International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 6 (2019), pp. 842-847
© International Research Publication House. http://www.irphouse.com

A Hybrid PCA-Fuzzy-ELM to Predict QSARs for the Inhibition of Dihydrofolate Reductase by Pyrimidines

B. Shanmuga Priya[1] **and Rajesh Reghunadhan**[2]
[1] Ph.D. Student, Bharathiar University, Coimbatore – 641046, Tamilnadu, India.
[2]Department of Computer Science, Central University of Kerala, Periya – 671316, Kasaragod, Kerala, India.

Abstract

Fully connected fuzzy inference systems (F-CONFIS) proposed recently by Wang et.al. when combined with Extreme Learning Machines (ELMs) outperforms other classifiers. We propose a hybrid PCA-Fuzzy-ELM method in this paper. In this hybrid method, firstly the dimension of the dataset is reduced using PCA, secondly a modified fully connected fuzzy inference system (MF-CONFIS) is designed and finally ELM is used to train the MF-CONFIS. The classification results based on benchmarking datasets shows the merits of the proposed hybrid PCA-Fuzzy-ELM. Finally, the proposed hybrid PCA-Fuzzy-ELM is used to Predict QSARs for the Inhibition of Dihydrofolate Reductase by Pyrimidines.

Keywords: Extreme Learning Machine, Fully Connected Fuzzy Inference System, Modified Fully Connected Fuzzy Inference System, Hybrid PCA-Fuzzy-ELM

I. INTRODUCTION

Classification, one of the major components in data mining, has been of greater interest to researchers for several decades. Many machine learning algorithms are available for classification, namely, neural networks [1], support vector machines [2], fuzzy rule based classifiers [3], K-nearest neighbourhood classifiers [4], Decision Trees [5], Bayesian network Classifiers [6], etc. For the datasets with more number of attributes, the computational/time complexity of the machine algorithm or the classifier increases and hence there is also a need of reducing the number of attributes (features) using methods like linear discriminant analysis (LDA), principle component analysis (PCA), etc.

Recently, Extreme learning machine (ELM) [7] was proposed by G.-B. Huang, et.al. ELMs are iteration less feedforward neural networks with random weights between the input layer and the hidden layer. The output weights between the hidden layer and the output layer can be easily found out by pseudo matrix inversions. These extreme learning machines are able to learn much faster from the known examples than the other learning algorithms. There are several extensions of the ELM including fuzzy ELM [8, 9]. The number of fuzzy rules, and hence the number of neurons in the hidden layer increases exponentially with the increase in the number of attributes and also with the increase in the number of membership functions per attribute.

This paper proposes a hybrid method, namely, PCA-FUZZY-ELM. In this proposed method the attributes (input variables) are dimension-reduced using PCA and the resultant attributes are fuzzified based on the idea to form a modified fully connected fuzzy inference system (MF-CONFIS). The consequent parts of the MF-CONFIS are found using ELM techniques. Experimental result on a benchmarking dataset using the hybrid PCA-Fuzzy-ELM shows the merits of the proposed method. The proposed method is also used to predict quantitative structure activity relationships (QSARs) for the Inhibition of Dihydrofolate Reductase by Pyrimidines and the results are promising.

II. RELATED WORK

This section discusses about extreme learning machines and fully connected fuzzy inference systems.

II.I. Extreme Learning Machines

Extreme learning machines are feedforward networks [10, 11], where the weights between the input layer and the hidden layer are randomly generated and the output weights are calculated by matrix inversion.

Let $X = \{x_1, x_2, \cdots, x_n\}$ be the set of n number of records, where $x_i = \{x_{i1}, x_{i2}, \cdots, x_{im}\}$ represent the m number of features for the i^{th} record. $t_i = \{t_{i1}, t_{i2}, \cdots, t_{ic}\}$ represent the class label for i^{th} record where c is the number of classes and each $t_{ij} \in \{0,1\}$.

