Dynamic clustering of time series with Echo State Networks*

Miguel Atencia[1], Catalin Stoean[2], Ruxandra Stoean[2], Roberto Rodríguez-Labrada[3], and Gonzalo Joya[1]

[1] Universidad de Málaga, Spain
{matencia, gjoya}@uma.es
[2] University of Craiova, Romania
{catalin.stoean, ruxandra.stoean}@inf.ucv.ro
[3] Centro para la Investigación y Rehabilitación de las Ataxias Hereditarias, Cuba
roberto@ataxia.hlg.sld.cu

Abstract. In this paper we introduce a novel methodology for unsupervised analysis of time series, based upon the iterative implementation of a clustering algorithm embedded into the evolution of a recurrent Echo State Network. The main features of the temporal data are captured by the dynamical evolution of the network states, which are then subject to a clustering procedure. We apply the proposed algorithm to time series coming from records of eye movements, called saccades, which are recorded for diagnosis of a neurodegenerative form of ataxia. This is a hard classification problem, since saccades from patients at an early stage of the disease are practically indistinguishable from those coming from healthy subjects. The unsupervised clustering algorithm implanted within the recurrent network produces more compact clusters, compared to conventional clustering of static data, and provides a source of information that could aid diagnosis and assessment of the disease.

Keywords: Echo State Networks, Clustering, k-means, Saccadic eye movement, Time series

1 Introduction

In this work we propose a novel methodology for unsupervised clustering of time series, by inputting the sequences to an Echo State Network (ESN), which is a recurrent neural network. Cluster analysis refers to the *discovery* of classes or categories within data that were previously unknown, thus it is included in the broad category of unsupervised learning tasks. Cluster analysis has a long history in the fields of statistics and machine learning (see e.g. the recent review [11] and references therein), and probably the best known clustering algorithm is k-means [6]. However, handling data with temporal features introduces a number

* Partially supported by the Spanish Ministry of Science, Innovation and Universities, through the Plan Estatal de Investigación Científica y Técnica y de Innovación, Project TIN2017-88728-C2-1-R, as well as the Universidad de Málaga.

of complications. First of all, the length of the sequence may be not known in advance, may be different for each sequence, or may even be infinite, which in all cases breaks the usual assumption that data is arranged as a rectangular (finite) matrix. Besides, time series can contain long term correlations, where very large chunks of data must be taken into account in order to fully capture the qualitative dynamical behaviour of the series. Finally, even if fixed-length data are available, considering time series as a vector neglects the temporal information, e.g. the correlation of data values at distant time points may be more significant that the find of similar values in successive components. Specific cluster analysis of temporal data has been tackled from the complementary viewpoints of time series [1] and data streams [3]. Many clustering algorithms for time series are applied on a *window* of data, which must then be carefully chosen: if it is too small, long-term relationships will be missed, whereas a large data window will introduce a signifcant computational cost.

This work is prompted by the need to analyse data coming from electroooculographic records, in order to implement a tool to aid in the diagnosis and assessment of spino-cerebellar ataxia of type 2 [12]. Data coming from ataxia patients in Cuba, as well as control, healthy subjects, have been analysed and labelled by physicians. Previously, data mining techniques have been applied to these data for the extraction of significant clinical events [2], but much work is needed in order to improve the accuracy of classification. In particular, the records from healthy individuals and patients at an early stage of the disease are almost indistinguishable, even for human experts.

Echo state networks are recurrent neural networks that can be classified into the paradigm of reservoir computing (see [8] and references therein). The most striking feature of ESNs is the absence of learning, in the conventional sense of modification of weights or connections between units. Instead, a set of units is fully connected by recurrent connections, whereas the connection values are constant. The feedback loops induce a dynamical behaviour that is intended to capture the important features of the time series that is presented as input. Echo State Networks have been applied to time series coming from medical applications [4], but there is much margin for improving our knowledge of both the fundamental analysis of these models, and their practical applicability. In particular, to the best of our knowledge, our proposal that a clustering algorithm is embedded inside the temporal evolution of an ESN is original.

