

Article

Learning Entropy: Multiscale Measure for Incremental Learning

Ivo Bukovsky

Czech Technical University in Prague, Technicka 4, 166 07, Prague 6, Czech Republic;

E-Mail: ivo.bukovsky@fs.cvut.cz; Tel.: +420-224-352-529; Fax: +420-224-352-674

Received: 26 July 2013; in revised form: 17 September 2013 / Accepted: 22 September 2013 /

Published: 27 September 2013

Abstract: First, this paper recalls a recently introduced method of adaptive monitoring of dynamical systems and presents the most recent extension with a multiscale-enhanced approach. Then, it is shown that this concept of real-time data monitoring establishes a novel non-Shannon and non-probabilistic concept of novelty quantification, *i.e.*, Entropy of Learning, or in short the Learning Entropy. This novel cognitive measure can be used for evaluation of each newly measured sample of data, or even of whole intervals. The Learning Entropy is quantified in respect to the inconsistency of data to the temporary governing law of system behavior that is incrementally learned by adaptive models such as linear or polynomial adaptive filters or neural networks. The paper presents this novel concept on the example of gradient descent learning technique with normalized learning rate.

Keywords: incremental learning; adaptation plot; multiscale; learning entropy; individual sample learning entropy; approximate learning entropy; order of learning entropy; learning entropy of a model; non-Shannon entropy; novelty detection; chaos; time series; HRV; ECG

Nomenclature

LE	Learning Entropy
ALE	Approximate Learning Entropy
ISLE	Individual Sample Learning Entropy
AISLE	Approximate Individual Sample Learning Entropy
OLE	Order of Learning Entropy
LEM	Learning Entropy of a Model
ApEn	Approximate Entropy (by Pincus)
SampEn	Sample Entropy (by Pincus)

AP	Adaptation Plot
GD	Gradient Descent

1. Introduction

Prediction and novelty detection of dynamical system behavior is a vital topic today. Time series are common representatives of observed behavior of complex dynamical systems and the non-stationarity and perturbations are the real-world drawbacks. Such variously caused novelties of newly measured samples of data affect the prediction accuracy and thus they can affect, e.g., control, diagnostics, medical treatment accuracy, and can interfere with many other signal processing objectives.

It has been shown more recently in [1,2] that novelty of individual samples of time series or even intervals of behavior of complex, high-dimensional and nonlinear dynamical systems can be efficiently monitored by relatively simple adaptive models (*i.e.*, low-dimensional neural network architectures). By real-time adaptation of such short-term predictors and by observing also the behavior of adapted parameters (neural weights), we are able to cognitively monitor and evaluate every new measured sample or even whole intervals of behavior with varying complexity (e.g., varying levels of chaos, noise, perturbations). Therefore, in this approach every new measured sample of data is evaluated with respect to its consistency to temporary governing law (dynamics) of a system, which is different from common statistical measures and furthermore, different from entropy based approaches that do not consider consistency of data with the governing law of behavior of data. Moreover, the cognitive approach presented in this paper is also different from existing methods of novelty detection that use learning systems, because this approach does not operate with residuals of the learning system. The terms adaptation, learning and incremental learning can be understood to be equal for clarity of explanations in this paper. However, the learning process of a learning system is generally understood to be a more complex cognitive process than just a parameter adaptation technique [3,4]. The novel concept of entropy in this paper is not limited only to the supervised adaptation, but the principal is applicable to any learning systems in general.

In literature, two fundamental streams of evaluating the entropy of data in dynamical systems in the sense of information contents (novelty) that is carried by measured samples of data can be tracked down, *i.e.*, the probability based approaches, e.g., [5] and the learning system based approaches, e.g., [6].

The first (probabilistic) stream is represented by the statistical approaches of novelty measures and by probabilistic approaches for evaluation of entropy. The Sample Entropy (SampEn) and the Approximate Entropy (ApEn) are the very typical and relevant examples [7,8]. These approaches are closely related to the multi-scale evaluation of fractal measures as discussed in [9–12] and thus to the power-law [13] concept, which is also the partial inspiration for the presented matter in this paper. The usefulness of multi-scale approach is also apparent from the coarse-graining based multi-scale extensions to SampEn in [14,15] and its further and very recent extension in [16]. Some more case studies utilizing SampEn, ApEn, and Multiscale Entropy (MSE) can be found in [17,18]. Another probabilistic approach to the evaluation of entropy as the conditional mutual information between present and past states is proposed as the Compensated Transfer Entropy in [19]. Work [20] can be referenced for the fault detection using a probabilistic entropy approach, and a probabilistic entropy approach to the concept shift (sometimes the concept drift) detection in sensory data is reported in [21].

The second stream is represented by the utilization of learning systems such as neural networks and fuzzy-neural systems, and this is also the main area relevant to the presented work in this paper. During the last three decades of 20th century, the works that were in focus regarding learning systems are [22–25], and for incremental learning approach can be referenced for example also the work [26]. Then, a particularly focused approach toward the utilization of learning systems has been rising with works [27–29], where nonlinear estimators and learning algorithm were utilized for the fault detection via the proposed utilization of a fault function that evaluates behavior of residuals of a learning system. Currently, significant research that shall be referenced is adaptive concept drift detectors proposed in [30–32] and the cognitive fault diagnosis system for sensory data published in [33]. Some readers might also see some analogies of the proposed approach in this paper to the Adaptive Resonance Theory [34]. Because, the proposed approach in this paper utilizes a memory of data behavior, which is represented by the online learning parameters of a learning system, and the unusual behavior of incrementally learning parameters is quantified and introduced as the novel entropy concept in this paper.

Up to the best of my knowledge, I am not aware of any works by other authors on non-probabilistic approaches for evaluation of entropy which are, in their very principal, free from any use of output residuals of a learning model; so that only the behavior of incrementally learning parameters of even imprecise learning models would serve for novelty evaluation in sense of information contents quantification (entropy).

This paper introduces novel concept of entropy and its calculation that neither is based on statistical approaches nor is it based on evaluation of error residual. This new approach operates only on parameter space of incrementally learning systems. The presented principle is purely based on evaluation of unusual behavior of incrementally learning parameters of a pre-trained model, regardless the error residual, *i.e.*, in principle regardless the prediction error itself or its behavior in time. This paper demonstrates the novel approach on Gradient Descent (GD) adaptation that is one of the most comprehensible incremental learning techniques. The very original and funding principals and some related results with Adaptation Plot (AP) have been published in [1,2,38,39] and the first multi-scale extension was proposed in [40]; those are the funding concepts of *Learning Entropy* (LE) and the *Approximate Individual Sample Learning Entropy* (AISLE) that are introduced in this paper.

The paper is organized as follows: the second section recalls two fundamental principles (techniques) that are necessary for evaluating the LE, *i.e.*, GD—a comprehensible example of an incrementally learning technique, and the technique of visualization of learning energy, *i.e.*, the AP. The third section derives the calculation of the novel measure of learning activity, *i.e.*, the measure of learning energy that an incrementally learning model displays for each newly measured sample of data and, thus, the principle of the *Individual Sample Learning Entropy* (ISLE). Then, a practical cumulative-sum technique for estimation of ISLE is introduced as the *Approximate Individual Sample Learning Entropy*. Consequently, the concept of the *Order of Learning Entropy* is introduced according to the order of the estimated time derivative of neural weights that serve to calculate the LE.

The fourth section shows experimental demonstrations including real-world data application of AISLE. The fifth section discusses results, furthermore, theoretical and practical aspects of LE, and it also discusses the fact that LE is not necessarily correlated to the magnitude of prediction error.

In terms of mathematical notation, variables are denoted as follows: small caps as “ x ” for a scalar, bold “ \mathbf{x} ” for a vector and, bold capital “ \mathbf{X} ” for a matrix. Lower indexes as in “ x_i ” or “ w_i ”, indicate the

element position in a vector. If a discrete time index is necessary to be shown, it comes as “ k ” in round brackets such as $\mathbf{x}(k)$, y denotes measured time series and \tilde{y} stands for a predictor output. Further notation, as such \mathbf{w} , represents a vector that contains all adaptable parameters, *i.e.*, weights of a predictor and, $\Delta\mathbf{w}$ is a vector of all adaptive weight increments that are the cornerstone quantities for evaluation of LE by incrementally learning models. The meaning of other symbols is given at their first appearance throughout the text. Time series of constant sampling are considered.

2. Funding Principles

This section reviews two fundamental principles for the latter introduced LE concept. The very fundamental principle is the supervised incremental learning of predictive models, *i.e.*, sample-by-sample adaptation of adaptive parameters (neural weights) to the evaluated signal. As the very cornerstone approach, the GD (incremental) learning is recalled such as for linear or polynomial predictors and neural networks. The second fundamental principle is the binary-marker visualization of how much must the (initially pre-trained) predictor adapt its weights to each sample of data to capture contemporary governing law, *i.e.*, this is the technique of the AP [1–2].

2.1. Predictive Models and Adaptive Learning

Though not limited to, the GD adaptation algorithm is the most fundamental technique for the evaluation of LE. Moreover, GD learning is very efficient especially when used with linear filters or low-dimensional neural network architectures (predictors). The use of GD is recalled particularly for linear predictors (filters) and for polynomial predictors (also called Higher-Order Neural Units HONUs [1,35–37]) in this subsection.

As for time series, let us consider the representation of a general prediction scheme as follows;

$$\tilde{y}(k+h) = f(\mathbf{x}(k), \mathbf{w}), \quad (1)$$

where $\tilde{y}(k+h)$ denotes predicted value at prediction horizon of h samples; $f(\cdot)$ is a general differentiable function (linear or polynomial predictor or a neural network) mapping the input vector $\mathbf{x}(k)$ to the predicted output; vector \mathbf{w} contains all adaptable parameters (weights) of a predictor.

To unify the terminology about general predictors (1) that are used for the purpose of LE, the following lemmas are given:

L1. Predictor (1) is a static model that performs direct prediction if input vector $\mathbf{x}(k)$ contains only the recent history of measured data.

L2. Predictor (1) is a dynamical (recurrent) model that performs indirect prediction if input vector $\mathbf{x}(k)$ contains also step-delayed values of \tilde{y} .

L3. Dimensionality of predictor (1) corresponds to the state space dimension for which the mapping $f(\cdot)$ is defined (for Equation (1) this relates to the numbers of inputs in \mathbf{x} including step-delayed feedbacks of \tilde{y} if a predictor is dynamical one).

L4. Dimensionality of a real data corresponds to the order of dynamics of real-data-generating system and it is further extended by other real inputs (that further increase dimensionality of behavior of real data).

L5. A low-dimensional predictor of time series is such predictor that is considerably less-dimensional than the embedding dimension of the time series itself.

Sample-by-sample GD adaptation scheme of predictor (1) can be defined using prediction error e , which is given as:

$$e(k) = y(k) - \tilde{y}(k), \tag{2}$$

so the individually adapted weight increments are calculated in order to decrease the square error criteria as follows:

$$\Delta w_i(k) = -\frac{1}{2} \cdot \mu \cdot \frac{\partial e(k)^2}{\partial w_i} = \mu \cdot e \cdot \frac{\partial \tilde{y}(k)}{\partial w_i} = \mu \cdot e \cdot \frac{\partial f(\mathbf{x}(k-h), \mathbf{w})}{\partial w_i}, \tag{3}$$

where, $\Delta w_i(k)$ is an adaptive weight increment of i^{th} weight, μ is the learning rate, and k is the discrete index of time that also denotes the reference time, *i.e.*, $e(k)$ is currently measured error, $\tilde{y}(k+h)$ is predicted value h samples ahead. For completeness, the updates of all weights at each sample time k can be in its simplest form (no momentum or regularization term) as follows:

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \Delta \mathbf{w}(k). \tag{4}$$

Recall, stability of the reviewed GD algorithm (2–4) can be practically improved by proper scaling input and output variables (e.g., z-score) and by various approaches for rescaling the learning rate μ (e.g., [41,42]).

In case the predictor is a dynamic model, *i.e.*, lemma L2, we may alternatively refer to GD as to Real Time Recurrent Learning (RTRL) technique [43] if the above GD scheme (1–4) is applied with recurrently calculated derivatives for feedback elements in input vector \mathbf{x} .

Equation (1) gives only a general form of a predictor for LE. The particular form of the mapping $f(\cdot)$ and configuration of inputs and feedbacks in input vector \mathbf{x} as in Equation (1), as well as the proper sampling period, are all case specific.

Nevertheless, it can be reasonable to start with linear adaptive filters as they are simplest and computationally efficient especially when vector \mathbf{x} should contain relatively higher number of inputs ($\sim > 20$). In case of linear filters, the predictor (1) yields the vector multiplication form of row vector \mathbf{w} and column vector \mathbf{x} as follows:

$$\tilde{y}(k+h) = \mathbf{w}(k) \cdot \mathbf{x}(k), \tag{5}$$

The weight updates are directly calculated for a linear model as follows:

$$\Delta \mathbf{w}(k) = \mu \cdot e(k) \cdot \mathbf{x}(k)^T, \tag{6}$$

where, T denotes vector transposition and where recurrently calculated partial derivatives (as if according to RTRL) are neglected, *i.e.*, $\partial \mathbf{x} / \partial \mathbf{w} = \mathbf{0}$.