Let p be the number of hidden nodes, g be the transfer function, and H^{Ψ} be the Moore-Penrose generalized inverse of H. The weights connecting the hidden layer and the output layer can be obtained by $\beta = H^{\Psi}T$, where

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{n1} & \beta_{n2} & \cdots & \beta_{nm} \end{bmatrix}$$

$$H = \begin{bmatrix} g\left(\sum_{i=1}^{m} w_{i1}x_{1i} + b_1\right) & g\left(\sum_{i=1}^{m} w_{ip}x_{1i} + b_p\right) \\ g\left(\sum_{i=1}^{m} w_{i1}x_{2i} + b_1\right) & g\left(\sum_{i=1}^{m} w_{ip}x_{2i} + b_p\right) \\ & \\ g\left(\sum_{i=1}^{m} w_{i1}x_{ni} + b_1\right) & g\left(\sum_{i=1}^{m} w_{ip}x_{ni} + b_p\right) \end{bmatrix}$$

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1c} \\ t_{21} & t_{22} & \cdots & t_{2c} \\ \vdots & \vdots & \vdots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nc} \end{bmatrix}$$

Latest advances and application of extreme learning machines are, but not limited to, Electroencephalography (EEG) signal classification [12], nonlinear learning over network using distributed-ELM [13], indoor localization algorithm based on online sequential ELM [14], multiple instance learning based on ELM [15], remote sensing image classification [16], multi-dimensional ELM [17], semi-supervised ELM [18], binary/ternary extreme learning machines [19], parallel online sequential ELM [20], multiple kernel ELM [21], non-redundant synergy pattern based graph classification [22], smile detection [23], facial age range estimation [24], classification of mental tasks from EEG signals [25], optimization of biodiesel engine performance [26], prediction of monthly effective drought index in eastern Australia [27], predictive control strategy [28], multivariate time series online sequential prediction [29], etc.

II.II. Fully Connected Fuzzy Inference Systems

Three layer fully connected fuzzy neural network (F-CONFIS) [30, 31] consists of three layers as shown in figure 1. F-CONFIS is an improvement over Adaptive Neuro Fuzzy Inference System (ANFIS) [32] in terms of speed (time complexity) and memory (space complexity).

The number of neurons in the input layer of F-CONFIS is m. There are L hidden neurons in the hidden layer. Each hidden layer receives m inputs. The transfer functions in the hidden layers are the exponentials of the sum of the inputs. The weights between the input layer and the hidden layer is given by

$$v_{ij} = \log\left(A_{ir_i(j)}(x_i)\right)$$

where $i = 1,2,\cdots,n$ represents the index of the input variables. $j = 1,2,\cdots,L$ represent the index of the membership function of the variable. The number of MFs for fuzzy variable x_i is R_i.

$r_1(l) = l\% R_1$ and

$r_i(l) = \left(l \big/ \prod_{k=1}^{i-1} R_k\right)\% R_1$ for $i = 2,\cdots,n$

The number of output neurons is c and equal to the number of classes.

The upper bound [30] of the number of patterns (P) or instances that F-CONFIS can learn with m inputs, c outputs and L neurons in the hidden layer and each input variable has s_i membership functions with P_{s_i} parameters is given by

$$P \le \frac{\left(\sum_{i=1}^{m} s_i P_{s_i} + Lc\right)}{c}$$

III. PROPOSED HYBRID PCA-FUZZY-ELM

This section proposes a hybrid PCA-FUZZY-ELM method with combination of (i) iteration less feedforward neural network (the ELM) for fast learning, (ii) fuzziness in the layer between input layer and hidden layer by the concept of modified fully connected fuzzy system (MF-CONFIS) for dealing with uncertainty, vagueness and ambiguity in the data, (iii) principle component analysis to reduce the dimensionality of the data set thereby indirectly reducing the number of fuzzy rules and the number of hidden neurons.

