Echo State Networks for Feature Selection in Affective Computing

P. Koprinkova-Hristova[1], L. Bozhkov[2], and P. Georgieva[3(✉)]

[1] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria
[2] Technical University of Sofia, Sofia, Bulgaria
[3] DETI/IEETA, University of Aveiro, Aveiro, Portugal
petia@ua.pt

Abstract. The Echo State Networks (ESNs) are dynamical structures designed initially to facilitate learning in Recurrent Neural Networks which are normally applied for time series modeling. In this paper we show that the ESN reservoirs can serve as an effective feature selection procedure that improved the discrimination of human emotion valence from EEG signals, a task that belongs to the research field of affective computing. A number of supervised and unsupervised machine learning techniques provided with the new feature vector extracted from ESN reservoir states were comparatively studied with respect to their discrimination accuracy. This novel application serves as a proof of concept for the possibility of extending the usability of the ESNs in classification or clustering frameworks.

Keywords: Echo state network · Feature selection · Affective computing · EEG data classification and clustering

1 Introduction

Echo State Networks (ESN) represent a class of recurrent neural networks (RNN) where the so called "reservoir computing" approach for training is formulated, [15]. The key idea of this biologically inspired approach is to mimic structures in human brain that seem to be composed by randomly connected dynamic non-linear neurons called reservoir whose output is usually linear combination of the current states of the reservoir neurons. The main advantage of the ESN is the simplified training algorithm since only weights of the connections from the reservoir to the readout neurons are subject to training. Thus instead of gradient descent learning much faster least squares method can be used.

Although the reservoir connections and their weights are randomly generated, in order to prevent improper behavior of ESN, the reservoir needs to possess the so called "echo state property" as formulated in [5]. The basic rule formulation is: the effect of input disturbances should vanish gradually in time, that means the dynamic reservoir must be stable. According to this rule, a reservoir weight matrix with spectral radius below one needs to be generated. However as pointed out in [15] this

Y. Demazeau et al. (Eds.): PAAMS 2015, LNAI 9086, pp. 131–141, 2015.
DOI: 10.1007/978-3-319-18944-4_11

condition will not guaranty ESN stable behavior in general. Therefore, many task-dependent recipes for improvement of reservoir connections were proposed.

Since one of the laws of thermodynamics says that any stable stationary state has a local maximum of entropy [3], it can be expected that maximization of entropy at the ESN reservoir output could increase its stability. This motivated several works proposing ESN reservoir improvement by its entropy maximization [16]. Other authors proposed the biologically motivated algorithm called Intrinsic Plasticity (IP) based on mechanisms of changing neural excitability in response to the distribution of the input stimuli [18], [19]. In [6] we have shown that in fact IP training achieves balance between maximization of entropy at the ESN reservoir output and its concentration around the pre-specified mean value increasing at the same time reservoir stability. During the investigations in [6] another interesting effect was observed: the reservoir neurons equilibrium states were concentrated in several regions. Then a question arose: is it possible to use this effect for classification or clustering purposes too? This initiated development of the proposed here algorithm for multidimensional data classification and clustering.

Since ESN are dynamic structures designed initially for time series modeling, using them for static data classification/clustering might seem odd. However the idea for using RNNs in this way is not new. There are examples in the literature like neural systems possessing multi-stable attractors [2] that perform temporal integration aimed at discrimination between multiple alternatives. In other works [1, 4] unsupervised learning procedures that minimize given energy function were proposed aiming at achievement of network equilibrium states that reflect given data structure.

Concerning ESN applications for classification or clustering, there are only few works available. In [20] it was proposed for the first time to use ESN as feature extraction stage of image classification. Their role was to “draw out” silent underlying features of the data to be used further to train a feedforward neural network classifier. In [17] the idea to exploit equilibrium states of the ESN reservoir in order to design multiple-clusters ESN reservoirs was proposed. It was inspired by complex network topologies imitating cortical networks of the mammalian brain. In [14] it was reported that using another kind of IP algorithm in combination with Spike-time Dependent plasticity (STDP) of synaptic weights changes the connectivity matrix of the network in such a way that the recurrent connections capture the peculiarities of the input stimuli so that the network activation patterns can be separated by an unsupervised clustering technique.