As regards selection of learning rate μ , the first technique that should be considered is the learning rate normalization that practically improves the stability of the weight update system (4, 6), so the weight updates can be actually calculated at every sample time as follows:

$$\Delta \mathbf{w}(k) = \frac{\mu}{1 + \|\mathbf{x}(k)\|} \cdot e(k) \cdot \mathbf{x}(k)^T, \tag{7}$$

where “ $\|\cdot\|$ ” denotes a vector norm \square more on techniques for learning rate normalization and adaptation of a regularization term (the unit in the denominator) can be found, e.g., in [41,42].

For evaluation of LE of nonlinear time series, polynomial adaptive predictors such as Higher-Order Neural Units can be recommended [1,35–37]; HONUs are attractive adaptive models because their mapping is customizable as nonlinear while they are linear in parameters => optimization of HONUs is of a linear nature, so HONUs do not suffer from local minima problem in the way as conventional neural networks do when GD learning technique is used.

Of course, because of various systems and according to various user experience, other types of predictors such as perceptron neural networks or any other kind of adaptive models, suitable for GD (but not limited to GD) adaptation, can be used as a cognitive (here the supervised), incrementally learning tool, for evaluation of LE. This subsection recalls the GD rule as a straightforward example of incremental learning technique, that is a comprehensible option for AP and latter for calculation of LE.

2.1. Adaptation Plot (AP)

The variability of weight increments $\Delta \mathbf{w}$ in (3) resp. (7) reflects the novelty of data that corresponds to the difficulty with which an adaptive model learns to every new sample of data. Therefore, the AP was introduced for GD and adaptive models (HONUs) in [1] and further in [2,38,39] as a visualization tool of adaptation activity (or novelty in data) of adaptive predictors.

AP is based on evaluation and visualization of unusual weight increments of sample-by-sample GD adapted models. It was shown through [1,2,38–40] that low-dimensional predictors can capture and evaluate important signal attributes. As such, unusual samples, very decent perturbations, unusual appearance or variations of level of chaos or noise, incoming inter-attractor transitions of hyper-chaotic systems, also hidden repeating patterns can be revealed and intervals of a similar level of chaos can be revealed in otherwise seemingly, similarly complicated signals.

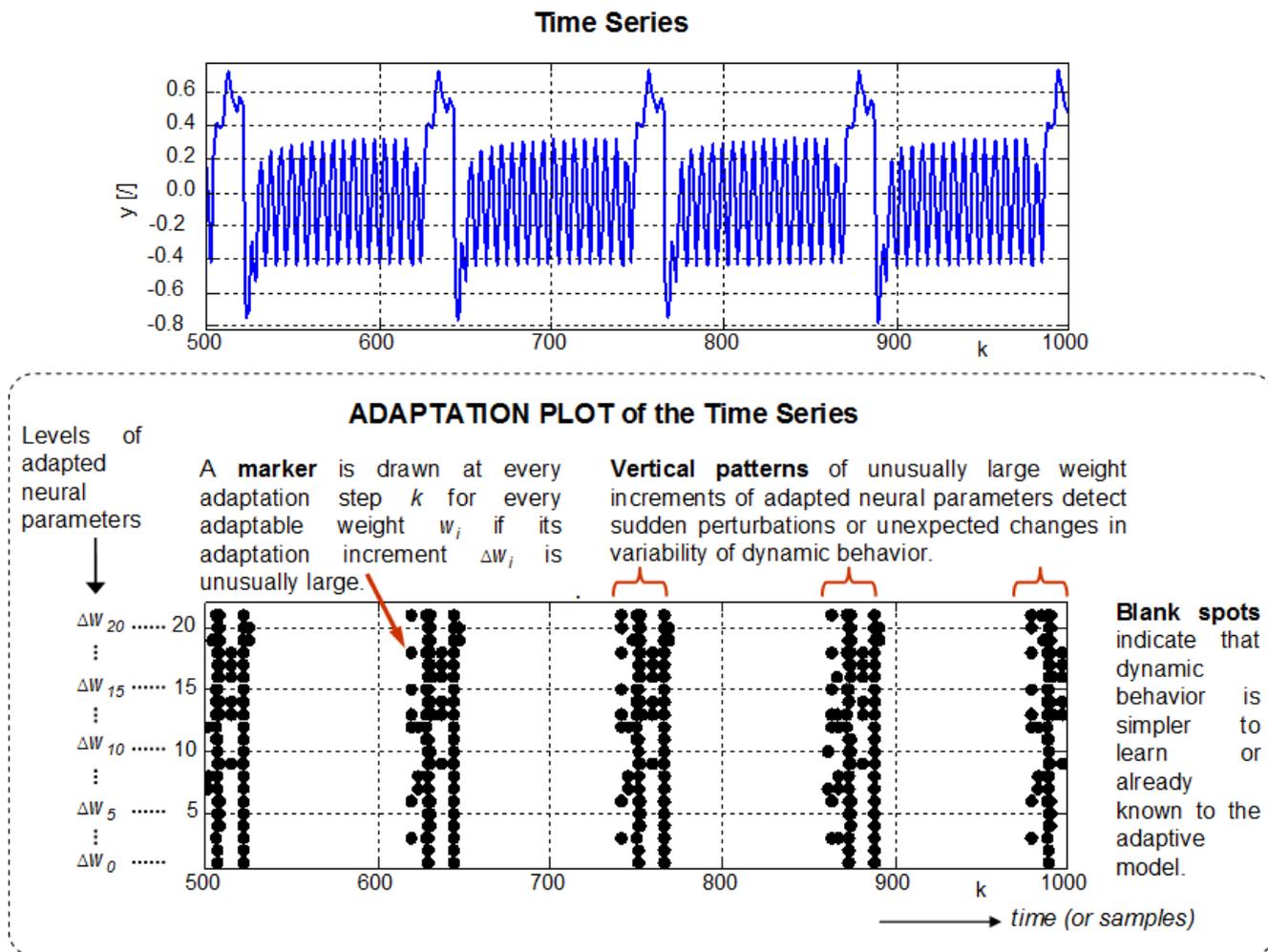
To clarify the principle of AP, the sensitivity parameter α for marker detection of AP has to be recalled. A governing law variability marker (a dot) in AP (Figure 1) is drawn at every sampling time k , if the corresponding weight increment exceeds its contemporary, usual magnitude, which can be in principle sketched by the following rule:

$$\text{if } \left(|\Delta w_i(k)| > \alpha \cdot \overline{|\Delta w_i(k)|} \right) \Rightarrow \text{draw a marker for weight } w_i \text{ at time } k, \quad (8)$$

where, α is the detection sensitivity parameter, and $\overline{|\Delta w_i(k)|}$ is a floating average of absolute values of recent m neural increments of i^{th} neural weight as follows:

$$\overline{|\Delta w_i(k)|} = \frac{1}{m} \cdot \left(\sum_{j=k-m+1}^k |\Delta w_i(j)| \right). \quad (9)$$

Figure 1. Adaptation Plot (AP) is a tool for universal evaluation of information hidden in adapted neural weights via transformation into a binary space (time series is cognitively transformed to patterns of binary features — AP markers (the dots)), for more on functionality of AP please see [1,2].



The mutually alternative explanations of the sensitivity α are as follows:

- The larger value of α , the larger magnitudes of weight increments (i.e., $|\Delta w|$) are considered to be unusual.
- The larger α , the more unusual data samples in signal are detected and visualized in AP.
- The larger α , the less sensitive AP is to data that do not correspond to the contemporary dynamics learned by a model.
- The larger α , the lower density of markers in AP.

The major single-scale weakness of AP is the need for manual tuning of the sensitivity parameter α , so the first multi-scale solution to AP has been proposed in [40] without connotations to any entropy concept.

In the next section, it is proposed that the multi-scale solution to novelty detection via AP and over a whole range of sensitivity detection establishes a novel entropy concept of LE.

3. Learning Entropy (LE)

In this section, the concept of Learning Entropy (LE) is introduced for supervised incremental learning. This novel entropy concept can utilize sample-by-sample adaptation of low-dimensional predictors [1,2,36–39] and uses the technique of the AP. Notice, the GD with normalized learning rate (e.g., [41,42]) is used in this paper for its clarity and for its good performance; however, LE is not principally limited to only GD technique, nor to supervised techniques in general.

In fact, LE is a cognitive entropy measure concept because the cognitively obtained knowledge about variation of temporary governing laws of the evaluated data is utilized.

Important distinction of this concept is that if a system behavior is very complex from statistical point of view, but it is deterministic from the point of view of its governing law, the information content (complexity, entropy) of the data is lower, the more deterministic the behavior is (deterministic chaos, forced nonlinear (chaotic) oscillator).

For example, if a predictive model can adapt fully to a governing law of deterministically chaotic time series, then the further data of time series have no new information to us (the new data are redundant because we know a governing law). However, if a deterministic (chaotic) time series becomes perturbed, the perturbed data (samples or intervals) have new information, *i.e.*, novel data have entropy that can be adaptively (cognitively) detected (e.g., by supervised GD learning).

3.1. Individual Sample Learning Entropy (ISLE)

In this subsection, LE is approached via GD (supervised learning) and it is demonstrated on the example of deterministically chaotic time series obtained from Mackey-Glass equation [44] in chaotic mode as particularly given in Equation (10):

$$\frac{dy(t)}{dt} = b \cdot y(t-\tau) \cdot \left(1 + y(t-\tau)^{10}\right)^{-1} - g \cdot y(t), \quad (10)$$

where t denotes the continuous index of time, and the chaotic behavior results from the setup of; $b = 0.2$, $g = 0.1$, and the lag $\tau = 17$. The time series was generated by Equation (10) and data were sampled with period $\Delta t = 1$ [time unit] as $\{y(k); k = 0, 1, \dots, 700\}$ where k denotes the discrete time index.

Lets introduce 1% perturbation at sample $k = 500$ as follows:

$$y(500) = y(500) + 0.01 \cdot y(500). \quad (11)$$

The time series with the detail of the perturbed sample at $k = 500$ is shown in Figure 2. First 200 samples is used to pre-train a low-dimensional predictive model (given random initial weights \mathbf{w}) by GD in 300 epochs. Then, the adaptive model runs adaptation only once on further data $k = 200 \dots 700$.

As a nonlinear and low-dimensional predictive model, static quadratic neural unit (QNU, [1,35,36]) is chosen for its good quality of nonlinear approximation and in-parameter linearity that theoretically avoids problem of local minima for adaptation [37]. QNU can be expressed in a long-vector multiplication form as follows:

$$\tilde{y}(k) = \sum \left\{ w_{i,j} \cdot x_i \cdot x_j ; i = 0 \dots n, j = i \dots n, x_0 = 1 \right\} = \mathbf{w} \cdot \mathbf{colx}, \quad (12)$$

where, the length of recent history of time series in input vector \mathbf{x} is chosen as $n = 4$ as follows:

$$\mathbf{x} = [x_0=1 \ x_1 \ \dots \ x_n]^T = [1 \ y(k-1) \ y(k-2) \ y(k-3) \ y(k-4)]^T \tag{13}$$

where, T stands for vector transposition and \mathbf{colx} is the long column vector representation of quadratic multiplicative terms that are pre-calculated from \mathbf{x} as follows:

$$\mathbf{colx} = \{x_i \cdot x_j ; i = 0 \dots n, j = i \dots n, x_0 = 1\} = [x_0^2 \ x_0x_1 \ x_0x_2 \ \dots \ x_ix_j \ \dots \ x_n^2]^T \tag{14}$$

Furthermore, \mathbf{w} is a row weight vector of the same length as \mathbf{colx} .