In Section 2 the formulation of Echo State Networks is briefly reviewed, emphasizing the need to choose the network hyperparameters in order to produce a rich dynamical behaviour. The characteristics of the dataset of electrooculographic records are gathered in Section 3, stressing the difficulty of detecting the disease when eye movements are not yet significantly impaired. The proposed algorithm is described in Section 4, contrasting its principled definition to usual methods for clustering of time series. Experimental results are discussed in Section 5, whereas Section 6 puts an end to the paper with some final remarks and directions for further research.

2 Echo state networks

In their original formulation, Echo State Networks are supervised classification models that are applied to data stemming from time series. Given an input signal $x(t)$, the objective is to predict at every instant t a target output $o(t)$, where $t = 0 \dots T \dots$ is the discrete time whose final time t_f is not necessarily finite or known in advance. It is often the case that the final aim is time series prediction, and then the target output is simply the one-step ahead input $o(t) = x(t+1)$. The architecture of the network consists of a fixed number d of units that store a value at every time step, so the *state* of the network is a d-dimensional vector $y(t) \in \mathbb{R}^d$, which is time varying. The output is computed from the reservoir state and the input by a feedforward connectionist layer, which is usually linear:

$$\hat{o}(t) = W_{\text{io}}\, x(t) + W_{\text{out}}\, y(t) + b \tag{1}$$

where b is a bias vector. Therefore, learning proceeds by solving a regression problem through minimization of the error $E(t) = o(t) - \hat{o}(t)$ thus computing the (time-varying) parameters W_{io}, W_{out}, b. Note that, the dependence of the error on parameters being linear, this regression problem is particularly simple.

In an ESN, the reservoir states $y(t)$ are dynamically updated by the following recursive rule:

$$y(t) = (1-l)\, y(t-1) + l\, \tanh\left(W_{\text{in}}\, x(t) + W\, y(t-1)\right) \tag{2}$$

where $l \in [0, 1]$ is a *leaking* hyperparameter. The distinguishing feature of ESNs, compared to e.g. backpropagation through time learning [5], is that the weight matrices W_{in}, W are constant, and fixed once and for all at the beginning of the reservoir time evolution. The rationale behind the construction of ESNs is that the dynamical evolution of the reservoir is able to grasp the important features of the input, which is thus *echoed* by the states, so the influence of the particular chosen values of the weight matrices is negligible. When this objective is achieved, the network is said to possess the *echo state* property.

In order to construct an ESN, choices must be made regarding the value of several hyperparameters. First of all, the reservoir size d must attain a trade-off between capacity and computational cost, and it is important that the number of units is related to the length and dimensionality of input data, since a reservoir too large fed by insufficient data may lead to overfitting. The already mentioned leaking rate l adjusts the balance between long and short term memory, and it can be regarded as a measure of the velocity of the dynamical evolution of the network. Finally, the strategy to define the weight matrix W is the choice that most influences the qualitative dynamical behaviour of the network. There are several—to some extent, equivalent—criteria to characterize the network dynamics with a single hyperparameter that rules the construction of the weight matrix. One of these concepts is the maximum Lyapunov exponent [13], which, if positive, signals that the dynamics is chaotic. Another useful measure of the complexity of the network dynamics is the spectral radius R, which is the maximum absolute value of the eigenvalues of the weight matrix W. A stability analysis by

linearization shows that if $R < 1$ then the network state tends to an equilibrium, in the absence of input. This behaviour is considered undesirable, since a single fixed point would not carry the whole information of the input. On the contrary, a very large value of R would cause the network states to grow without bound. The rule of thumb for fixing the spectral radius suggests that its value should be slightly larger than 1, to produce a rich dynamics while the states remain bounded [8]. Anyway, these results can hardly be rigorously applied in practice since, on the one hand, the influence of inputs cannot be neglected and, on the other hand, the presence of the hyperbolic tangent introduces a non-linearity.

Since the objective of this paper is unsupervised learning, we omit the output layer, together with the regression computation to fix the parameters W_{io}, W_{out}, b. Our focus is on the network dynamics being able to capture the most significant features of the time series, which will allow for clustering.

