Podado Pre-Entrenamiento de Máquinas de Aprendizaje Extremo

Nicolás Nieto[1], Guido Bracalenti[1], Iván Gareis[1,2], and Hugo L. Rufiner[1,2]

[1] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH−UNL/CONICET, Argentina
nnieto@sinc.unl.edu.ar,
[2] Laboratorio de Cibernética, Fac. de Ing., Univ. Nacional de Entre Ríos, Argentina

Resumen Las Máquinas de Aprendizaje Extremo son una eficiente herramienta de aprendizaje maquinal, con tiempos de entrenamiento reducidos y buena capacidad de generalización. Este tipo de redes suele poseer una gran cantidad de neuronas en su capa oculta, no siendo necesariamente todas de utilidad. En este trabajo se propone un nuevo método de podado pre-entrenamiento que utiliza información de las distribuciones de probabilidad asociadas a las activaciones de cada neurona. Para la evaluación del método propuesto se utiliza como ejemplo de aplicación la clasificación de señales de electroencefalografía, registradas durante tareas de habla imaginada. Se muestran resultados preliminares que evidencian la potencialidad del método propuesto.

Keywords: Máquinas de aprendizaje extremo, Podado de redes neuronales, Estadística de alto orden, Habla Imaginada

1. Introducción

Las Máquinas de Aprendizaje Extremo (ELM por sus siglas en inglés) son redes neuronales de una única capa oculta que se caracterizan por inicializar de forma aleatoria los pesos de la capa de entrada y por determinar de forma analítica las conexiones de la capa oculta con la capa de salida [1]. Esto les permite realizar un entrenamiento más rápido respecto a los perceptrones multicapa *tradicionales*, que poseen una estructura similar pero con un aprendizaje iterativo. Muchas veces para lograr un buen desempeño con las ELM se requiere un número relativamente elevado de neuronas aleatorias en la capa oculta. Sin embargo, no todas las neuronas resultan útiles e incrementan innecesariamente el tiempo de entrenamiento. Es por ello que, a fin de disminuir la cantidad de neuronas, se aplican métodos de podado que permitan eliminar las neuronas que resultan menos útiles. El principal desafío de estas técnicas es encontrar la manera de *medir* la utilidad de cada neurona. Este problema ha sido abordado anteriormente por distintos autores [2,3,4,5]. En este trabajo se propone caracterizar el comportamiento de las activaciones de las neuronas de la capa oculta mediante estadística de alto orden, que podría indicar cuales aportan información valiosa.

Se desarrolla un nuevo método de podado pre-entrenamiento que utiliza la estimación de la curtosis de las funciones de distribución de probabilidad (FDP) de dichas activaciones. Para evaluar el desempeño del método se utilizan datos reales de electroencefalografía (EEG), registradas en la tarea de habla imaginada [6]. La clasificación de señales de EEG ha sido uno de los problemas donde se han utilizado las ELM [7,8,9].