In the proposed algorithm, the dimension of the training data is reduced using PCA. Then MF-CONFIS is constructed as shown in figure 2. The number of neurons in the input layer of MF-CONFIS is m. The number of output neurons is c and equal to the number of classes. There are L hidden neurons in the hidden layer. Each hidden layer receives m inputs. The transfer functions in the hidden layers are the sigmoid of the m^th^ root of the exponential of the sum of the inputs. The weights between the input layer and the hidden layer is given by $v_{ij} = \log\left(A_{ir_i(j)}(x_i)\right)$ where $i = 1,2,\cdots,n$ represents the index of the input variables. $j = 1,2,\cdots,L$ represent the index of the membership function of the variable. The number of MFs for fuzzy variable x_i is R_i.

$r_1(l) = l\% R_1$ and

$r_i(l) = \left(l \big/ \prod_{k=1}^{i-1} R_k\right)\% R_1$ for $i = 2,\cdots,n$

Then MF-CONFIS is trained using ELM with the known dimension-reduced dataset. Then for any new unknown dataset the PCA is used to reduce the dimension and then passed to the trained MF-CONFIS to predict the class in which the data belongs. Figure 3 shows the diagram for the proposed hybrid PCA-Fuzzy-ELM. Steps of the proposed hybrid PCA-Fuzzy-ELM are shown in Algorithm 1.

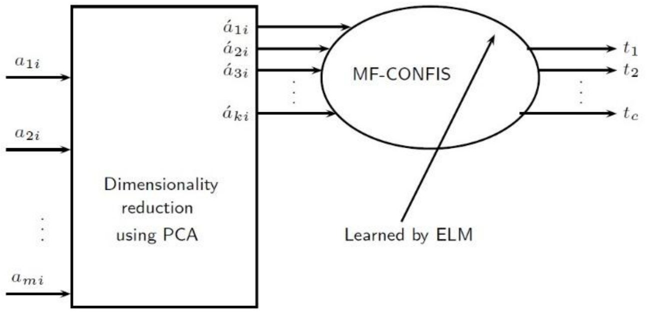

Figure 3: Hybrid PCA-FUZZY-ELM

III.I Classification of Small Round Blue-Cell Tumors

The cancer microarray data set (small round blue-cell tumors (SRBCTs) [33]) is one of the highly challenging dataset with four distinct diagnostic categories namely, Ewing's family of tumors (EWS), neuroblastoma (NB), non-Hodgkin lymphoma (Burkitt's lymphoma, BL) and rhabdomyosarcoma (RMS). It consists of 83 samples (29-EWS, 11-BL, 18-NB, 25-RMS) with 2308 genes. Leave one out classification (LOOC) method is used for dividing the samples into training and testing set.