3 Dataset of saccadic movements

The motivating problem for the current work is to aid to the diagnosis of spinocerebellar ataxia of type 2 (SCA2), which is a degenerative disorder that causes uncoordinated movement, among other neurological symptoms. In particular, weakening eye muscles produce a slowing of fast eye movements (saccades). The onset of the disease is progressive and its duration usually ranges between 10 and 15 years [9]. At initial stages, patients can undergo relatively mild symptoms, or even none at all, thus diagnosis may be delayed until clinical manifestations are severely disabling. Contrarily, early diagnosis can help establish a supportive program that includes physical therapy and life style adaptations. Since clinical manifestations are often inconclusive, the definitive diagnosis can only be established by genetic testing. However, specific genetic analysis is costly and cannot be used as a general screening method in areas where the disease is prevalent, such as Cuba. Instead, the examination of the degree of alteration of saccadic movements provides a cheap and accessible diagnostic tool.

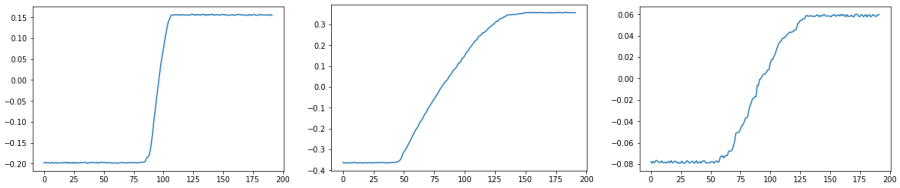

Fig. 1. Sample saccades (at different vertical scales) corresponding to a healthy individual, a presymptomatic subject, and an ill patient, from left to right. Observe the different amplitude, speed, and steadiness.

Saccades are measured by means of an electrooculograpy device, which samples the weak electrical potentials due to eye movements, when the subject is

trying to track the trajectory of an object on a screen. Typically, the test is repeated with several amplitudes that lead to different angular displacements of the subject's eyes. In this work we have used a database that contains 88 registers, which have been labelled as C (control, healthy individuals), P (presymptomatic subjects, including those with very mild degeneration), and D (diseased patients with severe clinical manifestations). The electrooculograpic records are analysed to isolate individual saccades, which are then preprocessed leading to a database of 6124 saccades, each containing 192 samples. The details of this preprocessing are included in a companion paper. As an example, three saccades each corresponding to one class are shown in Figure 1.

It is important to emphasize that labelling has been performed by medical experts from their subjective analysis, knowledge of the subject's family history, genetic tests, or other means, because observation of saccades alone does not allow to establish a diagnosis. In particular, individual saccades from presymptomatic patients are often indistinguishable from those recorded in healthy subjects. Thus from the point of view of classification as a supervised learning task, the problem at hand can be considered as *ill-posed*, since some instances from presymptomatic patients possess almost exactly the same data features as those from healthy subjects. Therefore a supervised classification method is prone to low accuracy, since data are mislabelled. This suggests the use of unsupervised techniques to provide a finer clustering than the one obtained from the standard classification procedure.

4 Dynamic clustering

The proposed procedure can be summarised as the application of a clustering method inside every step of the ESN. Starting from a dataset with n saccades, n identically initialized ESNs are evolved, each accepting as input the series of the corresponding saccade. Then, at every evolution step t, the reservoir states of all ESNs are stacked to form a new database, which is subject to a clustering method, thus obtaining k clusters, each represented by its centroid. At the next time step, the clustering process is repeated, but now the initial clusters are set to those coming from the previous step $t-1$. The final produced clustering is the one that results from the last time step in the series. A formal algorithmic description of the proposed method is shown in Algorithm 1.

The rationale behind the described algorithm is that reservoir states should capture time series dynamics, without the need to explicitly compute frequency-domain features, such as Fourier transforms. In contrast, methods based upon sliding windows on the series suffer from a limited bandwidth and cannot achieve long-term memory. Therefore our novel proposal would constitute a sort of hybrid between time-domain and frequency-domain techniques. In the description above, we have not specified which clustering method is used. In principle, any unsupervised clustering technique would work, but the minimal adaptations would be needed for iterative, partitioning methods, where the cluster centroids obtained at the previous step can be used for the next initialization.

Algorithm 1 Dynamic clustering through evolution of the Echo State Network.