The idea described in this paper was motivated initially from stability analysis of ESN and proposed for the first time in [7]. It exploits similar reservoir properties reported by other works but looking from a different point of view: to consider combinations between steady states of each two neurons in the reservoir as numerous two-dimensional projections of the original multidimensional data fed into the ESN input; next to use these low dimensional projections for classification or clustering of the original multidimensional data. The ESN feature selection methodology proposed in [7] was successfully tested on a number of different data sets [8-13] to solve clustering problems.

In this paper we go further and apply it for the first time to a binary classification problem. We also compare classification and clustering technique for discrimination of positive and negative emotional states of multiple subjects.

2 Echo State Networks Basics

The basic structure of an ESN, presented in Fig. 1, consists of a reservoir of randomly connected dynamic neurons with sigmoid nonlinearities fres (usually hyperbolic tangent):

$$r(k) = f^{res}\left(W^{in} in(k) + W^{res} r(k-1)\right) \tag{1}$$

and a linear readout f^{out} (usually identity function) at the output:

$$out(k) = f^{out}\left(W^{out}\left[in(k) \quad r(k)\right]\right) \tag{2}$$

Here k denotes discrete time instant; $in(k)$ is a vector of network inputs, $r(k)$ - a vector of the reservoir neurons states and $out(k)$ – a vector of network outputs; n_{in}, n_{out} and n_r are the dimensions of the corresponding vectors in, out and r respectively; W^{out} is a trainable $n_{out} \times (n_{in}+n_r)$ matrix; W^{in} and W^{res} are $n_r \times n_{in}$ and $n_r \times n_r$ matrices that are randomly generated and are not trainable. In some applications direct connection from the input to the readout is omitted.

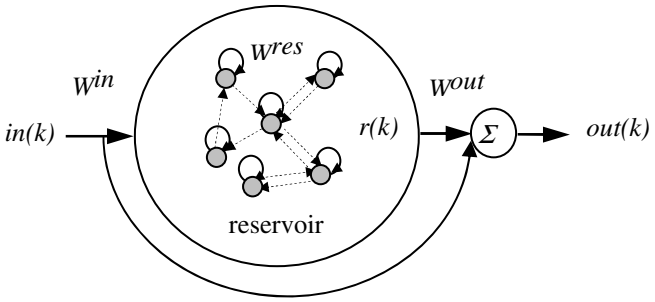

Fig. 1. Echo state network basic structure

The key idea is that having rich enough reservoirs of nonlinearities will allow to approximate quite complex nonlinear dependence between input and output vectors by tuning only the linear readout weights. Hence the training procedure is simplified to solving in one step Least Squares task [5].

Although this idea seems to work well, it appears that initial tuning of reservoir connections to the data that will be fed into the ESN helps to improve its properties. In [18], [19] was proposed a reservoir tuning approach called “intrinsic plasticity” (IP). It is aimed at maximization of information transmission through the ESN that is equivalent to its output entropy maximization. Motivation of this approach is related

to known biological mechanisms that change neural excitability according to the distribution of the input stimuli. The authors proposed a gradient method for adjusting the biases and an additional gain term aimed at achieving the desired distribution of outputs by minimizing the Kullback-Leibler divergence:

$$D_{KL}\left(p(r), p_d(r)\right) = \int p(r) \log\left(\frac{p(r)}{p_d(r)}\right) \tag{3}$$

That is a measure for the difference between the actual *p(r)* and the desired $p_d(r)$ probability distribution of reservoir neurons output *r*. Since the commonly used transfer function of neurons is the hyperbolic tangent, the proper target distribution that maximizes the information at the output according to [18] is the Gaussian one with a prescribed small variance σ and a zero mean μ:

$$p_d(r) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right) \tag{4}$$

Hence equation (3) can be rearranged as follows:

$$D_{KL}(p(r), p_d(r)) = -H(r) + \frac{1}{2\sigma^2} E\left((r-\mu)^2\right) + \log\frac{1}{\sigma\sqrt{2\pi}} \tag{5}$$

Where *H(r)* is entropy, the last term is constant and the second one determines the deviation of the output from the desired mean value. Thus minimization of (5) will lead to compromise between entropy maximization and minimization of distance between μ and *r*.