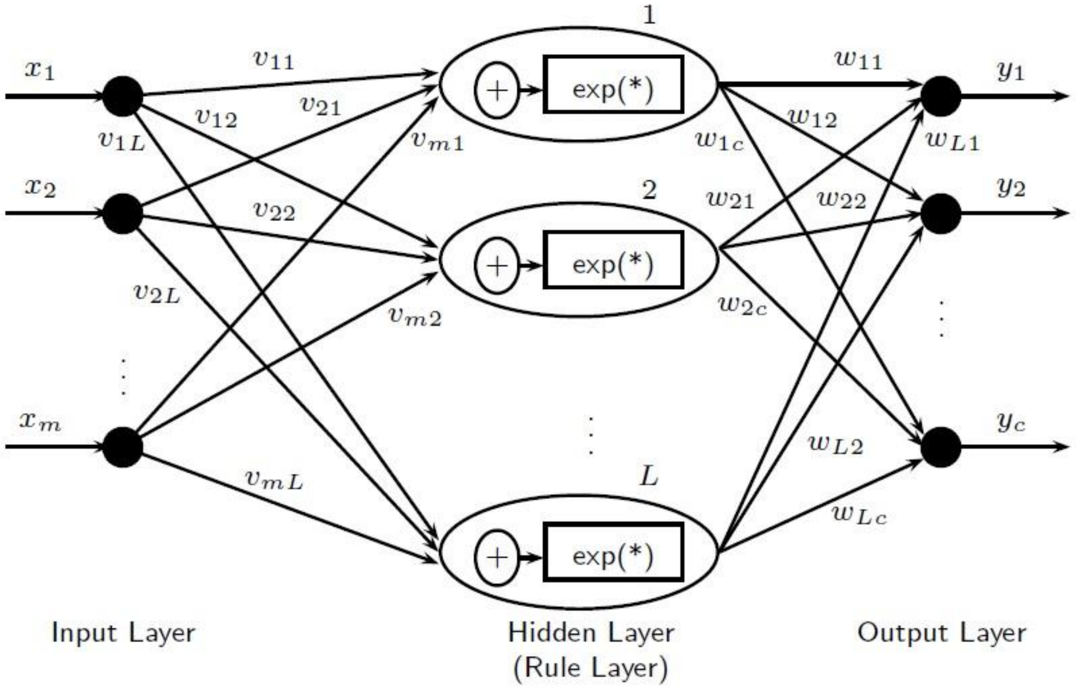

Figure 1: Fully Connected Fuzzy Inference System (F-CONFIS).

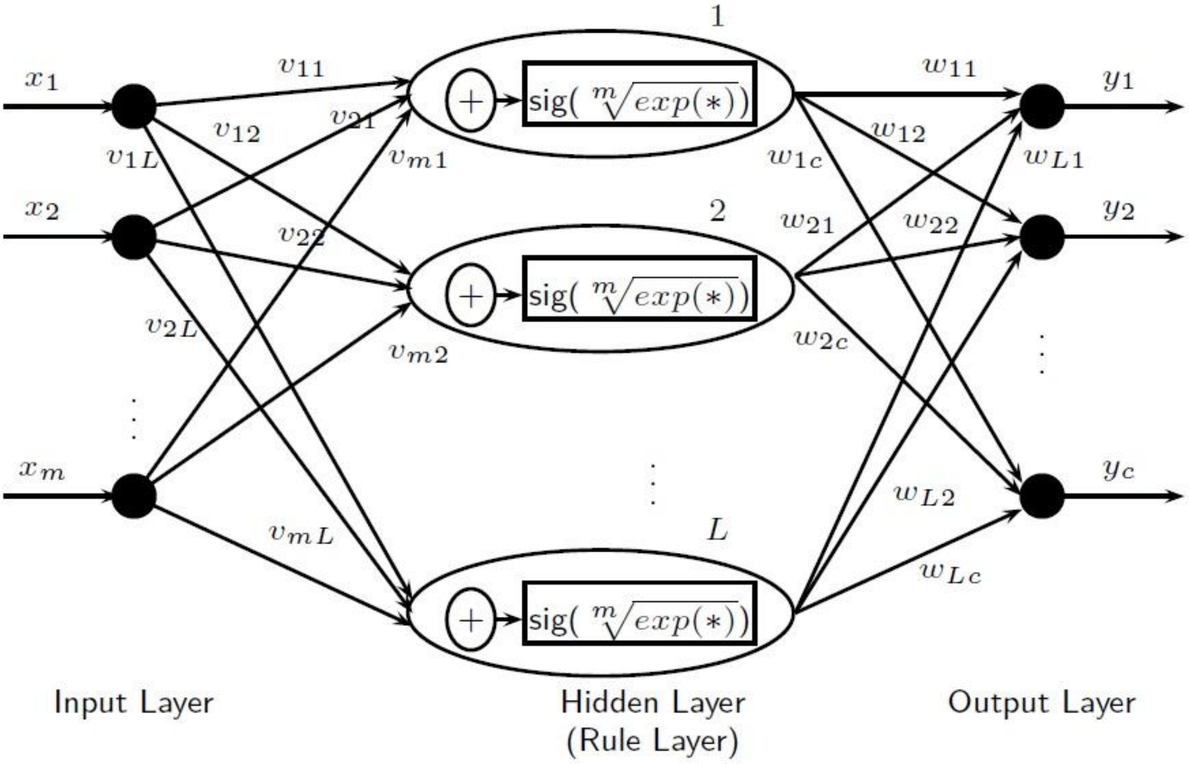

Figure 2: Modified Fully Connected Fuzzy Inference System (MF-CONFIS)