Require: Data set of n saccades X, number of clusters k, final time t_f
Ensure: k centroids

```
Initialize ESN weight matrices W_i, W and replicate n identical instances
for t = 0 to t_f do
    for each saccade do
        Update the corresponding ESN instance by Equation (2)
    end for
    if t = 0 then
        Initialize centroids
    else
        Set initial centroids to centroids resulting from step t − 1
    end if
    Build the dataset Y of n reservoir states
    Compute centroids at step t from clustering of dataset Y
end for
```

We have implemented the experiments with the well-known method k-means, which is computationally rather efficient. In contrast, hierarchical methods not only would require some tweaking in the algorithm, but also they usually lead to unaffordable computational cost.

A significant advantage of our proposal is the ability to deal with variable-length time series. Since the evolution of all ESNs, each corresponding to one series, is simultaneous, shorter input sequences should be set to a null value. However, the corresponding reservoir states would continue their evolution autonomously, hopefully having memorised the dynamical behaviour acquired from the series. This feature is particularly advantageous in the example described in Section 3 since, although apparently all saccades have the same length, this is the result of the preprocessing, which includes a somewhat arbitrary clipping to the established length, whereas saccades themselves are intrinsically variable-length time series.

5 Experimental results

In order to provide a proof of concept for the proposed approach, we have implemented the dynamical clustering method embedded into the ESNs evolution. As an illustrative example of the key issues of the method, we use as input the saccades time series. Four kinds of experiments have been performed to assess the performance gain resulting from the introduction of the ESN, compared to conventional forms of unsupervised clustering:

– First of all, a baseline has been established by performing clustering on the vectors comprising the whole saccades, disregarding the fact that components are temporally related.

– Also, for the sake of comparison, clustering has been carried out on an instantaneous snapshot of the reservoir states, which is obtained by averaging of the states along the whole evolution.
– Another fixed view of the states results from considering the instantaneous state at the final time, once the input presentation ends.
– Finally, our main proposal is the repeated implementation of clustering iteratively at every step along the ESNs evolution.

For all experiments, the k-means algorithm has been chosen for clustering, due to its simplicity and computational efficiency.

After preliminary experimentation, the number of clusters is set to $k = 10$. The hyperparameters of the ESNs have been set as follows: the reservoir size is 1/5 of the length of the saccades, i.e. $d = 38$, which is considered enough to memorize the input dynamics while keeping a limited computational cost; the leaking rate is $l = 0.3$; finally, several runs have been performed with different values of the weight matrix spectral radius R, which is known to be a critical hyperparameter in the dynamical behaviour of the ESN. For the sake of brevity, only experiments with $R = 2$ and $R = 3$ are here reported.

A critical concern in unsupervised learning is the assessment of the quality of the result since, unlike in supervised classification, there is no obvious accuracy measure. Among the large numbers of clustering quality measures, we have chosen the silhouette coefficient [10]. For each data instance i assigned to cluster A, the silhouette coefficient s_i is defined by

$$\begin{aligned} a_i &= \frac{1}{|A|-1} \sum_{j \in A-\{i\}} d(i,j) \\ b_i &= \min_{B \neq A} \frac{1}{|B|} \sum_{j \in B} d(i,j) \\ s_i &= \frac{b_i - a_i}{\max(a_i, b_i)} \end{aligned} \tag{3}$$

It is easy to see that $-1 \leq s_i \leq 1$ and values of s_i intuitively represent the certainty that the instance i really belongs to cluster A. For the clusterings obtained in the four procedures described above, we have computed the average of the silhouette coefficients, and results are shown in Table 1. It can be observed that the clustering obtained from the reservoir states is significantly more meaningful that the one directly resulting from saccades data. This reinforces the notion that the Echo State Network, with its recurrent dynamics, is able to extract and memorise temporal features of data, which are not obvious when saccades are simply considered as a vector.

A second set of experiments aims at obtaining information from the unsupervised clustering about the health status of subjects, rather than individual saccades, thus providing supplementary information to aid a diagnosis. Therefore, we now incorporate the knowledge of the class labels by computing a *severity index* from the clustering results, by means of an averaging. Specifically, the following procedure is performed:

Clustering data	Spectral radius	
	$R = 2$	$R = 3$
Original time series	0.42	
Reservoir state average	0.43	0.51
Final reservoir state	0.52	0.49
Dynamical reservoir states	**0.54**	0.53

Table 1. Values of the silhouette coefficient for clustering with the original saccade time series, the average of the reservoir states of the corresponding ESNs, the final reservoir state after the complete presentation of the saccade, and the repeated process at every step of the ESNs evolution.