The sample-by-sample updates of \mathbf{w} , can then be calculated by the GD according to (1–6) as follows:

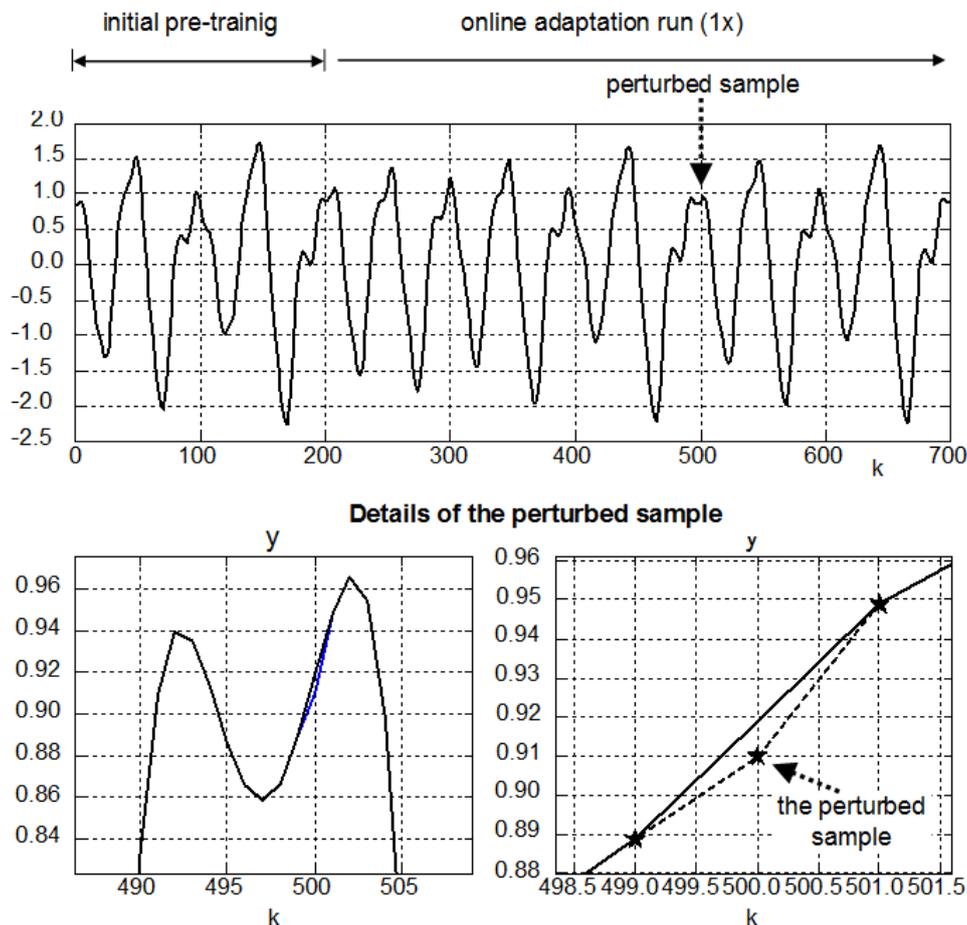
$$\Delta \mathbf{w}(k) = \mu \cdot e(k) \cdot \mathbf{colx}(k)^T . \tag{15}$$

In regards to selection of the learning rate μ , a variation of the learning rate normalization (e.g., [41]) practically improves the stability of the weight update system (6), so the weight updates Equation (15) can be actually calculated at every sample time as follows:

$$\Delta \mathbf{w}(k) = \frac{\mu}{1 + \mathbf{colx}(k)^T \cdot \mathbf{colx}(k)} \cdot e(k) \cdot \mathbf{colx}(k)^T . \tag{16}$$

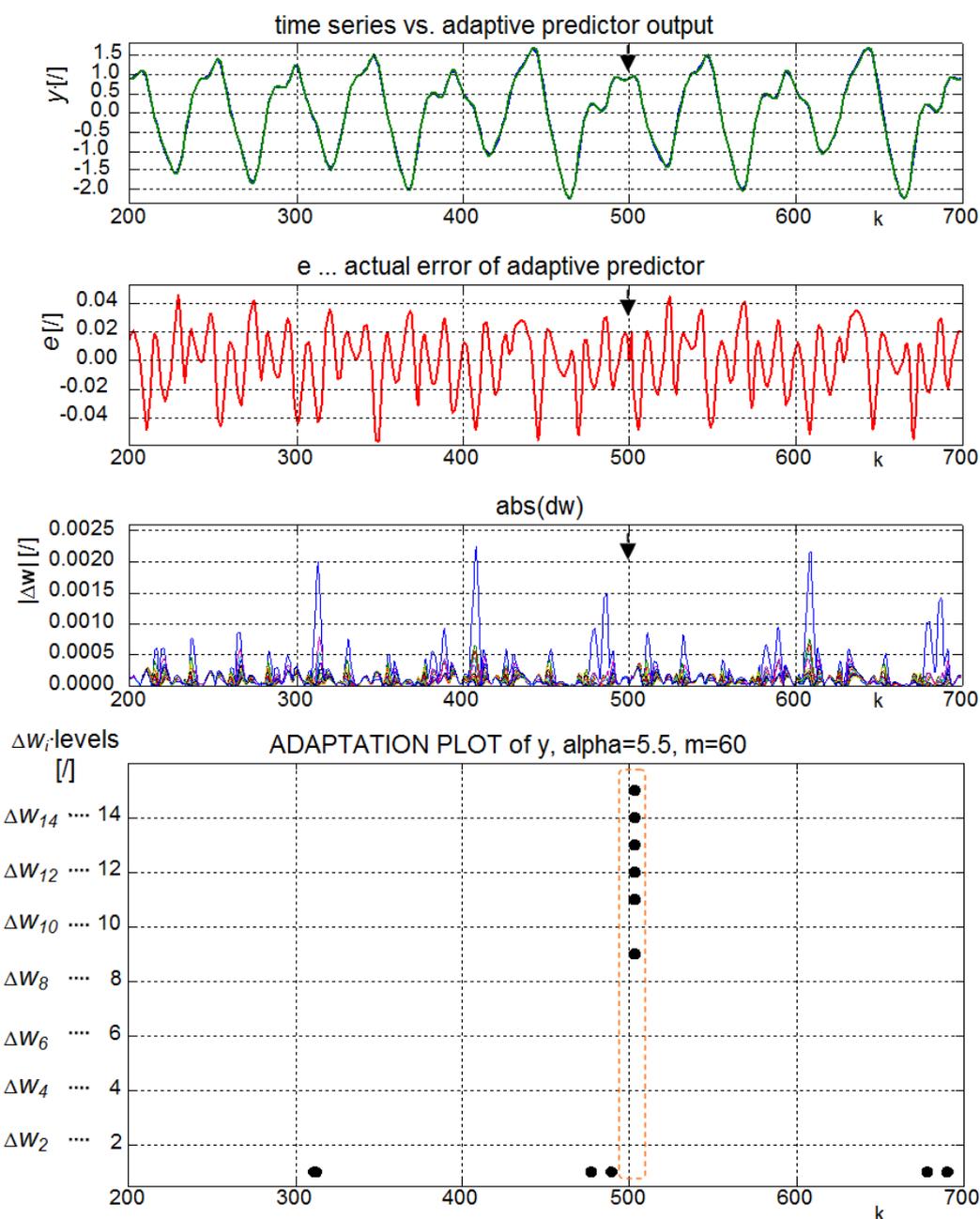
where $\mu = 1$ is used in following experiments for initial pre-training, and $\mu = 0.1$ is used for adaptive detection online.

Figure 2. Mackey-Glass time series in chaotic mode (10) with the detail of perturbation (11) at $k = 500$.



For the time series shown above and for the given setup, the QNU has 15 neural weights and the corresponding AP is shown in Figure 3. The AP in the bottom axis of Figure 3, shows that the six weight increments Δw_9 and Δw_{11-15} of adaptive model were unusually large for the sensitivity $\alpha = 5.5$, *i.e.*, the pre-trained adaptive model (12) captured the perturbed sample at $k = 500$, while the prediction error $e(k = 500:503)$ was of even a smaller than usual magnitude. Therefore, the markers in AP, visualize, activity in which the model learns to each newly measured sample, even when the adaptive model is not absolutely accurate (for another example please see Figure 13 in [40]).

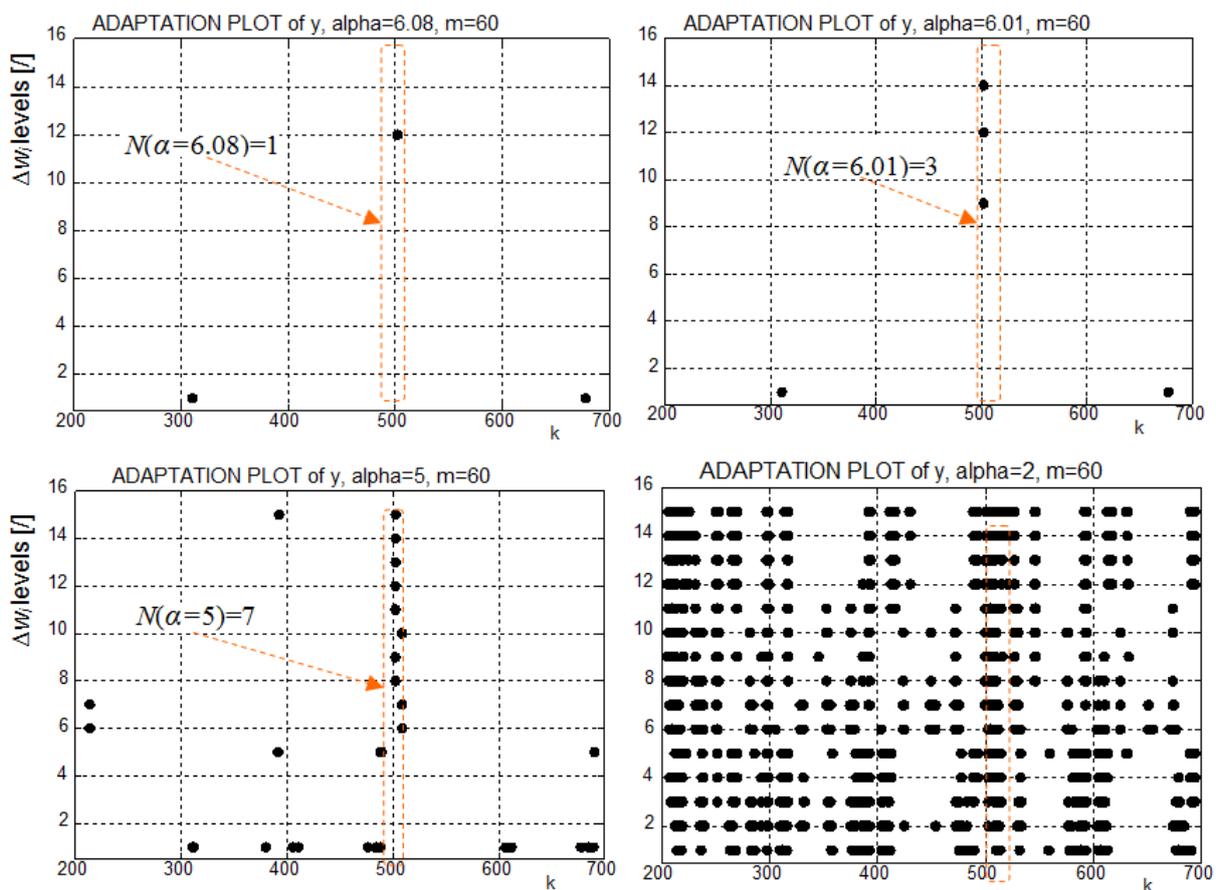
Figure 3. The bottom axis is the AP with manually tuned α for perturbed time series in Figure 2. The six AP markers for weights w_9 and w_{11-15} at $k = 503$ correspond to the perturbed sample at $k = 500$ (11) (while the above magnitudes of prediction error and the adaptation weight increments do not indicate anything at first sight).



The particular time series in top axes in Figure 3 has a significant frequency component of about 60 samples and the most of the signal events chaotically recur within this interval. Therefore, parameter m that calculates recently usual magnitudes of weight increments Equation (9) is pragmatically set to $m = 60$ for this time series in this paper. More discussion on choice of parameter m follows in Section 5 of this paper.

The most critical parameter to obtain a meaningful and useful result with the AP (such as in Figure 3) is the detection sensitivity α . To overcome this single-scale weakness of AP, *i.e.*, the dependence on proper selection of α , a multi-scale approach can be adopted. Naturally, the markers in AP appear according to α , and the more sensitive detection is (*i.e.*, the smaller α), the more markers appear for more unusual samples of data [40], *i.e.*, for samples of higher LE. The dependence of AP markers is demonstrated in Figure 4.

Figure 4. The APs of the chaotic time series (Figure 2) with perturbed sample at $k = 500$ for increasing detection sensitivity (*i.e.*, decreasing $\alpha = 6.08, 6.01, 5, 2$); the number of AP markers $N(\alpha)$ related to perturbed sample at $k = 500$ tends to increase with increasing sensitivity more rapidly than for usual (not novel) samples.



For $\alpha = 6.08$ in Figure 4, there are three AP markers that indicate some unusually large learning activity at $k = 311, 503, 678$. However, for constant α , the detection must be made more sensitive to reveal the perturbed sample ($k = 500$) in contrary to the other seemingly novel points, which is manually made in Figure 4 by redrawing AP for $\alpha = 6.01$ and $\alpha = 5$. For $\alpha = 2$, the detection becomes too sensitive and the AP is not useful anymore.

To become independent of single-scale issue of manual selection of α , the multi-scale approach was proposed in [40]. It is further recalled in this section for the example of the above time series, and it is newly related to the concept of LE.

In order to enable us to perform multi-scale analysis of AP markers, we may consider the power-law concept, e.g., [13]. For AP markers at instant time k as a function of sensitivity parameter α , where the detection sensitivity is increasing with decreasing α , i.e., we can assume theoretical power-law approximation as follows:

$$N(\alpha) \cong (\alpha)^{-H}, \quad \log(N(\alpha)) \cong -H \cdot \log(\alpha), \tag{17}$$

where, the exponent H characterizes the nonlinear change of quantity of AP markers along the varying sensitivity of detection α , and N is the quantity of AP markers (here the vertical sum at instant k as in Figure 4) for the given sensitivity of detection α .

