Contribution ID: **45** Type: **not specified**

Minimal Complexity Extreme Learning Machines

Learning sparse representations and minimizing model complexity have gained much interest recently. Parsimonious models are expected to generalize well, are easier to implement, and lead to smaller test times. The recently proposed Minimal Complexity Machine (MCM) showed that for training data $X = \{(x_i, y_i)|x_i \in \mathbb{R}^n, y_i \in \mathbb{R}, i = 1, 2, ...M\}$, minimizing h^2, where
\begin{equation}
h = \frac{\max_{i = 1, 2, ..., M} \|u^T x^i + v\|}{\min_{i = 1, 2, ..., M} \|u^T x^i + v\|}.
\end{equation}
leads to a hyperplane classifier $u^T x + v = 0$ with a small VC dimension. This task was shown to be equivalent to

\begin{equation}
\min_{w, b, h} h + C \cdot \sum_{i = 1}^M q_i
\end{equation}
\begin{equation}
h \geq y_i \cdot [{w^T x^i + b}] + q_i, ~i = 1, 2, ..., M
\end{equation}
\begin{equation}
y_i \cdot [{w^T x^i + b}] + q_i \geq 1, ~i = 1, 2, ..., M
\end{equation}
\begin{equation}
q_i \geq 0, ~i = 1, 2, ..., M.
\end{equation}

Models such as the Extreme Learning Machine (ELM) and Random Vector Functional Link Network (RVFLN) have been adapted to a number of applications and offer several advantages. Typically, the ELM solves
\begin{equation}
\min_{\beta, \xi} \; \frac{1}{2} ||\beta||^2 + \frac{1}{2} C \sum_i^M \xi_i^2
\end{equation}
\begin{equation}
h(x_i) \beta = y_i - \xi_i , \; i=1, 2, ..., M
\end{equation}

The last layer of the ELM network conventionally involves the computation of a pseudo-inverse; the hidden layer output matrix H is computed as a solution to $H\beta = Y$, where $H(w_1, w_2, ..., w_{\hat{n}}, b_1, b_2, ..., b_{\hat{n}}, x_1, x_2, ..., x_M) = g(w_i \cdot x_i + b)$, $\beta_i = [\beta_{i1}, \beta_{i2}, ..., \beta_{in}]^T$ is the weight vector connecting the i^{th} hidden node and output nodes, $w = [w_{i1}, w_{i2}, ..., w_{in}]^T$ is the weight vector connecting the i^{th} hidden node and input nodes, and Y is the vector of y_i's.

We propose combining the ELM with the MCM. This allows us to build classifiers or regressors with lower complexity in terms of VC dimension, which induce sparsity in the connections between the neurons of the final layer of the network. This has shown to not only improve generalization, but also create sparser networks which depict models closer to human cognition. Numerical stability issues associated with the calculation of the pseudo-inverse are also avoided.

Primary authors: Prof. DR, Jayadeva (Department of Electrical Engineering, Indian Institute of Technology, Delhi, India); Mr SOMAN, Sumit (PhD Candidate)

Presenter: Mr SOMAN, Sumit (PhD Candidate)

Track Classification: Student