In order to achieve those effects two additional reservoir parameters - gain *a* and bias *b* (both vectors with n_r size) - are introduced as follows:

$$r(k) = f^{res}\left(diag(a) W^{in} in(k) + diag(a) W^{res} r(k-1) + b\right) \tag{6}$$

The IP training is a procedure that adjusts vectors *a* and *b* using gradient descent.

3 Affective Computing and Data Set Description

We consider learning to discriminate emotional states of human subjects, based on their brain activity observed via Event Related Potentials (ERPs). ERPs are transient components in the EEG generated in response to a stimulus. ERPs were collected while subjects were viewing high arousal images with positive or negative emotional content. This problem is important because such classifiers constitute "virtual sensors" of hidden emotional states, which are useful in psychology science research and clinical applications [21], [22].

A total of 26 female volunteers participated in the study. The signals were recorded while the volunteers were viewing high arousal images with positive and negative valence. For each image, signals from 21 EEG channels were sampled at 1000Hz and stored (see Table 1). The signals were recorded while the volunteers were viewing pictures selected from the International Affective Picture System (IAPS) repository.

A total of 24 high arousal (IAPS rating> 6) images with positive valence (M=7.29 +/- 0.65) and negative valence (M=1.47 +/- 0.24) were selected. Each image was presented 3 times in a pseudo-random order and each trial lasted 3500 ms: during the first 750 ms, a fixation cross was presented, then one of the images was presented during 500 ms and at last a black screen appeared during 2250 ms. The raw EEG signals were first filtered (band-pass filter between 0.1 and 30Hz.), eye-movement corrected, baseline compensated and segmented into epochs using NeuroScan software. The single-trial signal length is 950 ms with 150ms before the stimulus onset. The ensemble average for each condition (positive/negative valence) was also computed and filtered using a Butterworth filter of 4th order with passband [0.5-15] Hz. Thus, the filtered ensemble average signals cover the frequency band ranges corresponding to Delta ([0.5 -4] Hz), Theta ([4 -8] Hz) and Alpha neural activity ([8 -12] Hz).

Temporal features (amplitudes and latencies) are extracted from the filtered, segmented and ensemble averaged ERP data. Starting by the localization of the first minimum after time t=0s, the features are defined as a sequence of the local positive and negative picks, and their respective latencies (time of occurrence). Twelve temporal features are stored (Table 2) corresponding to the amplitudes of the first three local minimums (A_{min1}, A_{min2}, A_{min3}), the first three local maximums (A_{max1}, A_{max2}, A_{max3}), and their associated latencies (L_{min1}, L_{min2}, L_{min3}, L_{max1}, L_{max2}, L_{max3}).

Table 1. Channels

Nº	EEG Channels
1	Ch 1 (FP1)
2	Ch 2 (FPz)
3	Ch 3 (FP2)
4	Ch 4 (F7)
5	Ch 5 (F3)
6	Ch 6 (Fz)
7	Ch 7 (F4)
8	Ch 8 (F8)
9	Ch 9 (T7)
10	Ch 10 (C3)
11	Ch 11 (Cz)
12	Ch 12 (C4)
13	Ch 13 (T8)
14	Ch 14 (P7)
15	Ch 15 (P3)
16	Ch 16 (Pz)
17	Ch 17 (P4)
18	Ch 18 (P8)
19	Ch 19 (O1)
20	Ch 20 (Oz)
21	Ch 21 (O2)

Table 2. Features

Nº	Features
1	$A_{min1(A1)}$
2	A_{max1}
3	A_{min2}
4	A_{max2}
5	A_{min3}
6	A_{max3}
7	L_{min1}
8	L_{max1}
9	L_{min2}
10	L_{max2}
11	L_{min3}
12	L_{max3}

As a result, the initial feature set is a matrix X with dimension of 252 columns (21 channels x12 features) and 52 lines (the ensemble averaged positive and negative labeled trials of 26 subjects). The ESN-based features selection discussed in the next section is applied on the normalized feature matrix

$$\bar{X} = \frac{X - mean(X)}{std(X)} \tag{7}$$

4 ESN for Feature Selection

The original feature matrix (7) was processed via ESN reservoir following the procedure developed in [8-13]. The two-step algorithm is outlined in Table 3:

Step 1: IP tuning of the ESN reservoir using original feature data set;