Similarly to common fractal measure approaches, the change of quantity of AP markers along the increasing sensitivity parameter α , can be quantified by estimation of characterizing exponent H as the slope of log-log plot as:

$$H = \lim_{\alpha \rightarrow \alpha_{max}} \left(-\frac{\log(N(\alpha))}{\log(\alpha)} \right), \tag{18}$$

where, α_{max} is a specific (theoretical) value of detection sensitivity for which the very first AP marker would appear for evaluated samples k . Thus, α_{max} can be loosely defined as follows:

$$\sum_k \{N(\alpha \leq \alpha_{max})\} \geq 1; \quad \sum_k \{N(\alpha > \alpha_{max})\} = 0; \quad \sum_k \{N(\alpha_{max})\} \geq 1. \tag{19}$$

For any α arbitrarily close to α_{max} , H becomes large if a measured sample of data is novel, i.e., if data is inconsistent with the governing law that has been temporarily learned by a predictor. For a particularly used predictor (1), the theoretical maximum $H = +\infty$ could be obtained correctly only for those samples of data where all AP markers vertically appear instantly, i.e., if $N(\alpha_{max}) = n$ where $N(\alpha > \alpha_{max}) = 0, \forall \alpha$ and where n is the number of all adaptable weights used for the AP.

If we introduce a new variable E to normalize H as follows:

$$E(k) = \frac{2}{\pi} \cdot \arctan(H(k)) \Rightarrow E \in \langle 0, 1 \rangle, \tag{20}$$

then, we have arrived in Equation (20) to the normalized entropy measure E that quantifies learning activity of a sample-by-sample adaptive models (1–4). Thus, the variable E in Equation (20) is a novel non-Shannon and non-probabilistic measure for evaluation of novelty of each single sample of data in respect to its consistency to the temporary governing law learned by a predictor. Thus, E in Equation (20) can be called the *Individual Sample Learning Entropy (ISLE)*.

ISLE can be evaluated for all samples in a window of AP and consequently even only for a custom selection of particular samples. For example, Table 1 and Figure 5 shows comparison of ISLE for three specific samples of data for which AP markers appeared as the very first for $\alpha_{max} \cong 6.08$ in Figure 4, but where only $k = 503$ has AP markers due to the novelty in data. We can see in Figure 6 that sample at $k = 503$ has much larger E than the other two data samples.

Table 1. The number of AP markers $N(\alpha)$ increases fastest for $k = 503$ because the incremental learning attempts to adapt the weights to the perturbation at $k = 500$ (Figure 2), $\alpha_{max} \cong 6.08$.

α	6.08	6	5	4	3	2	1
$N(\alpha), k = 311$	1	1	1	1	1	2	4
$N(\alpha), k = 503$	1	3	7	12	14	14	15
$N(\alpha), k = 678$	1	1	1	1	1	2	3

Figure 5. The limit slope H (18) for Table 1 is largest when an adaptive model starts learning a new governing law (here the perturbation at $k = 500$) that causes unusually large weight increments (here calculated with normalized learning rate by Equation (16)).

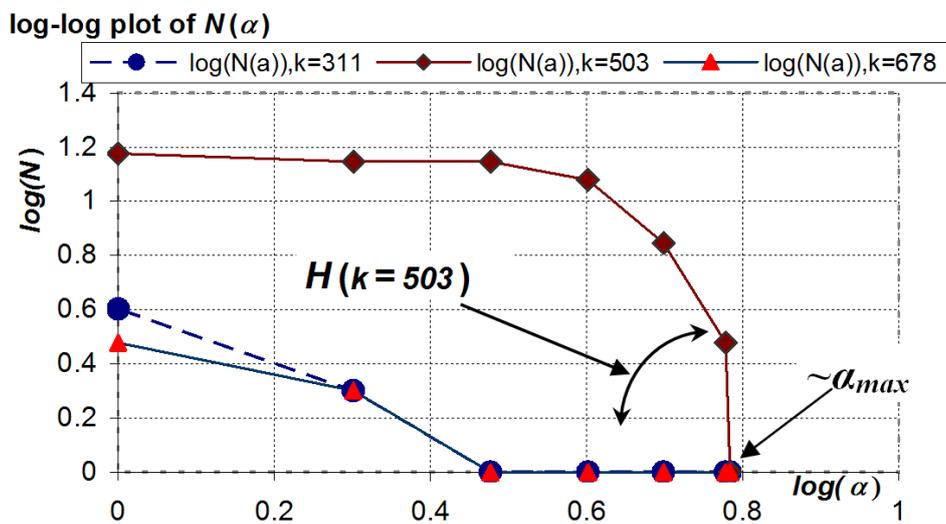
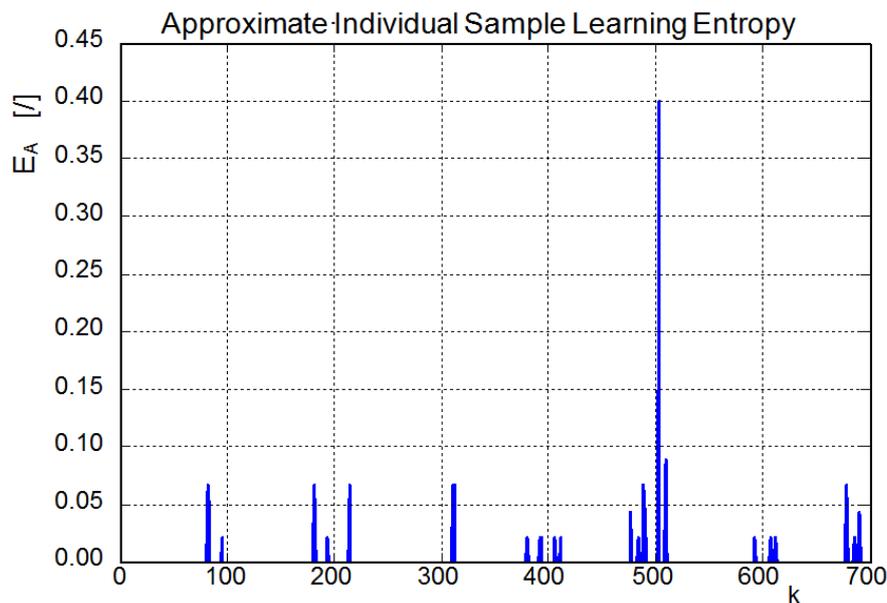


Figure 6. AISLE calculated as E_A by Equation (21) for the signal in Figure 2 and for sensitivities $\alpha = 3 * [1.1^7, 1.1^6, 1.1^5]$, $m = 60, \mu_0 = 1, \mu = 0.1$; the perturbation at $k = 500$ is followed by the rapid increase of LE.



Naturally, if an adaptive model is familiar with a temporary governing law of behavior of data, and if the measured samples of data are consistent with the governing law, then the adaptive model does not need to unusually adapt its parameters (weights) and E is low.

Importantly, evaluation of LE is not conditioned by the fact that the predictor must be precise and perfectly pre-trained. Inversely, if $E = 0$ is only achieved for all samples in an AP window, it does not necessarily imply that a predictor is precisely pre-trained and has zero prediction errors. $E = 0$ would just imply that the adaptive predictor (1) is familiar with data regardless its prediction error in Equation (2). $E = 0$ would mean that weights are constant (rounded to a decimal digit) during adaptation for all k even though the prediction error is not zero; this practically happens often with GD for not too large learning rate μ (this can be practically verified on pre-training data).

In particular, this section demonstrated the calculation of the LE via the sample-by-sample adaptation and it is applicable to every new sample measured and it can detect novelty of individual samples. Therefore, this particular technique by Equations (1–20) results in evaluation of the novelty measure that can be called the *Individual Sample Learning Entropy (ISLE)*.

However, the above estimation of ISLE via E by Equation (20) is rather a theoretical and explanatory matter, because proper estimation of slope H (18) depends on finding of α_{max} . The next subsection resolves this issue.

3.2. Approximate Individual Sample Learning Entropy (AISLE)

A practical technique to approximate E as a normalized measure of ISLE for every newly measured sample y is introduced in this subsection. This approximate technique does not require discovering of proper α_{max} complying strictly with (19). E can be approximated by E_A for every sample of data $y(k)$ as follows:

$$E_A(k) = \frac{1}{n \cdot n_\alpha} \sum \{N(\alpha); \alpha \in \mathbf{\alpha}\}, \quad \mathbf{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_{n_\alpha}], \quad \alpha_1 > \alpha_2 > \dots > \alpha_{n_\alpha}, \quad (21)$$

which is a sum of markers over a range of sensitivities α that is normalized by the number of weights n and by the number of selected sensitivity parameters n_α , thus $1/(n \cdot n_\alpha)$ is a normalization term to achieve $E_A \in (0, 1)$, i.e., n markers can appear n_α times for a sample of data $y(k)$.

For every individual sample, the measure E_A in Equation (21) approximates E in (20), because larger values of E_A , corresponds to a steeper slope H (Figure 5). Thus E_A can be called the *Approximate Individual Sample Learning Entropy (AISLE)*.

Also similarly to E in (20), E_A introduced in Equation (21) is also a normalized measure and its possible maximum value $E_A = \max = 1$ can be obtained if all AP markers for a sample $y(k)$ appear for all detection sensitivities in $\mathbf{\alpha}$. The possible minimum value of E_A of a sample $y(k)$ is zero if no markers appear [similarly as to E in Equation (20)].

For the example given in Table 1, the setup for calculation of E_A is $n = 15$, $n_\alpha = 7$, $\mathbf{\alpha} = [6.08, 6, 5, 4, 3, 2, 1]$, for which E_A results as follows: $E_A(k = 311) = 0.105$, $E_A(k = 503) = 0.629$, $E_A(k = 678) = 0.095$ (see corresponding slopes H in Figure 5). However, to better approximate the limit slope H , the elements of $\mathbf{\alpha}$ shall be selected closer to the approximate neighborhood of α_{max} , i.e., around $\alpha_{max} \approx 6$ while $\alpha = 2$ is already too far from α_{max} , as it is shown in Figure 4. Thus, E_A can be

estimated for α that is reduced to the neighborhood of $\alpha_{max} \cong 6$ and both cases are thus compared by Figure 5 and 6. Even though that the result in Figure 6 is dependent on selection of α , it clearly reveals perturbation of sample $y(k = 500)$.

In Figure 6, we can see that other samples of data show also some smaller LE (AISLE in particular). Those data are not perturbed and they are fully consistent with governing law (10), yet they display nonzero LE. The reason for this is that the predictor (12–16) is a low-dimensional one and so it is not able to fully learn the governing law. However, it clearly detects and evaluates an inconsistent (here the perturbed) sample. In the next subsection, the further extension to the above introduced definition and calculation of LE that improves its evaluation accuracy is proposed.

3.3. Orders of Learning Entropy (OLEs)

When weights of a learning system are adapted by an incremental learning, the weights fluctuate in the weight state space with energy that the weight-update system has.

The weight update system receives its (learning) energy, from the measured data, *i.e.*, from the input vector $\mathbf{x}(k)$ and the target $y(k)$ in case of supervised learning. The more inconsistent newly measured samples $y(k)$ to the current knowledge of the learning system, *i.e.*, the higher LE of the samples, the more energy the weight increments $\Delta\mathbf{w}$ receives.

In other words, the weight update system resembles an engine, with its fuel being the input data. Then, the LE is the actual (time-varying) octane number of the fuel.

Weight increments $\Delta\mathbf{w}$ are the key variables for LE. During incremental learning, each weight w_i behaves with energy of various orders that can be defined for the AP and thus for evaluation of the LE as follows:

- Order learning energy of weight w_i corresponds to exceeding the floating average of its m recent magnitudes $\overline{|w_i(k)|} = \frac{1}{m} \cdot \sum_{j=k-m}^{k-1} |w_i(j)|$,
- 1st Order learning energy of w_i corresponds to exceeding the floating average of its m recent first derivative magnitudes $\left| \frac{dw_i(k)}{dt} \right| \approx |\Delta w_i(k)|$ (this is the case of rule (8)),
- 2nd Order learning energy of w_i corresponds to exceeding the floating average of its m recent second order derivative magnitudes $\left| \frac{d^2 w_i(k)}{dt^2} \right| \approx |\Delta^2 w_i(k)| = |\Delta w_i(k) - \Delta w_i(k-1)|$, see (22), and similarly,
- 3rd Order learning energy of w_i relates to $\left| \frac{d^3 w_i(k)}{dt^3} \right| \approx |\Delta^3 w_i(k)| = |\Delta w_i(k) - 2 \cdot \Delta w_i(k-1) + \Delta w_i(k-2)|$,
- 4th Order learning energy of w_i to $\left| \frac{d^4 w_i(k)}{dt^4} \right| \approx \Delta^4 w_i(k) = |\Delta w_i(k) - 3 \cdot \Delta w_i(k-1) + 3 \cdot \Delta w_i(k-2) - \Delta w_i(k-3)|$
- etc.