Algorithm 1 Hybrid PCA-FUZZY-ELM

Let $R = \{r_1, r_2, \cdots, r_n\}$ *be the set of 'n' records/instances/patterns of a dataset. Each record contains values for 'm' number of attributes/variables/parameters. For example i*th *record can be given as* $r_i = \{a_{1i}, a_{2i}, \cdots, a_{mi}\}$, *where* a_{ji} *is the value of j*th *attribute in the i*th *record.* $t_i = \{t_{i1}, t_{i2}, \cdots, t_{ic}\}$, *represents the class label for ith record where c is the number of classes with each* $t_{ij} \in \{0,1\}$, *and* $T = transpose(\{t_1, t_2, \cdots, t_n\})$ *with size* $n \times c$

1. *Transform the 'm' attributes values to 'k' linearly uncorrelated attributes by keeping only the first 'k' principal components by applying principle component analysis (PCA), so that* $r_i = \{\hat{a}_{1i}, \hat{a}_{2i}, \cdots, \hat{a}_{ki}\}$, *where* $\hat{a}_{ji}$ *is the value of j*th *uncorrelated attribute in the i*th *record.*
2. *Construct MF-CONFIS by choosing the number of membership functions for each of the 'k' attributes according to Wang et. al. [30] in such a way that the number of records is less than* $\left(\sum_{i=1}^{k} s_i P_{s_i} + M\prod_{j=1}^{k} s_j\right)\Big/ c$, *where* s_i *is the number of membership functions for the ith variable and* P_{s_i} *is the number of parameters of the* s_i^{th} *membership function.*
3. *Choose the centre and spread of the membership functions in such a way that the following H matrix have less number of elements with the value equal to zero, where 'sig' is the sigmoid function.*

$$H = \begin{bmatrix} sig\left(\sqrt[k]{e^{\left(\sum_{i=1}^{k}\log\left(A_{ir_i(1)}(x_{1i})\right)\right)}}\right) & \cdots & sig\left(\sqrt[k]{e^{\left(\sum_{i=1}^{k}\log\left(A_{ir_i(p)}(x_{1i})\right)\right)}}\right) \\ sig\left(\sqrt[k]{e^{\left(\sum_{i=1}^{k}\log\left(A_{ir_i(1)}(x_{2i})\right)\right)}}\right) & \cdots & sig\left(\sqrt[k]{e^{\left(\sum_{i=1}^{k}\log\left(A_{ir_i(p)}(x_{2i})\right)\right)}}\right) \\ \vdots & \vdots & \vdots \\ sig\left(\sqrt[k]{e^{\left(\sum_{i=1}^{k}\log\left(A_{ir_i(1)}(x_{ni})\right)\right)}}\right) & \cdots & sig\left(\sqrt[k]{e^{\left(\sum_{i=1}^{k}\log\left(A_{ir_i(p)}(x_{ni})\right)\right)}}\right) \end{bmatrix}$$

$$= \begin{bmatrix} sig\left(\sqrt[k]{\prod_{i=1}^{k} A_{ir_i(1)}(x_{1i})}\right) & \cdots & sig\left(\sqrt[k]{\prod_{i=1}^{k} A_{ir_i(p)}(x_{1i})}\right) \\ sig\left(\sqrt[k]{\prod_{i=1}^{k} A_{ir_i(1)}(x_{2i})}\right) & \cdots & sig\left(\sqrt[k]{\prod_{i=1}^{k} A_{ir_i(p)}(x_{2i})}\right) \\ \vdots & \vdots & \vdots \\ sig\left(\sqrt[k]{\prod_{i=1}^{k} A_{ir_i(1)}(x_{ni})}\right) & \cdots & sig\left(\sqrt[k]{\prod_{i=1}^{k} A_{ir_i(p)}(x_{ni})}\right) \end{bmatrix}$$

4. *Determine the output weights by considering MF-CONFIS as an extreme learning machine using* $\beta = H^{\Psi}T$ *, where* H^{Ψ} *is the Moore-Penrose generalized inverse of H.*

We consider first eight projected features provided by PCA. Three membership functions are used to represent each of the eight features (Figure 4 shows the membership function used for each of these eight features). Thus, 3^8 rules can be formed and hence there needs 3^8 hidden neurons (L).

