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, Pertia – 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 weight initialization. The weights between the hidden 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. \( t_i = \{t_{i1}, t_{i2}, \cdots, t_{ik}\} \) represent the class label for the \( i^{th} \) 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 \( H^{\Psi} \) 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

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

\[
H = \begin{bmatrix}
g\left(\sum_{i=1}^{m} w_{il}x_{i1} + b_1\right) \\
g\left(\sum_{i=1}^{m} w_{il}x_{i2} + b_1\right) \\
\vdots \\
g\left(\sum_{i=1}^{m} w_{il}x_{im} + b_1\right)
\end{bmatrix}
\]

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.
### 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

\[
\mathbf{w}_{ij} = \log(\mathbf{A}_{ji}(\mathbf{x}_i))
\]

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

\[
r_1(l) = l \% R_1 \quad \text{and} \quad r_j(l) = \left( \prod_{k=1}^{l-1} R_k \right) l \% R_i \quad \text{for} \quad i = 2, \ldots, n
\]

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 \( s_j \) membership functions with \( P_{s_j} \) parameters is given by

\[
P \leq \left( \frac{m}{\sum_{j=1}^{n} s_j P_{s_j} + Lc} \right) c
\]

### III.1 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

1. Transform the 'm' attributes values to 'k' linearly uncorrelated attributes by keeping only the first 'k' principal component analysis (PCA), so that \( r_i = [\hat{a}_{i1}, \hat{a}_{i2}, \ldots, \hat{a}_{ik}] \), where \( \hat{a}_{ij} \) is the value of the '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 \( \sum_{i=1}^{k} s_i P_{ij} + M \prod_{j=1}^{k} s_j \) for every 'i', where 's' is the number of membership functions for the 'i'th variable and \( P_{ij} \) is the number of parameters of the 'j'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.

\[
H = \begin{bmatrix}
\text{sig} \left( \frac{\sum_{i=1}^{k} \log(A_{ij}(t)(x_{1}))}{\prod_{i=1}^{k} A_{ij}(t)(x_{1})} \right) & \ldots & \text{sig} \left( \frac{\sum_{i=1}^{k} \log(A_{ij}(t)(x_{1}))}{\prod_{i=1}^{k} A_{ij}(t)(x_{1})} \right) \\
\text{sig} \left( \frac{\sum_{i=1}^{k} \log(A_{ij}(t)(x_{2}))}{\prod_{i=1}^{k} A_{ij}(t)(x_{2})} \right) & \ldots & \text{sig} \left( \frac{\sum_{i=1}^{k} \log(A_{ij}(t)(x_{2}))}{\prod_{i=1}^{k} A_{ij}(t)(x_{2})} \right) \\
\vdots & \ddots & \vdots \\
\text{sig} \left( \frac{\sum_{i=1}^{k} \log(A_{ij}(t)(x_{n}))}{\prod_{i=1}^{k} A_{ij}(t)(x_{n})} \right) & \ldots & \text{sig} \left( \frac{\sum_{i=1}^{k} \log(A_{ij}(t)(x_{n}))}{\prod_{i=1}^{k} A_{ij}(t)(x_{n})} \right)
\end{bmatrix}
\]

4. Determine the output weights by considering MF-CONFIS as an extreme learning machine using \( \beta = H^\Psi T \), where \( H^\Psi \) 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).

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Average Accuracy over 100 runs in %</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELM</td>
<td>80.72</td>
<td>0.3969</td>
</tr>
<tr>
<td>LibSVM [36]</td>
<td>84.75</td>
<td>0.9400</td>
</tr>
<tr>
<td>J48 [36]</td>
<td>88.75</td>
<td>0.7900</td>
</tr>
<tr>
<td>SMO [36]</td>
<td>89.50</td>
<td>3.2200</td>
</tr>
<tr>
<td>Random Forest [36]</td>
<td>89.75</td>
<td>1.6600</td>
</tr>
<tr>
<td>Logistic Regression [36]</td>
<td>91.50</td>
<td>1.6600</td>
</tr>
<tr>
<td>IBk [36]</td>
<td>92.25</td>
<td>0.5000</td>
</tr>
<tr>
<td>Hybrid PCA-Fuzzy-ELM</td>
<td>92.77</td>
<td>0.2605</td>
</tr>
</tbody>
</table>
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](image)

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>
<thead>
<tr>
<th>Methods</th>
<th>Average Accuracy over 100 runs in %</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLEM [37]</td>
<td>73.80</td>
<td>0.095</td>
</tr>
<tr>
<td>ELM</td>
<td>77.31</td>
<td>0.0548</td>
</tr>
<tr>
<td>Hybrid PCA-Fuzzy-ELM</td>
<td>80.09</td>
<td>0.0599</td>
</tr>
</tbody>
</table>

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