Table 3. Algorithm to obtain the new feature vector as a vector of equilibrium states of neurons in the ESN reservoir

```
in(1:features number,1:examples number)=original_features;
n_in=features number; n_out=1; n_r=chosen number;
esn=generate_esn(n_in, n_out, n_r);
for it=1:number of IP iterations
        for i=1:examples number
                esn=esn_IP_training(esn, in(:,i));
        end
end
for i=1:examples number
        r(0)=0;
        for k=1:chosen number of steps
                r(k)=sim_esn(esn, in(:,i),r(k-1));
        end
        r_e(i)=r(k);
end
esn_features=r_e;
```

The structure of ESN was determined according to the size of the original feature matrix (7) so that the size of the ESN input vector corresponds to the number of the original features (in this case 252 features). Since we explore only the reservoir output, the size of the readout doesn't matter and it was set to one. The size of the reservoir varies starting from 10 up to 500 neurons in order to study its influence on the accuracy of the emotion valence discrimination. Our experimental results with 10, 30, 50, 100, 150, 300 and 500 neurons are visualized in the next section. The IP tuning was done by presenting one by one the feature vector of all training examples to the ESN input over a predefined number of iterations (we used 10 iterations) and adjusting the gain and the bias terms using gradient rules from [18].

Step 2: Calculating of the equilibrium states of all reservoir neurons.

Since it is hard to solve analytically the equation for equilibrium states corresponding to each input data in_c:

$$r_e = \tanh\left(diag(a)W^{in}in_c + diag(a)W^{res}r_e + b\right),$$

the equilibrium states r_e were determined by simulations for a previously chosen number of steps until reaching of steady state (in our experiments 25 steps were enough). The achieved reservoir neurons equilibrium states were kept as the new feature vector called further *esn_features*.

5 Emotion Valence Discrimination – Experimental Results

In this section we use the ESN extracted features (*esn_features*) in order to discriminate the positive and negative emotion valence applying supervised (classification) or unsupervised (clustering) learning techniques. Two approaches related with the new feature space were studied:

Approach 1: Using all possible 2D combinations between equilibrium states $r_e(i)$ and $r_e(j)$ of every two neurons *i* and *j* from the ESN reservoir as a 2D feature vector.

This approach actually maps the original feature data set into a bigger space of reservoir equilibriums, i.e. we first expand the feature data set and then select the best 2D projections among all possible combinations.

Approach 2: Using all reservoir equilibrium states *esn_features* as the new feature vector.

In contrast to the first approach, Approach 2 maps the original feature data set into a smaller size reservoir and thus the new feature set has a smaller dimension. This approach is analog to the PCA (Principal Component Analysis) where a feature reduction is first performed before the classification or clustering.

In the next section Approach 1 and Approach 2 are applied to two basic clustering algorithms, k-means and fuzzy C-means (FCM).

5.1 Data Clustering

In Fig. 2 are summarized the discrimination accuracies of k-means and FCM. It should be noted that Approach 1 produces a variety of 2D feature sets and in Fig.2 are presented the accuracy results only for the best 2D feature sets. Among the huge number of 2D feature combinations, only few of them achieve these results. For comparative purposes we also present the clustering accuracy of k-means and FCM using the original feature matrix (7).

From Fig. 2 it can be concluded that for all reservoir sizes (nr=10, 30, 50, 100, 150, 300, 500) the best clustering was obtained with Approach 1 (a combination of 2D