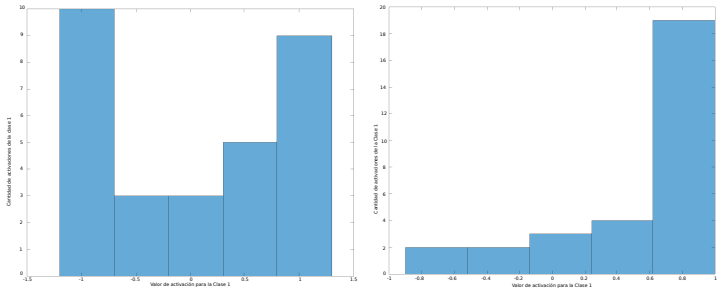
2. Máquinas de Aprendizaje Extremo

Las ELM son redes neuronales de aprendizaje supervisado, propuestas inicialmente por Huang et. al., que cuentan teóricamente con capacidad de aproximación universal [1]. Sean $(\boldsymbol{x}_i, t_i)$ los N ejemplos disponibles para un problema de clasificación, donde $\boldsymbol{x}_i \in \mathbb{R}^d$ es el i-ésimo patrón (d es la dimensión de los patrones de entrada) y $t_i \in \mathbb{R}$ su correspondiente etiqueta. Sea también $g_j : \mathbb{R}^d \times \mathbb{R}^L \times \mathbb{R} \to \mathbb{R}$ la función de activación de la j-ésima neurona oculta, donde L es la cantidad de neuronas de la capa oculta. Utilizando esta notación la salida de la j-ésima neurona oculta a la i-ésima entrada puede escribirse como $h_{i,j} = g_j(\boldsymbol{x}_i; \boldsymbol{w}_j, b_j)$, donde $\boldsymbol{w}_j \in \mathbb{R}^L$ y $b_j \in \mathbb{R}$ son los parámetros asociados a la neurona j, fijados aleatoriamente. En el caso particular de este trabajo todas las neuronas ocultas fueron aditivas con funciones de activación sigmoideas. Teniendo en cuenta estas consideraciones se puede escribir $h_{i,j} = S(\boldsymbol{x}_i \cdot \boldsymbol{w}_j + b_j)$, donde $S(x) = (1 - e^{-x})/(1 + e^{-x})$. Definiendo la matriz de activaciones ocultas como $\boldsymbol{H} = (h_{i,j}) \in \mathbb{R}^{N \times L}$, el vector de salidas de la red puede calcularse como $\boldsymbol{o} = \boldsymbol{H}\hat{\boldsymbol{\beta}}$, donde $\hat{\boldsymbol{\beta}}$ es el vector de pesos de la capa de salida. Sea $\boldsymbol{t} = [t_1, \cdot, t_N]^T$ el vector de salidas deseadas óptimo, el $\hat{\boldsymbol{\beta}}$ es el que minimiza la expresión $||\boldsymbol{o} - \boldsymbol{t}||$. Típicamente el conjunto óptimo de pesos de salida se obtiene en forma analítica como $\hat{\boldsymbol{\beta}} = \boldsymbol{H}^{\dagger}\boldsymbol{t}$, siendo $\boldsymbol{H}^{\dagger}$ la inversa generalizada de Moore-Penrose.

3. Materiales y Método Propuesto

La base de datos utilizada cuenta con señales de EEG de tres sujetos en la tarea de imaginar el movimiento de los labios para pronunciar los fonemas /a/, /u/ y un estado de *control* sin acción. Tras un filtrado espacial [10] se obtienen series temporales que maximizan la varianza entre cada uno de los pares de clases: /a/ *vs control*, /a/ *vs* /u/ y /u/ *vs control*.

A fin de poder caracterizar el comportamiento de las neuronas ocultas frente a los datos, en la Fig. 1 se presentan los histogramas de activación condicional por clase típicos. La Fig.1a corresponde a un comportamiento de tipo bimodal. En este caso las activaciones parecen no tener información relevante respecto de los datos de entrada, dando un número similar de activaciones positivas y negativas para la misma clase. En la Fig.1b se presenta un comportamiento de tipo unimodal, en el cual sí parece apreciarse una correlación entre las activaciones y los datos de entrada para una clase dada. El hecho de que estos histogramas sean los condicionales por clase, implica que una misma neurona podría tener



(a) Bimodal. (b) Unimodal.

Figura 1: Histogramas de activación condicional típicos de las neuronas aleatorias de la capa oculta para una clase determinada.

una FDP bimodal para ambas clases, bimodal para una clase y unimodal para la otra, o bien unimodal para ambas clases. La hipótesis consiste entonces en suponer que las neuronas que presenten un comportamiento bimodal pueden ser eliminadas. Con el fin de poder identificar este tipo de neuronas se estimó la curtosis de las FDP de las activaciones para cada clase. Este momento estadístico brinda información sobre la forma de la distribución. Un mayor valor implica una concentración de activaciones alrededor de la media (FDP más *picuda*). Sea $\hat{h}_j^{(\ell)}$ la media de las activaciones de la neurona j tomando como entrada los patrones correspondientes a la clase ℓ. La curtosis de la FDP asociada a las activaciones de la neurona j y la clase ℓ puede calcularse como,

$$K_j^{(\ell)} = \frac{\frac{1}{n_\ell}\sum_{i|t_i=c_\ell}(h_{i,j} - \hat{h}_j^{(\ell)})^4}{(\frac{1}{n_\ell}\sum_{i|t_i=c_\ell}(h_{i,j} - \hat{h}_j^{(\ell)})^2)^2}, \tag{1}$$

donde c_ℓ es la etiqueta y n_ℓ es el número de patrones de entrenamiento asociados a la clase ℓ. Utilizando la Ec. (1) podemos definir la figura de mérito utilizada para evaluar la neurona j como $M_j = K_j^{(c_1)} + K_j^{(c_2)}$. De esta forma se penaliza aquellas neuronas con un comportamiento del tipo bimodal, cuya curtosis es pequeña, frente a las de activación unimodal.