From the above point of view, the originally introduced rule of AP (8) can be extended for the second order LE as follows:

$$\text{if } \left(\left| \Delta w_i(k) - \Delta w_i(k-1) \right| > \alpha \cdot \left| \overline{\Delta^2 w_i(k)} \right| \right) \Rightarrow \text{draw a marker for weight } w_i \text{ at time } k, \tag{22}$$

where, the recently usual second derivative (acceleration) of weights is as:

$$\left| \overline{\Delta^2 w_i(k)} \right| = \frac{1}{m} \left(\sum_{j=k-m}^{k-1} \left| \Delta w_i(j) - \Delta w_i(j-1) \right| \right), \tag{23}$$

Similarly, for the 3rd order LE it would be as follows:

$$\left| \overline{\Delta^3 w_i(k)} \right| = \frac{1}{m} \left(\sum_{j=k-m}^{k-1} \left| \Delta^2 w_i(j) - \Delta^2 w_i(j-1) \right| \right), \tag{24}$$

and so on for higher orders.

To distinguish among the above modifications in which the LE is calculated via adaptation plot rule as shown in Equation (8) or Equation (22), the *Order of Learning Entropy* (OLE) is introduced in this section and its most common cases are summarized in Table 2, where the details of the formulas has been indicated above in this section.

Table 2. Orders of LE (OLE) and Corresponding Detection Rules, see Equations (8,9,22–24).

OLE	Notation	Detection Rule for AP Markers
0	E^0, E_A^0	$ w_i(k) > \alpha \cdot \overline{ w_i(k) }$
1	E^1, E_A^1	$ \Delta w_i(k) > \alpha \cdot \overline{ \Delta w_i(k) }$
2	E^2, E_A^2	$ \Delta^2 w_i(k) = \Delta w_i(k) - \Delta w_i(k-1) > \alpha \cdot \overline{ \Delta^2 w_i(k) }$
3	E^3, E_A^3	$ \Delta^3 w_i(k) = \Delta^2 w_i(k) - \Delta^2 w_i(k-1) > \alpha \cdot \overline{ \Delta^3 w_i(k) }$
4	E^4, E_A^4	$ \Delta^4 w_i(k) = \Delta^3 w_i(k) - \Delta^3 w_i(k-1) > \alpha \cdot \overline{ \Delta^4 w_i(k) }$

Figure 7 and Figure 8 show the results of AISLE for the above first five Orders of Learning Entropy estimates for data in time series (10), this time with two perturbations as follows:

$$y(k=475) = y(k=475) - .05 \quad \text{and} \quad y(k=500) = y(k=500) - 0.05 \tag{25}$$

Figure 7 demonstrates the impact of various LE orders as they can significantly improve detection of inconsistent samples of data. Moreover, Figure 7 and Figure 8 also demonstrates that Zero-Order Learning Entropy, which deals just with weights themselves, does not have the cognitive capability to evaluate the learning effort of the predictor, *i.e.*, E_A^0 does not detect the unusual samples at $k = 475, 500$, nor the AP (bottom Figure 8) reflects inconsistent data.

This subsection introduced Orders of LE as they relate to the time derivatives orders of adaptable parameters w_i . It was demonstrated that useful LE Orders are especially starting from 1st Order and higher (Figure 7) which is consistent to the results of experiments that were made through recent years with the AP [1,2,38–40].

Figure 7. AISLE of Orders of time series (10) with two perturbations of magnitude 0.05 at $k = 475$ and $k = 500$, $\alpha = [15, 14, \dots, 1]$, $m = 60$, $\mu_0 = 1$, $\mu = 0.1$; the zero order AISLE shown in top axes is not capable to capture the inconsistent data at $k = 475, 500$ (see the E_A^0 in Figure 8), the higher orders can improve novelty detection significantly.

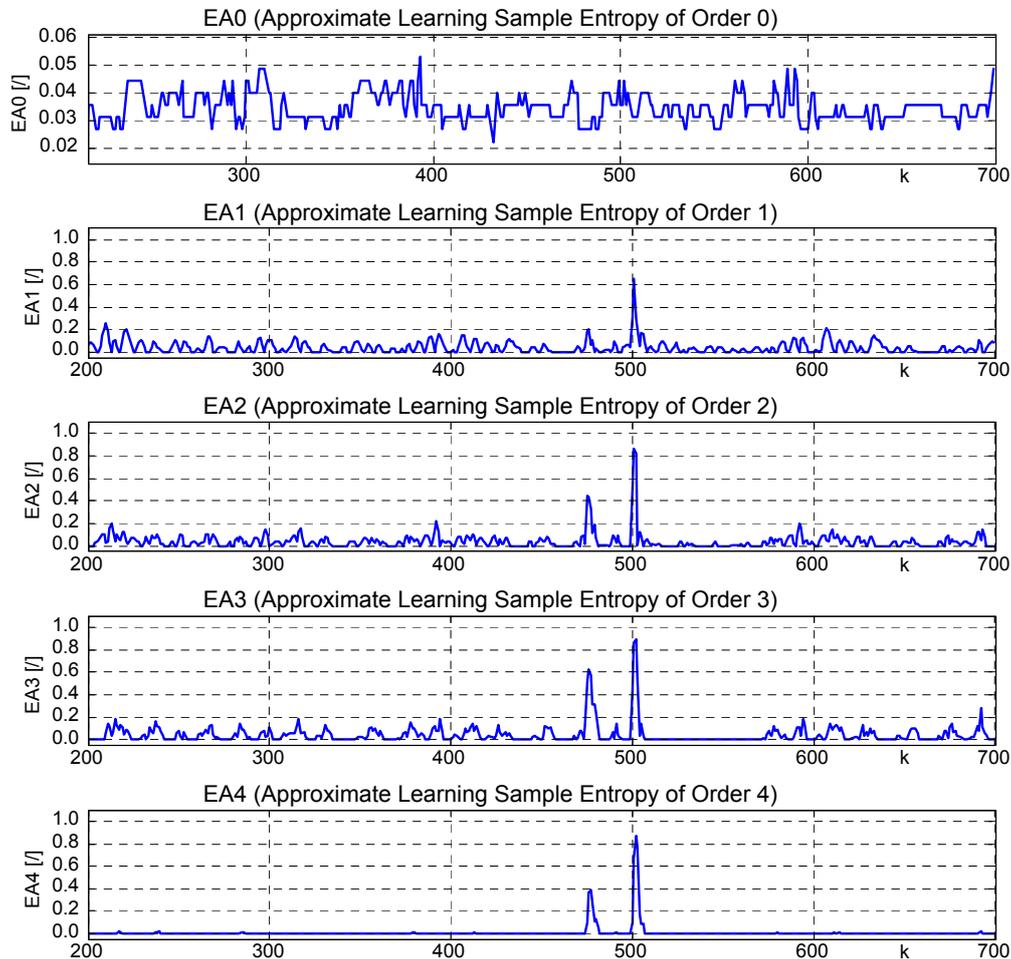
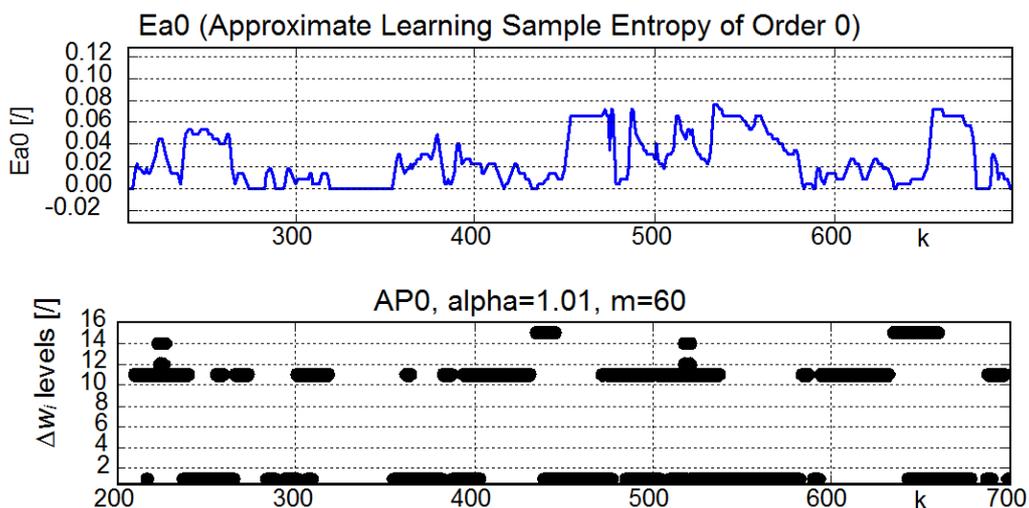


Figure 8. (Top) E_A^0 calculated for $\alpha = \{1.01^{[15, 14, \dots, 1]}\}$ that is closer to corresponding α_{max} of zero order ISLE and (bottom) the AP for $\alpha = 1.01$; Zero-Order learning entropy E^0 does not capture the inconsistent data (signal and other setup as in Figure 7).

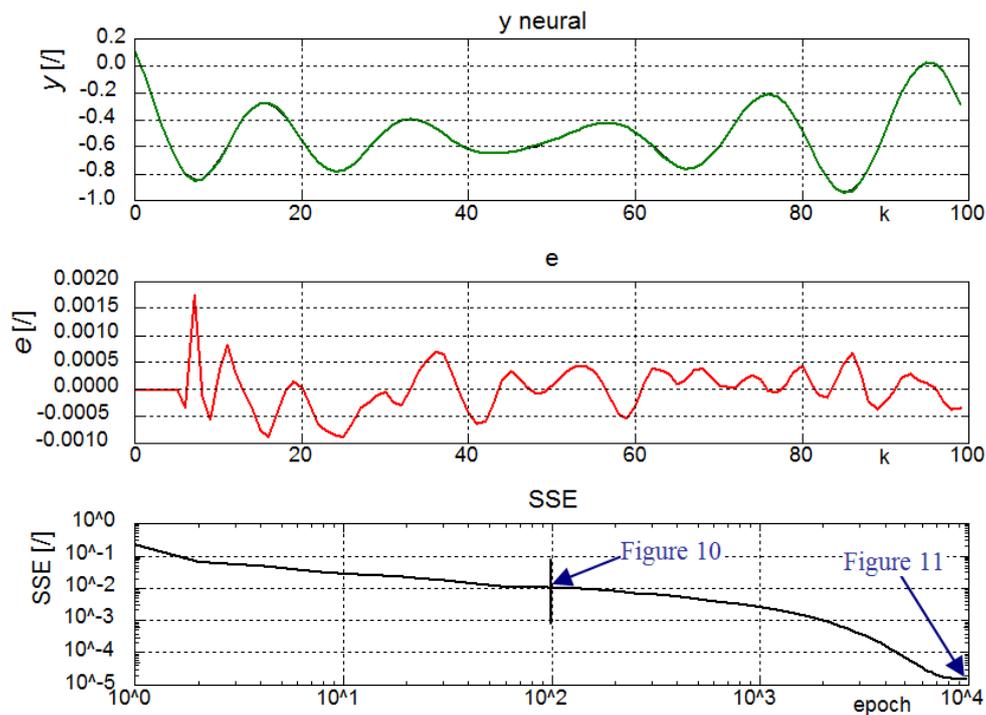


4. Experimental Analysis

4.1. A Hyper-Chaotic Time Series

Another theoretical example of time series where the AISLE can be clearly demonstrated is the time series obtained from hyper-chaotic coupled Chua’s circuit [45] (with some more details and results on AP in [2]). The dimension of the used coupled Chua’s circuit in continuous time domain is 6 and its embedding dimension shall be at least $2 \times 6 + 1 = 13$ according to the Taken’s theorem. Let’s choose the static QNU of lower dimension (embedding $n = 6$) as the learning model in sense of Equation (1) and in particularly according to Equations (12–14). Its proper pre-training by GD (16) for the first only 100 samples of the time series is shown in Figure 9; the sum of square errors (SSE) approaches $1E-5$ after last epoch of pre-training.

Figure 9. Pre-training of low-dimensional QNU (12–16), on first 100 samples of hyper-chaotic time series in 10,000 epochs; $\mu = 1$; $n = 6 \Rightarrow 28$ weights; quality of pre-training (SSE) affects LE (Figure 10 vs. Figure 11).



Let us now introduce a slight perturbation in two samples as follows:

$$y(k=475)=y(k=475)-.02 \text{ and } y(k=500)=y(k=500)-0.02 \tag{26}$$

Then, the evaluation of AISLE for the less properly and more properly pre-trained learning model (Figure 9) is given in Figure 10 and Figure 11, respectively.

Figure 10. AISLE for the less properly pre-trained model; only 200 epochs of pre-training (Figure 9) naturally result in that the LE appears larger for more samples than just the perturbed ones at $k = 457, 500$ Equation (26), see also Figure 11.