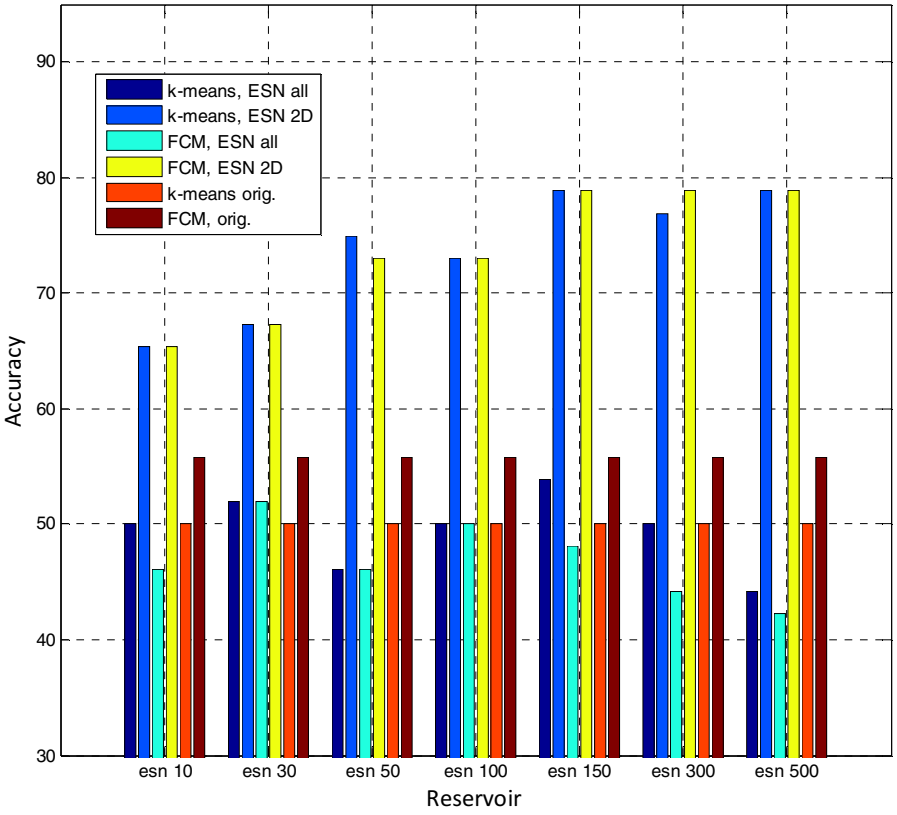

Fig. 2. Accuracy of all clustering algorithms using different features sets

feature sets). The clustering accuracy using all ESN reservoir states (*esn_features*) seems comparable and even worse (especially in the case of bigger reservoir size) than those obtained by direct clustering of the original features. Higher the reservoir size is, better is the clustering accuracy in the 2D feature scenario which goes close to 80%. Another interesting observation is that FCM outperforms k-means clustering when directly applied to the original feature matrix, while using the ESN extracted features seem to make both approaches similar. In the case of using all *esn_features* k-means outperforms slightly FCM while in the case of 2D feature vector both algorithms achieve similar accuracy.

5.2 Data Classification

The same ENS models were tested for the case of supervised learning to discriminate the two emotion valences. In order to eliminate the problem of choosing the "wrong" or the "lucky" model, we applied a number of standard classifiers, namely Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN), Naïve Bayes (NB),

Support Vector Machines (SVM) and Decision Trees (DT). Due to the limited number of examples (only 26 subjects), cross validation with leave-one-out subject is adopted. In order to increase the statistical confidence of the obtained results, classification based on the majority votes of the classifiers (LD, kNN, NB, SVM, and DT) was also done. We call this hierarchical classification methodology VOTE.