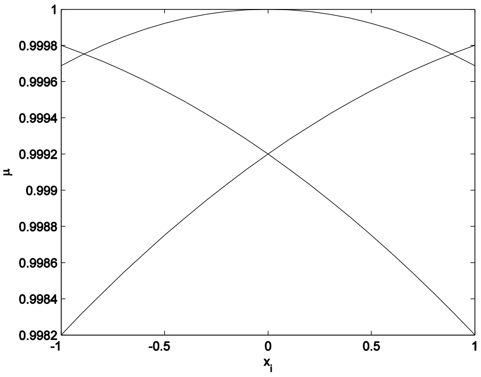

Figure 4: Membership functions of SRBCT

Table 1 shows the classification accuracy for the small round blue-cell tumor dataset using ELM (with 100 hidden neurons), other results from the literature and our proposed hybrid PCA-FUZZY-ELM.

Table 1. Classification of SRBCT dataset.

Methods	Average Accuracy over 100 runs in %	Standard deviation
ELM	80.72	0.3969
LiBSVM [36]	84.75	0.9400
J48 [36]	88.75	0.7900
SMO [36]	89.50	3.2200
Random Forest [36]	89.75	1.6600
Logistic Regression [36]	91.50	1.6600
IBk [36]	92.25	0.5000
Hybrid PCA-Fuzzy-ELM	92.77	0.2605

IV. PREDICTING QSARs FOR THE INHIBITION OF DIHYDROFOLATE REDUCTASE BY PYRIMIDINES

Learning quantitative structure activity relationships (QSARs) between pairs of compounds for the inhibition of Dihydrofolate Reductase by Pyrimidines [34, 35] is one of the challenging chemoinformatics dataset. There are three positions of possible substitutions for each drug and there are 9 attributes for each substitution position, namely, polar, size, flex, h-doner, h-acceptor, pi-doner, pi-acceptor, polarisable, sigma. There are 27 attributes in each drug. The total number of attributes in each instance is 54 (Since each instance have 2 drugs) and there are two classes. Five training and five testing dataset are available in [35, 36] and are used as it is. We consider first eight projected features provided by PCA. Three membership functions are used to represent each of the eight features (Figure 5 shows the membership function used for each of these eight features). Thus, 3^8 rules can be formed and hence there needs 3^8 hidden neurons (L).

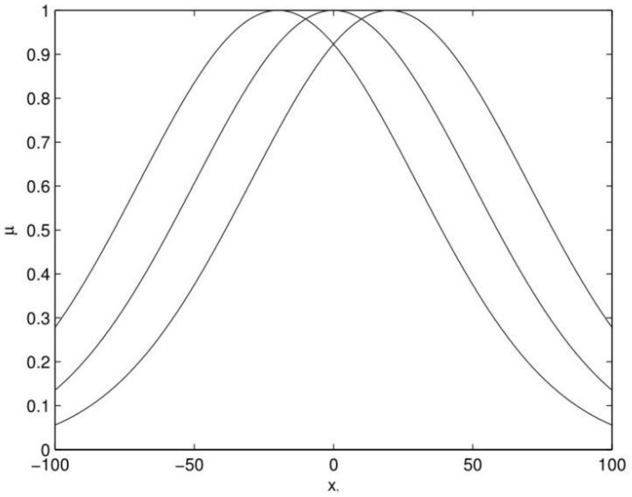

Figure 5: Membership functions for the Pyrimidine dataset

Table 2 shows the classification accuracy for the inhibition of dihydrofolate reductase by pyrimidines dataset using GOLEM [37], ELM (with 500 hidden neurons) and our proposed hybrid PCA-FUZZY-ELM. The results show that the proposed hybrid PCA-Fuzzy-ELM is better than ELM

Table 2. Classification of Pyrimidine dataset.