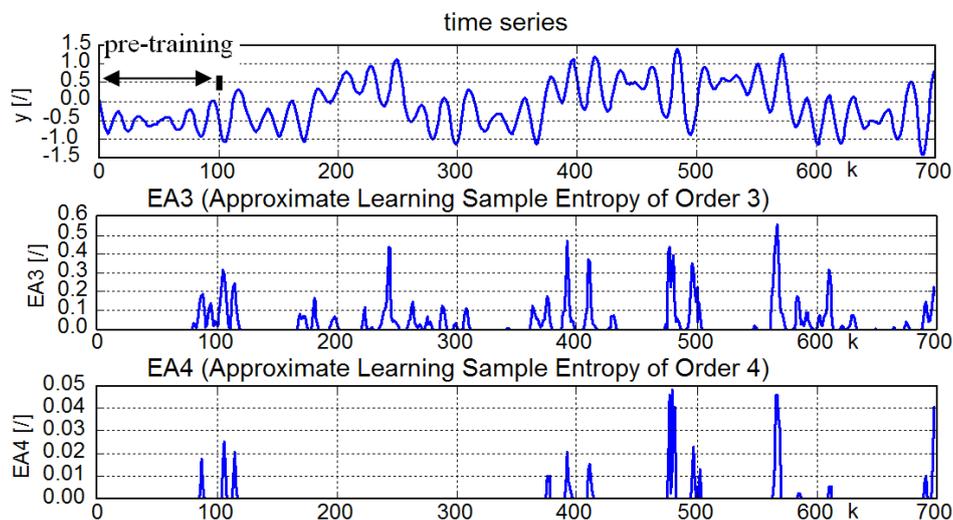
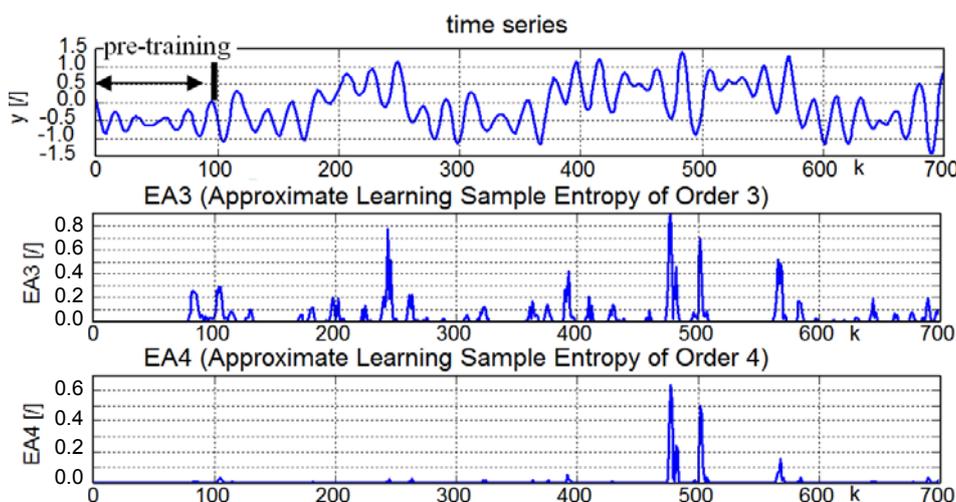


Figure 11. AISLE for more properly pre-trained learning model than in Figure 10. (here $1E + 4$ epochs, Figure 9); two slight perturbations at $k = 457, 500$ Equation (26) are followed by larger AISLE (esp. of 4th order).



However, the design of a proper learning model and its correct pre-training are crucial and non-trivial tasks for correct evaluation of LE and it may require an expert in adaptive (learning) systems or in neural networks. Nevertheless, from our experiments with AP and HONU [1,2,36–40] it appears that the very precise pre-training of the learning model is practically not always too crucial and that the structure of a learning model can be designed quite universally, e.g., with HONUs as they are nonlinear mapping predictors that are naturally linear in parameters. A practical rule of thumb for the above introduced HONU and GD can be to keep pre-training as long as the error criteria keep decreasing, *i.e.*, until the learning model learns what it can in respect to its quality of approximation vs. the complexity of the data. The effect of more proper pre-training can be demonstrated by comparison of Figure 10 with Figure 11, where we can see that more proper pre-training naturally filters out the

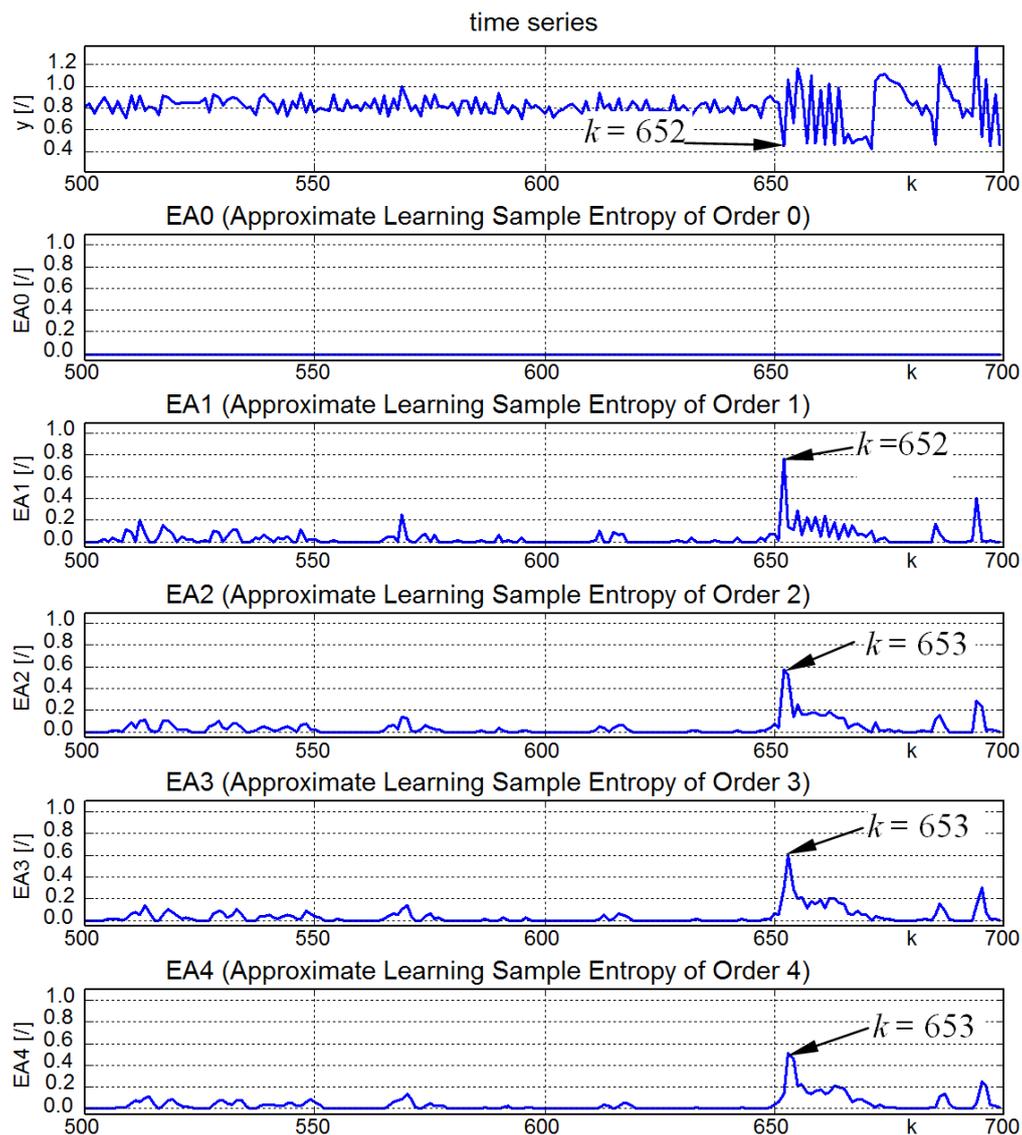
inconsistent samples at $k = 475, 500$, from all the other samples that are naturally generated by the hyper-chaotic behavior.

This subsection demonstrated the calculation of the LE on a theoretical hyper-chaotic time series and it demonstrated the influence of the quality of pre-training on its evaluation. Two real-world data examples are given in the next subsection.

4.2. Real Time Series

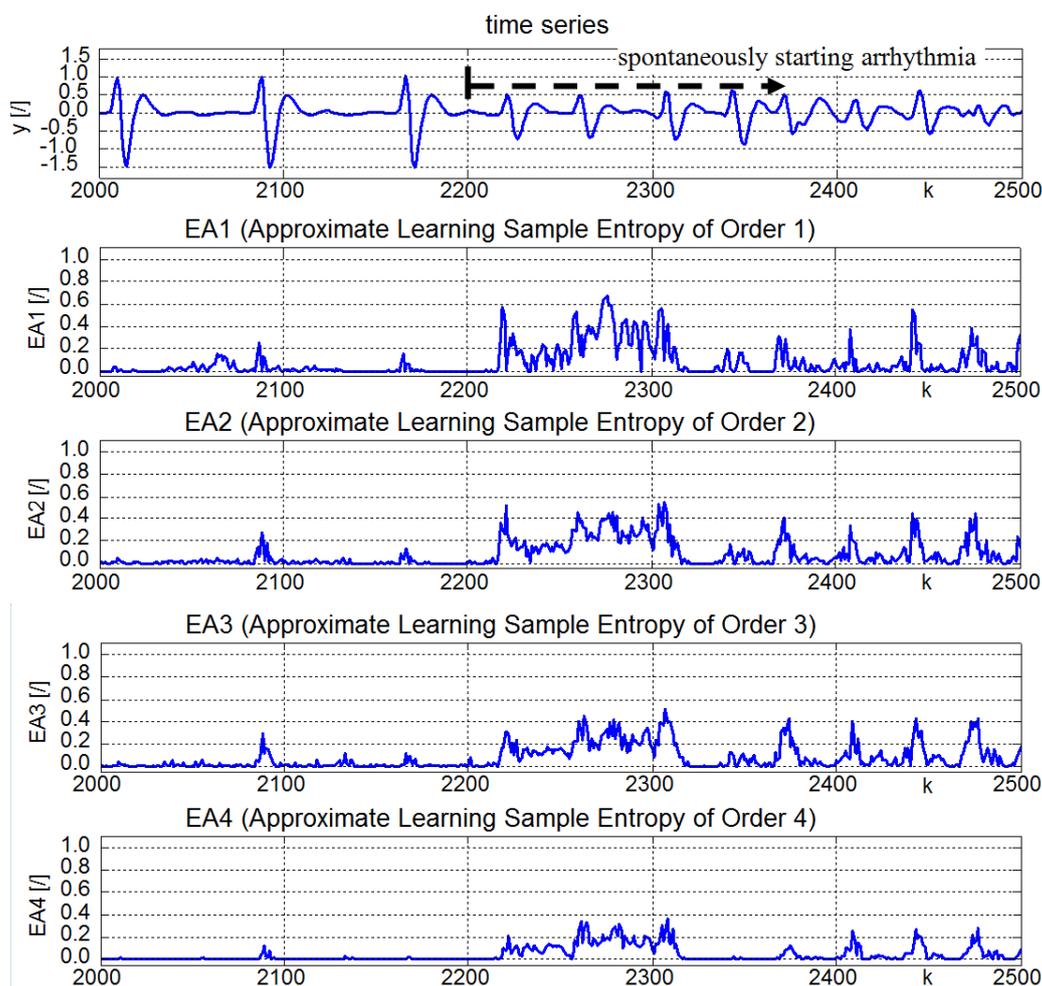
Heart beat tachograms (R-R diagrams) and ECG signals are complex and non-stationary time series that are generated by a multidimensional and multilevel feedback control system (the cardiovascular system) with frequent external and internal perturbations of various kind. First, this subsection demonstrates potentials for real-world use of LE on real-time novelty detection of heart beat samples in R-R diagram retrieved from [46] using static QNU and GD learning via Equations (12–16).

Figure 12. AISLE for R-R diagram [46]; the learning model is static QNU, $n = 5$, pre-training samples = 200, epochs = 100, $\mu = 0.001$; the peaks of AISLE correspond or directly follow the inconsistent sample at $k = 652$.



The results and the pre-training setup is given in Figure 12, which shows that the new pattern of heart rate behavior starting at $k = 652$ has been detected. Second, demonstration of potentials of LE is shown on real-time sample-by-sample monitoring of ECG with algorithm modification for quasi-periodic signals. LE is shown for a real-time sample-by-sample monitoring of ECG time series with spontaneous onset of ventricular tachycardia (233 Hz, data courtesy of [47]) using Linear Neural Unit (LNU) with normalized GD, *i.e.*, adaptive linear filter (5–7), (13). The dimensionality of the used LNU is $n = 80$ Equation (13), so the calculation of AISLE for also a higher dimensional predictor is demonstrated below in Figures 13, 14 and A1.

Figure 13. Capturing the onset of spontaneous ventricular tachycardia in animal ECG by the AISLE of static Linear Neural Unit (LNU, $n = 80$, pre-training epochs = 500, pre-training samples $k = 0.500$, full data in Figure 16), data courtesy of [47].