4. Resultados y Discusión

En la Fig. 2 se compara el ELM clásico (rojo) y con podado (azul) en la tarea de clasificación entre /a/ *vs control* para el sujeto 1 de [6]. Se utilizaron 60 ejemplos de entrenamiento y se dejaron 40 ejemplos para el conjunto de prueba. Los experimentos se corrieron con 100 inicializaciones de $\boldsymbol{w}_j$ y b_j distintas, reportándose media y desvío finales para diferentes cantidades de neuronas en la capa oculta. El algoritmo selecciona el 20 % de las mejores neuronas previo a la etapa de entrenamiento. Para analizar qué tipo de neuronas resultaron elimina-

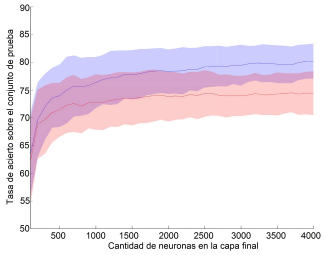
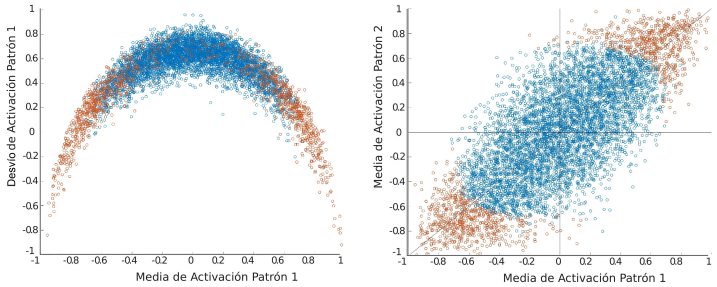


Figura 2: Tasa de acierto *vs* cantidad de neuronas finales en la capa oculta.



(a) Media *vs* Desvío. (b) Media c_1 *vs* Media c_2.

Figura 3: Neuronas eliminadas/podadas (azul), seleccionadas (naranja)

das se realizó un gráfico que muestra la relación entre el desvío y la media de la distribución para una clase (Fig.3a) y otro gráfico donde se muestran la media de las activaciones para cada clase (Fig. 3b). Se puede apreciar que el método propuesto elimina aquellas neuronas que poseen una media cercana a cero para ambas distribuciones, característica de las neuronas con una FDP bimodal.

5. Conclusiones y Trabajo Futuro

Se ha mostrado que con el método propuesto es posible identificar en forma exitosa las neuronas más útiles a partir de información estadística de la matriz de activaciones $\boldsymbol{H}$. Además, gracias al podado de las neuronas con FDP bimodal, el método mejora la capacidad de clasificación de las ELM. Resulta relevante destacar que a priori las neuronas con FDP condicionales unimodales similares para ambas clases no aportarían información discriminativa, sin embargo el algoritmo propuesto no las eliminaría. Si bien este problema escapa al alcance de este trabajo, ha sido contemplado y se proyecta abordarlo con métodos de doble podado. Por otra parte, se están evaluando otros estadísticos, como la asimetría, para agregar información adicional a la curtosis que pueda resultar de utilidad para el podado.

Referencias

1. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.
2. Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse, "Op-elm: optimally pruned extreme learning machine," *IEEE transactions on neural networks*, vol. 21, no. 1, pp. 158–162, 2010.
3. A. S. Alencar, A. R. R. Neto, and J. P. P. Gomes, "A new pruning method for extreme learning machines via genetic algorithms," *Applied Soft Computing*, vol. 44, pp. 101–107, 2016.
4. L. D. Tavares, R. R. Saldanha, D. A. Vieira, and A. C. Lisboa, "A comparative study of extreme learning machine pruning based on detection of linear independence," in *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*, pp. 63–69, IEEE, 2014.
5. M. Sánchez-Gutiérrez, E. M. Albornoz, H. L. Rufiner, and J. G. Close, "Post-training discriminative pruning for rbms," *Soft Computing*, pp. 1–15, 2017.
6. C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural networks*, vol. 22, no. 9, pp. 1334–1339, 2009.
7. S. Ding, N. Zhang, X. Xu, L. Guo, and J. Zhang, "Deep extreme learning machine and its application in eeg classification," *Mathematical Problems in Engineering*, vol. 2015, 2015.
8. L. Gao, W. Cheng, J. Zhang, and J. Wang, "Eeg classification for motor imagery and resting state in bci applications using multi-class adaboost extreme learning machine," *Review of Scientific Instruments*, vol. 87, no. 8, p. 085110, 2016.
9. B. Min, J. Kim, H.-j. Park, and B. Lee, "Vowel imagery decoding toward silent speech bci using extreme learning machine with electroencephalogram," *BioMed research international*, vol. 2016, 2016.
10. J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filters for single-trial eeg classification in a movement task," *Clinical neurophysiology*, vol. 110, no. 5, pp. 787–798, 1999.