Methods	Average Accuracy over 100 runs in %	Standard deviation
GOLEM [37]	73.80	0.095
ELM	77.31	0.0548
Hybrid PCA-Fuzzy-ELM	80.09	0.0599

V. CONCLUSION

In this paper, a hybrid PCA-Fuzzy-ELM is proposed. It has high applications in the classification of high dimensional dataset. The results of classification/prediction on benchmarking dataset show the merits of the proposed hybrid PCA-Fuzzy-ELM. The results can be further improved by considering more projected features from PCA.

ACKNOWLEDGEMENT

The authors would like to thank CMS College of Arts and Commerce, Central University of Kerala and Bharathiar University for the research support.

REFERENCES

[1] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural networks 2(5), 359-366 (1989)

[2] Steinwart, I.: Consistency of support vector machines and other regularized kernal classifiers. IEEE Trans. On Information Theory 51(1), 128-142 (2005)

[3] Angelov, P.P., Zhou, X.: Evolving fuzzy-rule-based classifiers from data streams. IEEE Trans. on fuzzy systems 16(6), 1462-1475 (2008)

[4] Gou, J., Yi, Z., Du, L., Xiong, T.: A local mean-based k-nearest centroid neighbor classifier. The Computer Journal, 2012, doi: 10.1093/comjnl/bxr131

[5] Tsang, S., Kao, B., Yip, K.Y., Ho, W.S.: Decision Trees for Uncertain Data. IEEE Trans. on Knowledge and Data Engineering 23(1), 64-78 (2009)

[6] Friedman, N., Geiger, D.: Bayesian Network Classifiers. Machine Learning 19, 131-163 (1997)

[7] Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: Theory and applications. Neurocomputing 70, 489-501 (2006)

[8] Lendasse, A., He, Q., Miche, Y., Huang, G.B.: Advances in Extreme Learning Machines. Neurocomputing 149 (Part A), 158-159 (2015)

[9] Wang, D.G., Song, W.Y., Li, H.X.: Aproximation properties of ELM-fuzzy systems for smooth functions and their derivaties. Neurocomputing 149 (Part A), 265-274 (2015)

[10] Huang, G., Huang, G.B., Song, S., You, K.: Trends in extreme learning machines - A review. Neural Networks 61, 32-48 (2015)

[11] Lin, S., Liu, X., Fang, J., Xu, Z.: Is extreme learning machine feasible? A theoretical assessment (Part II). IEEE Trans. on Neural Networks and Learning Systems 26(1), 21-34 (2015)

[12] Peng, Y., Lu, B.L.: Discriminative manifold extreme learning machine and applications to image and EEG signal classification. Yong Peng and Bao-Liang Lu, To appear in: Neurocomputing, 2015

[13] Huang, S., Li, C.: Distributed Extreme Learning Machine for Nonlinear Learning over network. Entropy 17(2), 818-840 (2015)

[14] Zou, H., Lu, X., Jiang, H., Xie, L.: A Fast and Precise Indoor Localization Algorithm Based on an Online Sequential Extreme Learning Machine. Sensors 15(1), 1804-1824 (2015)

[15] Wang, J., Cai, L., Peng, J., Jia, Y.: A novel multiple instance learning method based on extreme learning machine. Computational Intelligence and Neuroscience 2015, Article ID 405890 (2015)

[16] Han, M., Liu, B.: Ensemble of extreme learning machine for remote sensing image classification. Neurocomputing 149 (Part A), 65-70 (2015)

[17] Mao, W., Zhao, S., Mu, X., Wang, H.: Multi-dimensional extreme learning machine. Neurocomputing 149 (Part A), 160-170 (2015)

[18] Zhou, Y., Liu, B., Xia, S., Liu, B.: Semi-supervised extreme learning machine with manifold and pairwise constraints regularization. Neurocomputing 149 (Part A), 180-186 (2015)