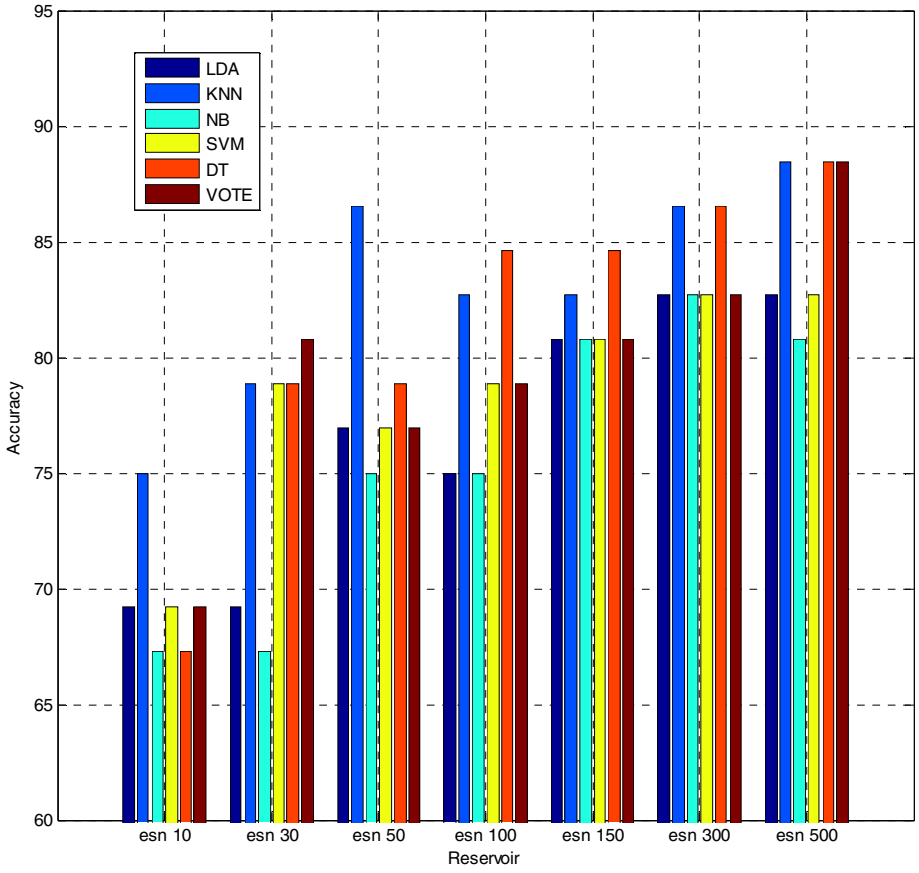

Fig. 3. Accuracy of all classification algorithms using different features sets

The results in Fig.3 show the same tendency of better accuracy for increasing number of neurons in the ENS reservoir. The intuition behind this is that higher the reservoir size, more binary combinations of neurons are produced and thus the probability of getting a good feature selection increases. An interesting observation but difficult to explain is the fact that the VOTE classifier does not always produce the best classification. In previous studies [23] we have obtained quite encouraging results with VOTE classifier in the framework of different feature selection scenarios. However, for bigger reservoir size, VOTE improves and approaches the expected performance.

Another interesting observation is the surprisingly good performance of the kNN and DT classifiers. Finally, comparing the bars in Fig.2 and Fig.3, it can be concluded

that the supervised learning (classification) outperforms significantly the unsupervised (clustering) approach, which is not a surprising result.

6 Conclusions

In this paper we propose the ESN as a mechanism for feature selection in two scenarios: i) map the original features into an expanded feature space defined by the number of the reservoir neurons (more neurons than ENS inputs) and choose the best combination of two neurons (2D projection) as the new features; ii) map the original features into a reduced feature space defined by the number of the reservoir neurons (less neurons than ENS inputs) and use all of them as the new features. Both scenarios were tested on the challenging problem of affective computing based on brain neural data (ERPs). In the 2D projection scenario it is always possible to find a combination of features that will cluster or classify the data with reasonable accuracy (close to 80% for the clustering and close to 89% for the classification task).

The computational complexity is however an unavoidable problem particularly when the reservoir size increases. Moreover from the very big number of neuron combinations (for example 124750 combinations for nr=500) only few of them reveal to be the proper choice.

Nevertheless, these proof of concept results encourage us to further test the ESN as a feature selection step prior to data classification/clustering in other applications.

References

1. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for Boltzmann machines. Cognitive Science **9**, 147–169 (1985)
2. Brody, C.D., Romo, R., Kepecs, A.: Basic mechanisms for graded persistent activity: Discrete attractors, continuous attractors, and dynamical representations. Current Opinion in Neurobiology **13**, 204–211 (2003)
3. Haddad, W.M., Chellaboina, V.S., Nersesov, S.G.: Thermodynamics: A Dynamical System Approach. Princeton University Press (2005)
4. Hinton, G.E., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
5. Jaeger, H.: Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach, GMD Report 159, German National Research Center for Information Technology (2002)
6. Koprinkova-Hristova, P., Palm, G.: ESN Intrinsic Plasticity versus Reservoir Stability. In: Honkela, T. (ed.) ICANN 2011, Part I. LNCS, vol. 6791, pp. 69–76. Springer, Heidelberg (2011)
7. Koprinkova-Hristova, P., Tontchev, N.: Echo State Networks for Multi-dimensional Data Clustering. In: Villa, A.E., Duch, W., Érdi, P., Masulli, F., Palm, G. (eds.) ICANN 2012, Part I. LNCS, vol. 7552, pp. 571–578. Springer, Heidelberg (2012)
8. Koprinkova-Hristova, P., Alexiev, K., Borisova, D., Jelev, G., Atanassov, V.: Recurrent neural networks for automatic clustering of multispectral satellite images. In: Bruzzone, L. (ed.) Proceedings of SPIE, Image and Signal Processing for Remote Sensing XIX, 88920X, vol. 8892 (October 17, 2013) doi:10.1117/12.