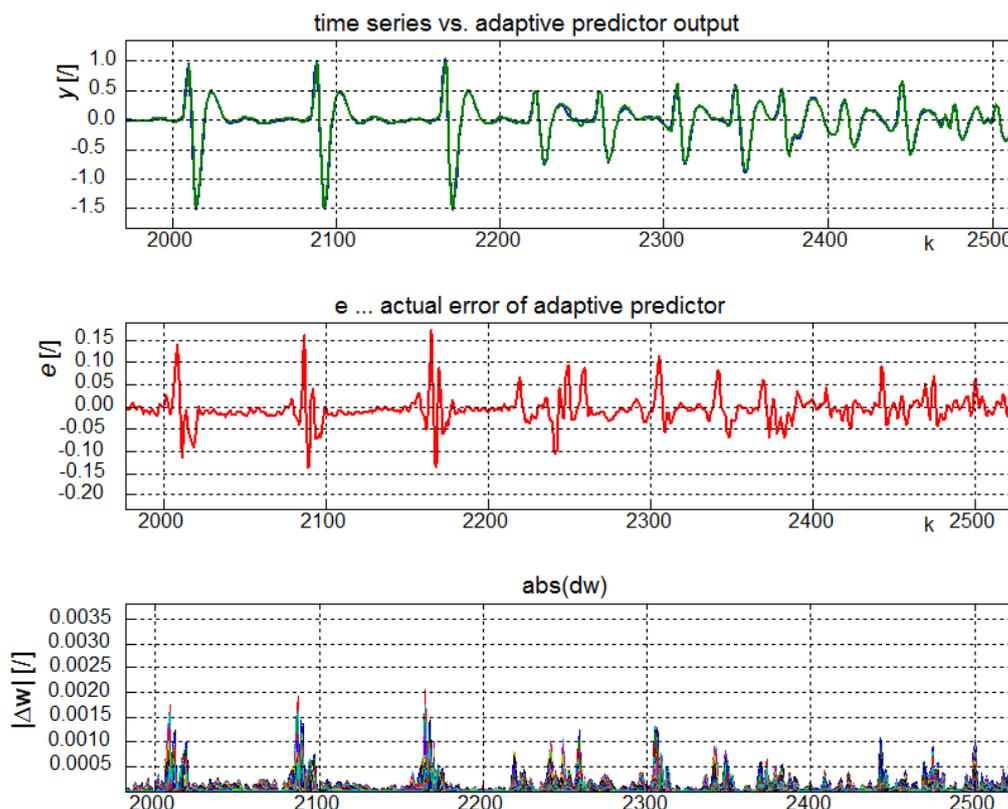


Because ECG is a quasi-periodical signal and the used LNU is not able to fully learn the governing law that drives the ECG, the evaluation of LE can be modified to compare neural weight increments with respect to the prevailing periodicity of the signal. The floating average of absolute values of recent m neural increments of i^{th} neural weight (originally given in (9)), can be modified as follows:

$$\overline{|\Delta w_i(k)|} = \frac{1}{m} \cdot \left(\sum_{j=k-90}^{k-70} |\Delta w_i(j)| \right), \tag{27}$$

where, the actual time range of averaging Δw is located into the neighborhood of maximum autocorrelation of ECG signal (here for the lag of 80 samples).

Figure 14. The typical feature of LE: while prediction error and weight increments reach seemingly regular magnitudes or even smaller ones, the LE can be correctly high regardless the actual error of the learning model (see Figures 13 and 15).



5. Discussion

The straightforward difference of the introduced LE from statistical entropy approaches is that behavior of a system may be statistically complex, but if the behavior is deterministic (e.g., deterministic chaos) and newly measured samples are consistent with the governing law, then these data does not carry any novel information. Due to the incremental learning (adaptation) during detection, the LE approach is potentially suitable for real time non-stationary systems.

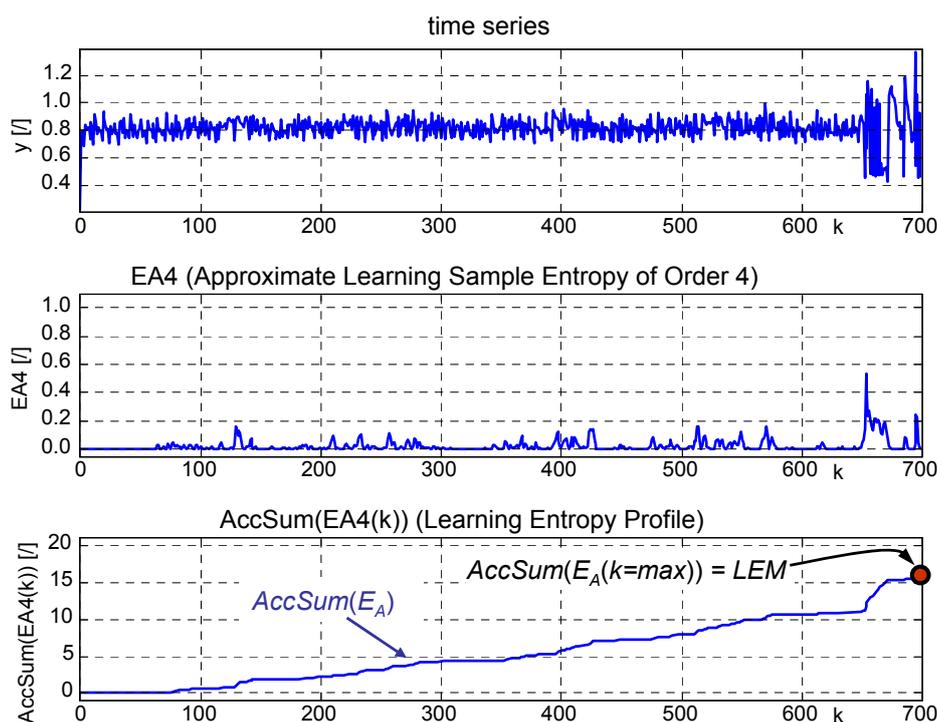
The weights in w and especially their higher-order time derivatives (estimated via weight increments Δw) correspond to the (learning) energy of an incrementally learning system, and it appears that higher orders of AISLE shall be generally considered. Naturally, higher orders of AISLE appear more reliable for novelty detection than the Zero-Order AISLE; however, higher orders AISLE appeared one sample delayed behind the first order AISLE (Figure 12).

The LE is a relative measure, that is related to the data to which it is applied as well as, it is related to the learning model and to its capability to extract the governing law from training data. Obtaining a useful LE can be a non-trivial task that requires a suitable (though not perfect) learning model (e.g., a low-dimensional model) that is capable to approximate a temporary governing law in data.

It is generally difficult to provide readers with the pragmatic rule of thumb that can be suggested for selection and pre-training of a model for LE, because the possible consequences of the insufficient training or the overtraining vary from case to case and may depend on used type of learning model and applied learning algorithm and on setup parameters, so the user’s experience might be necessary. Nevertheless, linear filters or relatively simple polynomial predictors that are linear in parameters can be a good option to start with experiments on the LE.

Regarding supervised learning (*i.e.*, predictors), it was demonstrated already in [40] on chaotic logistic equation that the actual prediction error is not necessary correlated to the inconsistency of data that is adaptively monitored. Details of adaptive predictor output, prediction error, and the magnitude of weight increments during incremental learning of ECG signal for Figure 13 are given in detail in Figure 14. By comparison of the two figures, we see that the onset of ventricular tachycardia does not introduce larger prediction error (Figure 14); however, the LE of the starting arrhythmia is high (Figures 13 and 15) for the learning model used.

Figure 15. Profile of LE of 4th order and the *Learning Entropy of a Model* (LEM) evaluated at $k = 700$ for QNU and for the time series from Figure 12.



While this papers resolves the single-scale issue of sensitivity detection parameter α via the multiscale approach, another point of discussion is selection of parameter m in formula (9) for a particular signal. A pragmatic rule for setting m according to the lowest frequency component of chaotic signals with quasi-periodic nature was indicated in subsection 3.1. Then an interval modification for choosing m for a quasi-periodical signal with significantly distinct intervals of behavior (e.g., the ECG signal) was demonstrated in Equation (27), where formula (9) was customized to calculate average increment magnitudes within the lag corresponding to the first maximum of autocorrelation function. For chaotic signals where the periodicity and maximum lag of autocorrelation

function is not clear (e.g., discrete time chaos of the logistic equation or R-R diagrams), the choice of parameter m can be a case dependant and future research is to be carried. Also, a possible multi-scale approach for m for LE can be considered for future research. Nevertheless, the author has observed that calculation of LE is practically much more robust to the selection of parameter m than to the sensitivity detection α . Therefore, the multi-scale approach for α and not for m is primarily introduced for LE in this paper.

The LE is also a promising concept for development and research of new measures that would evaluate particular learning models and data. While this paper introduces individual sample focused LE, *i.e.*, ISLE and AISLE, there is strong potential for interval-based measures of LE. For example, by introducing the accumulated plot of LE, *i.e.*, the cumulative sum of AISLE, as follows:

$$AccSum(E_A(k)) = \sum_{i=1}^k (E_A(i)), \quad (28)$$

one may obtain the LE profile of a particular adaptive model and of particular data. When evaluated for the whole time series, $AccSum(E_A)$ would summarize the signal and might be possibly used to distinguish between different models or adaptation techniques. Then the very last point of the profile, *i.e.*, $AccSum(E_A(k = \max))$, could be interpreted as the *Learning Entropy of a Model (LEM)*, see Figure 15.

In other words, the LEM quantifies the familiarity of an adaptive model with data. Practically, the familiarity of a learning model with data corresponds to the generalization ability of the model. Adopting the fact that the best learning model shall have the best generalization ability ($LEM = \min$) and the lowest prediction error, a general function of the convolution of Learning Entropy and the prediction error appears be a promising direction for continuing research that, however, exceeds the scope of this paper.

There are certainly many other issues that should be discussed regarding evaluation of LE and the further introduced concepts. The proper evaluation of LE depends on a number of factors where users experience with learning systems and signal processing is important. However, the objective of this paper is to introduce the LE as a new non-probabilistic concept of online novelty detection via evaluation of data sample inconsistency with contemporary governing law that is incrementally learned by a learning system.

6. Conclusions

This paper is the first work that introduces the concept of LE as a non-probabilistic online measure for relative quantification of novelty of individual data samples in time series. This normalized and multi-scale measure evaluates the inconsistency of individual samples of data as an unusual learning activity of an incrementally learning model over all adaptable parameters. It is a multi-scale measure because the learning activity is evaluated over the whole range of detection sensitivity parameter α that is the key parameter for online visualization of unusual learning activity (in AP). Evaluation of unusual learning activity was proposed for estimation of various orders of time derivatives of weights that reflects the learning energy of an incrementally learning model, thus Orders of LE were introduced.

A particular technique for calculation of Approximated Individual Sample Learning Entropy was introduced. AISLE represents a loose analogy to approximate sample entropy in sense of using cumulative

sums instead of approximation of the slope in a log-log coordinates. The whole explanation and the technique of calculation of AISLE is demonstrated on a straightforward example of supervised incremental learning, *i.e.*, on GD in this paper. As learning models, static linear and polynomial neural units (quadratic polynomials) were demonstrated in this paper as they are good to start with LE for their comprehensibility and in-parameter-linearity which is a good feature for GD learning. Examples of calculating the AISLE for theoretical chaotic time series as well as for two bio-signals were presented.

The major objective of this paper was a comprehensible introduction of the LE and its calculation rather than competition to conventional entropy approaches. The LE is introduced as a missing concept among probabilistic entropy approaches that usually do not consider a governing law in otherwise statistically complex data. In principle, the concept of LE is not only limited to supervised learning. There are strong potentials for LE for neural networks, signal processing, adaptive control, fault and concept drift detection, and big data applications. The evaluation of Learning Entropy will depend on many factors including users experience with adaptive systems and the detailed summary exceeds the introductory focus of a single paper.