1. For each saccade, the severity value 0, 1, or 2 is assigned according to whether the saccade is labelled as C, P, or D, respectively.
2. For each cluster, these values are averaged for all the saccades that are assigned to the same cluster. The assignment results from minimum distance to the cluster centroid, among all clusters. The obtained value $SI(c) \in [0, 2]$ is considered the *severity index* of the cluster c.

The rationale behind this computation is that the classification in three disease stages C, P, D is too coarse and there is significant overlap in the characteristics of individual saccades, especially between healthy subjects and presymptomatic ones. The distribution of labels and the severity indexes of the ten clusters are shown in Figure 2 and Table 2, respectively. It is obvious that classes are not identically distributed among clusters, which suggests that unsupervised clustering has captured some information of the disease stage that could be used to inform a diagnosis. In order to rigorously confirm this postulate, we have performed a χ^2 independence test. The corresponding contingency table is shown in Table 3. The hypothesis test results in a p-value of $2.39\,10^{-56}$, i.e. negligible, thus the hypothesis that clustering is statistically independent from class labels can be rejected. A deeper analysis of the form of this dependence, and the validation of the severity index methodology by medical experts is left for further research, so we have not performed here a systematic experimentation that results in a competitive classification method.

Cluster #	0	1	2	3	4	5	6	7	8	9
S.I.	1.051	0.696	1.179	0.950	1.073	0.761	1.080	1.089	0.872	0.720

Table 2. Values of the severity index for the ten clusters, shown in the same order as in Figure 2.

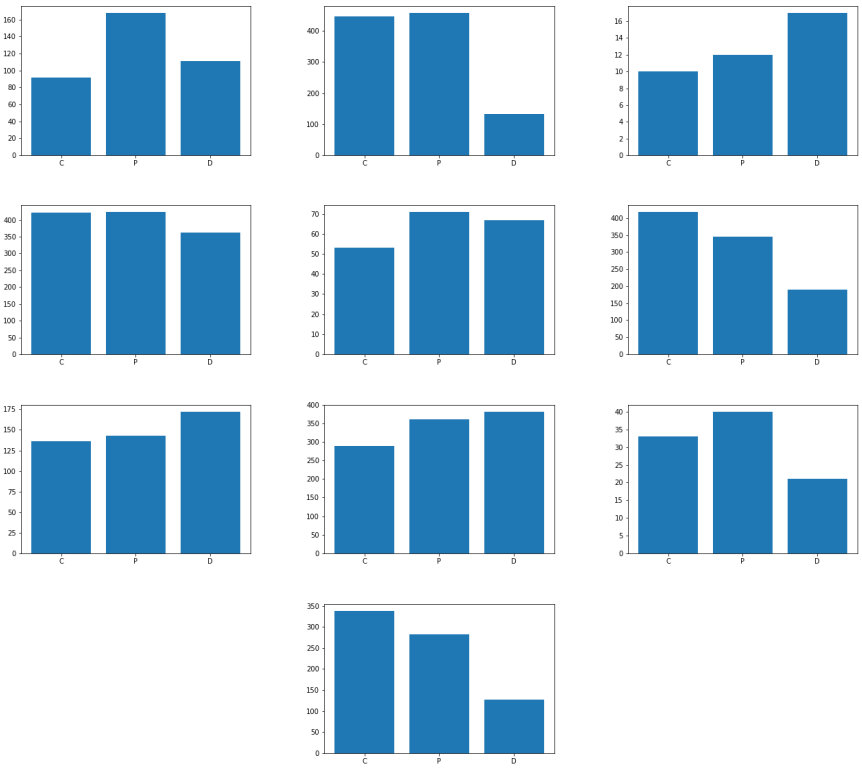

Fig. 2. Distribution of labels in the ten clusters obtained from dynamic clustering with $R = 3$. Clusters are shown in the same order as in Table 2.

Class / Cluster #	0	1	2	3	4	5	6	7	8	9
C	92	447	10	422	53	418	136	289	33	338
P	168	458	12	424	71	345	143	361	40	283
D	111	132	17	362	67	190	172	381	21	128

Table 3. Contingency table between health status classes and unsupervised clusters.