9. Koprinkova-Hristova, P., Angelova, D., Borisova, D., Jelev, G.: Clustering of spectral images using Echo state networks. In: 2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications, IEEE INISTA 2013, June 19-21, Albena, Bulgaria (2013). doi:10.1109/INISTA.2013.6577633
10. Koprinkova-Hristova, P., Doukovska, L., Kostov, P.: Working regimes classification for predictive maintenance of mill fan systems. In: 2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications, IEEE INISTA 2013, June 19-21, Albena, Bulgaria (2013) doi:10.1109/INISTA.2013.6577632
11. Koprinkova-Hristova, P., Alexiev, K.: Echo State Networks in Dynamic Data Clustering. In: Mladenov, V., Koprinkova-Hristova, P., Palm, G., Villa, A.E., Appollini, B., Kasabov, N. (eds.) ICANN 2013. LNCS, vol. 8131, pp. 343–350. Springer, Heidelberg (2013)
12. Koprinkova-Hristova, P., Alexiev, K.: Sound fields clusterization via neural networks. In: 2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications, INISTA 2014, June 23–25, Alberobello, Italy, pp. 368–374 (2014)
13. Koprinkova-Hristova, P., Alexiev, K.: Dynamic Sound Fields Clusterization Using Neuro-Fuzzy Approach. In: Agre, G., Hitzler, P., Krisnadhi, A.A., Kuznetsov, S.O. (eds.) AIMSA 2014. LNCS, vol. 8722, pp. 194–205. Springer, Heidelberg (2014)
14. Lazar, A., Pipa, G., Triesch, J.: Predictive Coding in Cortical Microcircuits. In: Kurková, V., Neruda, R., Koutník, J. (eds.) ICANN 2008, Part II. LNCS, vol. 5164, pp. 386–395. Springer, Heidelberg (2008)
15. Lukosevicius, M., Jaeger, H.: Reservoir computing approaches to recurrent neural network training. Computer Science Review **3**, 127–149 (2009)
16. Ozturk, M., Xu, D., Principe, J.: Analysis and design of Echo state networks. Neural Computation 19, 111–138 (2007)
17. Peng, X., Guo, J., Lei, M., Peng, Yu.: Analog Circuit Fault Diagnosis with Echo State Networks Based on Corresponding Clusters. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) ISNN 2011, Part I. LNCS, vol. 6675, pp. 437–444. Springer, Heidelberg (2011)
18. Schrauwen, B., Wandermann, M., Verstraeten, D., Steil, J.J., Stroobandt, D.: Improving reservoirs using intrinsic plasticity. Neurocomputing **71**, 1159–1171 (2008)
19. Steil, J.J.: Online reservoir adaptation by intrinsic plasticity for back-propagation-decoleration and echo state learning. Neural Networks **20**, 353–364 (2007)
20. Woodward, A., Ikegami, T.: A reservoir computing approach to image classification using coupled echo state and back-propagation neural networks. In: Proc. of 26th Int. Conf. on Image and Vision Computing, Auckland, New Zealand, November. 29–December 1, 2011, pp. 543–458 (2011)
21. Calvo, R.A., D'Mello, S.K.: Affect Detection: An Interdisciplinary Review of Models, Methods, and their Applications. IEEE Transactions on Affective Computing **1**(1), 18–37 (2010)
22. Georgieva, O., Milanov, S., Georgieva, P., Santos, I.M., Pereira, A.T., da Silva, C.F.: Learning to decode human emotions from ERPs. Neural Computing and Applications, Springer, On-line Access (2014). doi:10.1007/s00521-014-1653-6
23. Bozhkov, L., Georgieva, P., Trifonov, R.: Brain Neural Data Analysis Using Machine Learning Feature Selection and Classification Methods. In: Mladenov, V., Jayne, C., Iliadis, L. (eds.) EANN 2014. CCIS, vol. 459, pp. 123–132. Springer, Heidelberg (2014)