Acknowledgments

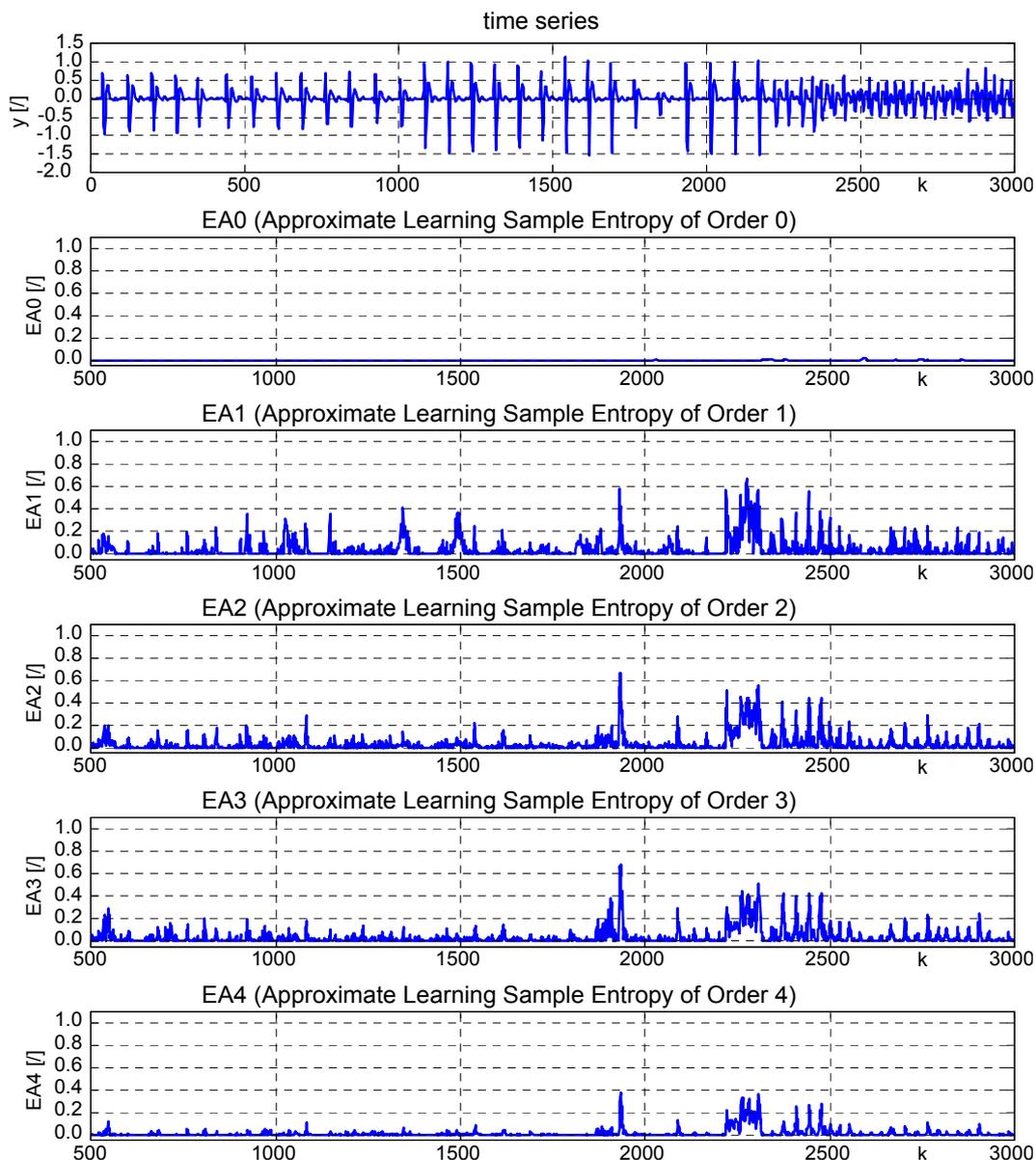
The author would like to thank Madan M. Gupta from the University of Saskatchewan for introducing him to higher-order neural units and for continuous support and consultancy. The author would also like to thank Witold Kinser from the University of Manitoba for introducing him to multi-scale analysis approaches and for continuous support and consultancy, the colleagues from Tohoku University, namely from the Yoshizawa-Sugita Lab (formerly Yoshizawa-Homma Lab) and from the Department of Radiological Imaging and Informatics, for their continuous and vital cooperation and support. Special thanks are due to The Matsumae International Foundation (Tokyo, Japan) that funded the author's cooperation with colleagues at Tohoku University in Japan in 2009, and this cooperation is still vitally continuing. The author would like to thank anonymous reviewers for their insightful and constructive remarks. The author has been partly supported by grant SGS12/177/OHK2/3T/12 and partly by project TA03020312. The author would like to acknowledge the language Python [48] and the related open-source communities for developing these open-source software and its libraries.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix

Figure A1. Full data for Figure 13 (excluding first 500 of pre-training samples), arrhythmia spontaneously starts around $k = 2200$), data courtesy of [47].



References

1. Bukovsky, I. Modeling of complex dynamic systems by nonconventional artificial neural architectures and adaptive approach to evaluation of chaotic time series. Ph.D. Thesis, Czech Technical University, Prague, Czech Republic, 2007.
2. Bukovsky, I.; Bila, J. Adaptive evaluation of complex dynamic systems using low-dimensional neural architectures. In *Advances in Cognitive Informatics and Cognitive Computing, Series: Studies in Computational Intelligence*; Zhang, D., Wang, Y., Kinsner, W., Eds.; Springer-Verlag: Berlin & Heidelberg, Germany, 2010; Volume 323, pp. 33–57.

3. Baker, W.; Farrell, J. An introduction to connectionist learning control systems. In *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*; White, D.A., Sofge, D.A., Eds; Van Nostrand and Reinhold: New York, USA, 1993; pp. 36–63.
4. Polycarpou, M.M.; Trunov, A.B. Learning approach to nonlinear fault diagnosis: Detectability analysis. *IEEE Trans. Autom. Control* **2000**, *45*, 806–812.
5. Markou, M.; Singh, S. Novelty detection: A review—Part 1: Statistical approaches. *Signal Process.* **2003**, *83*, 2481–2497.
6. Markou, M.; Singh, S. Novelty detection: A review—Part 2: Neural network based approaches. *Signal Process.* **2003**, *83*, 2499–2521.
7. Pincus, S.M. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301.
8. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circul. Physiol.* **2000**, *278*, 2039–2049.
9. Kinsner, W. Towards cognitive machines: Multi-scale measures and analysis, *Intern. J. Cognit. Inf. Natural Intell.* **2007**, *1*, 28–38.
10. Kinsner, W. A unified approach to fractal dimensions. *Intern. J. Cognit. Inf. Natural Intell.* **2007**, *1*, 26–46.
11. Kinsner, W. Is entropy suitable to characterize data and signals for cognitive informatics? *Intern. J. Cognit. Inf. Natural Intell.* **2007**, *1*, 34–57.
12. Zurek, S.; Guzik, P.; Pawlak, S.; Kosmider, M.; Piskorski, J. On the relation between correlation dimension, approximate entropy and sample entropy parameters, and a fast algorithm for their calculation. *Physica A* **2012**, *391*, 6601–6610.
13. Schroeder, M., R. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*; Freeman: New York, NY, USA, 1991.
14. Costa, M.; Goldberger, A.L.; Peng, C.K. Multi-scale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* **2002**, *89*, 68102.
15. Costa, M.; Goldberger, A.L.; Peng, C.K. Multi-scale entropy analysis of biological signals. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* **2005**, *71*, 021906.
16. Wu, S.-D.; Wu, Ch.-W.; Lin, S.-G.; Wang, Ch.-Ch.; Lee, K.-Y. Time series analysis using composite multi-scale entropy. *Entropy* **2013**, *15*, 1069–1084.
17. Gonçalves, H.; Henriques-Coelho, T.; Rocha, A.P.; Lourenço, A.P.; Leite-Moreira, A.; Bernardes, J. Comparison of different methods of heart rate entropy analysis during acute anoxia superimposed on a chronic rat model of pulmonary hypertension. *Med. Eng. Phys.* **2013**, *35*, 559–568.
18. Wu, S.-D.; Wu, Ch.-W.; Wu, T.-Y.; Wang, Ch.-Ch. Multi-scale analysis based ball bearing defect diagnostics using mahalanobis distance and support vector machine. *Entropy* **2013**, *15*, 416–433.
19. Faes, L.; Nollo, G.; Porta, A. Compensated transfer entropy as a tool for reliably estimating information transfer in physiological time series. *Entropy* **2013**, *15*, 198–219.
20. Yin, L.; Zhou, L. Function based fault detection for uncertain multivariate nonlinear non-gaussian stochastic systems using entropy optimization principle. *Entropy* **2013**, *15*, 32–52.
21. Vorburger, P.; Bernstein, A. *Entropy-based concept shift detection*. In Proceedings of the 2006 IEEE International Conference on Data Mining (ACDM 2006), Hong Kong, China, 18–22 December 2006; pp.1113–1118.

22. Willsky, A. A survey of design methods for failure detection in dynamic systems. *Automatica* **1976**, *12*, 601–611.
23. Gertler, J. Survey of model-based failure detection and isolation in complex plants. *IEEE Contr. Syst. Mag.* **1988**, *8*, 3–11.
24. Isermann, R. Process fault detection based on modeling and estimation methods: A survey. *Automatica* **1984**, *20*, 387–404.
25. Frank, P.M. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy—A survey and some new results. *Automatica* **1990**, *26*, 459–474.
26. Widmer, G.; Kubat, M. Learning in the presence of concept drift and hidden contexts. *Machine Learn.* **1996**, *23*, 69–101.
27. Polycarpou, M.M.; Helmicki, A.J. Automated fault detection and accommodation: A learning systems approach. *IEEE Trans. Syst. Man Cybern.* **1995**, *25*, 1447–1458.
28. Demetriou, M.A.; Polycarpou, M.M. Incipient fault diagnosis of dynamical systems using online approximators. *IEEE Trans. Autom. Control* **1998**, *43*, 1612–1617.
29. Trunov, A.B.; Polycarpou, M.M. Automated fault diagnosis in nonlinear multivariable systems using a learning methodology. *IEEE Trans. Neural Networks* **2000**, *11*, 91–101.
30. Alippi, C.; Roveri, M. Just-in-time adaptive Classifiers—Part I: Detecting nonstationary changes. *IEEE Trans. Neural Networks* **2008**, *19*, 1145–1153.
31. Alippi, C.; Roveri, M. Just-in-time adaptive Classifiers—Part II: Designing the Classifier. *IEEE Trans. Neural Networks* **2008**, *19*, 2053–2064.
32. Alippi, C.; Boracchi, G.; Roveri, M. Just-In-time Classifiers for recurrent concepts. *IEEE Trans. Neural Networks Learn. Syst.* **2013**, *24*, 620–634.
33. Alippi, C.; Ntalampiras, S.; Roveri, M. A Cognitive fault diagnosis system for distributed sensor networks. *IEEE Trans. Neural Networks Learn. Syst.* **2013**, *24*, 1213–1226.
34. Grossberg, S.; Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks* **2013**, *37*, 1–47.
35. Gupta, M.M.; Liang, J.; Homma, N. *Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory*; John Wiley & Sons: New Jersey, USA, 2003.
36. Bukovsky, I.; Bila, J.; Gupta, M.M.; Hou Z.-G.; Homma, N. Foundation and Classification of Nonconventional Neural Units and Paradigm of Nonsynaptic Neural Interaction. In *Discoveries and Breakthroughs in Cognitive Informatics and Natural Intelligence*; Wang, Y., Ed.; IGI Publishing: Hershey, PA, USA, 2009.
37. Bukovsky, I.; Homma, N.; Smetana, L.; Rodriguez, R.; Mironovova M.; Vrana S. *Quadratic neural unit is a good compromise between linear models and neural networks for industrial applications*. In Proceedings of the 9th IEEE International Conference on Cognitive Informatics (ICCI 2010), Beijing, China, 7–9 July 2010.
38. Bukovsky, I.; Anderle, F.; Smetana, L. *Quadratic neural unit for adaptive prediction of transitions among local attractors of Lorenz systems*. In Proceedings of the 2008 IEEE International Conference on Automation and Logistics, Qingdao, China, 1–3 September 2008.
39. Bukovsky, I.; Bila, J. *Adaptive evaluation of complex time series using nonconventional neural units*. In Proceedings of the 7th IEEE International Conference on Cognitive Informatics (ICCI 2008), Stanford, CA, USA, 14–16 August 2008.

40. Bukovsky, I.; Kinsner, W.; Bila, J. *Multi-scale analysis approach for novelty detection in adaptation plot*. In Proceedings of the 3rd Sensor Signal Processing for Defence (SSPD 2012), London, UK, 25–27 September 2012.
41. Mandic, D.P. A generalised normalised gradient descent algorithm. *IEEE Signal Process. Lett.* **2004**, *11*, 115–118.
42. Choi, Y.-S.; Shin, H.-Ch.; Song, W.-J. Robust regularization for normalized lms algorithms. *IEEE Trans. Circuits Syst. Express Briefs* **2006**, *53*, 627–631.
43. Williams, R.J.; Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1989**, *1*, 270–280.
44. Mackey, M.C.; Glass, L. Oscillation and chaos in physiological control systems. *Science* **1977**, *197*, 287–289.
45. Cannas, B.; Cincotti, S. Hyperchaotic behaviour of two bi-directionally coupled chua's circuits. *Inter. J. Circuit Theory Appl.* **2002**, *30*, 625–637.
46. R-R diagram, record #: 222. PhysioBank: MIT-BIH Arrhythmia Database. Available online: <http://www.physionet.org/physiobank/database/mitdb> (accessed on 5 April 2001).
47. Yoshizawa-Homma Lab. <http://www.yoshizawa.ecei.tohoku.ac.jp/~en> (accessed on 5 July 2013).
48. Van Rossum, G.; de Boer, J. Linking a stub generator (AIL) to a prototyping language (Python). In Proceedings of the Spring 1991 EurOpen Conference, Troms, Norway, 20–24 May 1991.

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).