6 Conclusions

In this work we have presented a novel method for unsupervised clustering of time series, based upon the introduction of an Echo State Network. This recurrent model possesses a natural dynamics, due to the presence of feedback connections, which captures the relevant features of the presented sequences. The state values then form a data table that goes through a conventional clustering algorithm *at every time step*. The method is applied to the task of analysing eye movement records in order to aid in the diagnosis of ataxia, where usual classification methods achieve limited results, due to the mixed up labelling of individual saccades. Simulation results, using conventional measures of cluster quality, show that the obtained grouping outperforms the static clustering arisen from considering temporal data as a vector. Also, complementing this result with the information from existing labels provides a severity index that could help medical experts in the assessment of the disease stage.

Several directions of research arise as a natural extension of the current work. First of all, in the implementation of the proposed method we have resorted to the well known k-means clustering method, due to its simplicity. However, a significant advantage of our proposal is its modularity regarding the clustering technique, i.e. a more advanced clustering method can be used instead of k-means, with some mild requirements. In particular, most iterative partitioning methods will be suitable for being included into the proposed scheme. In this regard, some advanced clustering techniques have been proposed that can outperform k-means for many applications, e.g. ISODATA [6], which has the advantage that the number of clusters must not be specified beforehand. Also, since the final aim is placing the patients in a linear scale regarding their disease stage, a Self Organizing Map [7] with a one-dimensional topology could be a promising candidate, since its principle is precisely the preservation of topology under clustering. Regarding the construction of the Echo State Network, further experimentation and analysis are needed in order to provide a systematic methodology. Our experiments show that the proposed method provides satisfactory clustering results when the spectral radius is fixed at values such as 2 and 3, i.e. considerably distant from 1. This is a significant finding that somewhat puts up to discussion the common knowledge that, in order to achieve the

echo state property, the Echo State Network should be set at the *edge of criticality* through values of the spectral radius only slightly larger than one. Our ongoing work aims at a theoretical explanation of these results.

References

1. Aghabozorgi, S., Seyed Shirkhorshidi, A., Ying Wah, T.: Time-series clustering - A decade review. Information Systems 53, 16–38 (2015)
2. Becerra-García, R.A., García-Bermúdez, R., Joya, G., Fernández-Higuera, A., Velázquez-Rodríguez, C., Velázquez-Mariño, M., Cuevas-Beltrán, F., García-Lagos, F., Rodráguez-Labrada, R.: Data mining process for identification of non-spontaneous saccadic movements in clinical electrooculography. Neurocomputing 250, 28–36 (2017)
3. Ding, S., Wu, F., Qian, J., Jia, H., Jin, F.: Research on data stream clustering algorithms. Artificial Intelligence Review 43(4), 593–600 (2015)
4. Gallicchio, C., Micheli, A., Pedrelli, L.: Deep Echo State Networks for Diagnosis of Parkinson's Disease. In: Verleysen, M. (ed.) Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'18). pp. 397–402. i6doc, Bruges (2018)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. The MIT Press, Cambridge, Massachusetts (2016)
6. Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31(8), 651–666 (2010)
7. Kohonen, T.: Essentials of the self-organizing map. Neural Networks 37, 52–65 (2013)
8. Lukoševičius, M.: A Practical Guide to Applying Echo State Networks. In: Montavon, G., Orr, G.B., Müller, K.R. (eds.) Neural Networks: Tricks of the Trade, vol. 7700, pp. 659–686. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
9. Pulst, S.M.: Spinocerebellar Ataxia Type 2. In: GeneReviews®. University of Washington, Seattle (1993), http://www.ncbi.nlm.nih.gov/pubmed/20301452
10. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20(C), 53–65 (1987)
11. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.T.: A review of clustering techniques and developments. Neurocomputing 267, 664–681 (2017)
12. Velázquez-Mariño, M., Atencia, M., García-Bermúdez, R., Sandoval, F., Pupo-Ricardo, D.: Architecture for neurological coordination tests implementation. In: Lecture Notes in Computer Science. vol. 10306, pp. 26–37 (2017)
13. Verstraeten, D., Schrauwen, B., D'Haene, M., Stroobandt, D.: An experimental unification of reservoir computing methods. Neural Networks 20(3), 391–403 (2007)