble d'épreuve prédit par itération successive. Au centre : les observations 2181-2280 (à gauche) et 3871-3870 (à droite) obtenues par prédiction itérée. En bas : les observations 4001-4100 (à gauche) et 5181-5280 (à droite) obtenues par prédiction itérée.

Références

[1] H.D.I. Abarbanel, *Analysis of observed chaotic data*, first ed., Springer, New York, 1996.

[2] A. Aussem, *Dynamical recurrent neural networks towards prediction and modeling of dynamical systems*, Neurocomputing **28** (1999), no. 3, 207–232.

[3] C. Cortes and V. Vapnik, *Support vector networks*, Machine Learning **20** (1995), 1–25.

[4] R. Fletcher, *Practical methods of optimization*, second ed., John Wiley & Sons Ltd., Chichester, 1987.

[5] T. Joachims, *Making large-scale svm learning practical*, Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges and A. Smola (ed.) (1999), 169–184.

[6] Olivier Michel and Patrick Flandrin, *Application of methods based on higher-order statistics for chaotic time series analysis*, Signal Proc. **53** (1996), no. 2-3, 133–148.

[7] K. Müller, A. Smola, G. Rätsch, B. Schölkopf, O. Kohlmorgen, and V. Vapnik, *Predicting time series with support vector machines*, Artificial Neural Networks - ICANN 97 (M. Hasler W. Gerstner, A. Germond and J.-D. Nicoud, eds.), Springer, 1997.

[8] S. Mukherjee, E. Osuna, and F. Girosi, *Nonlinear prediction of chaotic time series using support vector machines*, IEEE Workshop on Neural Networks for Signal Processing VII (N. Morgan J. Principe, L. Giles and E. Wilson, eds.), IEEE Press, 1997, p. 511.

[9] A.E. Omidvar, *Configuring radial basis function network using fractal scaling process with application to chaotic time series prediction*, Chaos, Sol. and Fract. **22** (2004), no. 4, 757–766.

[10] T. Sauer, *Time series prediction by using delay coordinate embedding*, Time series prediction : Forecasting the future and understanding the past (A.S. Weigend and N.A. Gershenfeld, eds.), Addison-Wesley, 1994, pp. 175–193.

[11] M. Small and C. K. Tse, *Minimum description length neural networks for time series prediction*, Physical Review E **66** (2002), no. 6, 066701.

[12] V. Vapnik, *The nature of statistical learning theory*, Springer Verlag, New York, 1995.

[13] B.W. Wah and M. Qian, *Violation guided neural-network learning for constrained formulations in time-series predictions*, Int'l Journal on Computational Intelligence and Applications **1** (2001), no. 4, 383–398.

[14] E.A. Wan, *Time series prediction by using a connectionist network with internal delay lines*, Time series prediction : Forecasting the future and understanding the past (A.S. Weigend and N.A. Gershenfeld, eds.), Addison-Wesley, 1994, pp. 195–217.

[15] A.S. Weigend and N.A. Gershenfeld (eds.), *Time series prediction : Forecasting the future and understanding the past*, Addison-Wesley, 1994.