[19] Heeswijk, M.V., Miche, Y.: Binary/ternary extreme learning machines. Neurocomputing 149 (Part A), 187-197 (2015)

[20] Wang, B., Huang, S., Qiu, J., Liu, Y., Wang, G.: Parallel online sequential ELM based on MapReduce. Neurocomputing 149 (Part A), 224-232 (2015)

[21] Liu, X., Wang, L., Huang, G.B., Zhang, J., Yin, J.: Multiple kernel extreme learning machine. Neurocomputing 149 (Part A), 253-264 (2015)

[22] Wang, Z., Zhao, Y., Wang, G., Li, Y., Wang, X.: On extending extreme learning machine to non-redundant synergy pattern based graph classification. Neurocomputing 149 (Part A), 330-339 (2015)

[23] An, L., Yang, S., Bhanu, B.: Efficient smile detection by extreme learning machine. Neurocomputing 149 (Part A), 354-363 (2015)

[24] Sai, P.K., Wang, J.G., Teoh, E.K.: Facial age range estimation with extreme learning machine. Neurocomputing 149 (Part A), 364-372 (2015)

[25] Liang, N.Y., Saratchandran, P., Huang, G.B., Sundararajan, N.: Classification of Mental Tasks from EEG signals using ELM. Int. J. Neur. Syst. 16(1), 29 (2006)

[26] Wong, P.K., Wong, K.I., Vong, C.M., Cheung, C.S.: Modeling and optimization of biodiesel engine performance using kernel-based extreme learning machine and cuckoo search. Renewable energy 74(C), 640-647 (2015)

[27] Deo, R.C., Sahin, M.: Application of the extreme learning machine algorithm for the prediction of monthly effective drought index in eastern Australia. Atmospheric Research 153, 512-525 (2015)

[28] Yang, Y., Lin, X., Miao, Z., Yuan, X., Wang, Y.: Predictive control strategy based on extreme learning machine for path-tracking of autonomous mobile robot. Intelligent Automation and Soft Computing 21(1), 1-19 (2015)

[29] Wang, X., Han, M.: Improved extreme learning machine for multivariate time series online sequential prediction. Engineering Applications of Artificial Intelligence 40, 28-36 (2015)

[30] Wang, J., Wang, C.H., Chen, C.L.P.: The Bounded Capacity of Fuzzy Neural Netowrks (FNNs) via a New Fully Connected Neural Fuzzy Infernce System (F-CONFIS) with its Applications. IEEE Trans. on Fuzzy Systems 22(6), 1373 - 1386 (2013)

[31] Wang, J., Wang, C.H., Philip, C.L.: On the BP training algorithm of fuzzy neural networks (FNNs) via its equivalent fully connected neural networks (FFNNs). Proc. of IEEE International Conference on Systems, Man, and Cybernetics (SMC), 1376-1381 (2011)

[32] Jang, S.R.: ANFIS - Adaptive network based fuzzy inference system. IEEE Trans. on SMC 23(3), 665-685 (1993)

[33] Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 7(6):673-9 (2001)

[34] Ross, K.D., Steven, M., Richard, L., Michael, S.J.E.: Drug Design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. Proc. Natl. Acad. Sci. USA. 89, 11322-11326 (1992)

[35] Ross, K.D., Jonathan, H.D., Michael, S.J.E.: A comparison of artificial intelligence methods for modelling QSARs. Applied Artificial Intelligence 9(2), 213-233 (1995)

[36] Onur Dagliyan, Fadime Uney-Yuksektepe, I. Halil Kavakli1, Metin Turkay, Optimization Based Tumor Classification from Microarray Gene Expression Data, PLoS ONE 6(2): e14579.

[37] Jonathan D.H, Ross D. K., Michael J.E. S., Quantitative structure-activity relationships by neural networks and inductive logic programming. I The inhibition of dihydrofolate reductase by pyrimidines, Journal of Computer-Aided Molecular Design, 8 (1994) 405-420