Modelado de Sistemas No Estacionarios usando Perceptrones Lineales Adaptativos

Lucas C. Uzal, María Inés Széliga y Alejandro Ceccatto

Instituto de Física Rosario, CONICET/UNR,
Bv 27 de Febrero 210 Bis, 2000 Rosario, República Argentina

Resumen Presentamos un nuevo método de modelado de sistemas con comportamientos no estacionarios lentos, asociados a acoplamientos débiles de los mismos con medioambientes cambiantes o alteraciones en sus parámetros internos. El método propuesto se aplica, a modo de ejemplo, al caso más sencillo del modelado de sistemas no estacionarios lineales mediante perceptrones adaptativos. Los resultados obtenidos muestran que la estrategia utilizada, basada en la teoría de regularización de Tikhonov, es muy eficiente y superior al método usual de modelado local del sistema en ventanas temporales. La extensión a sistemas no lineales es directa, estando actualmente en desarrollo trabajos en esa dirección.

1. Introducción

El problema típico de regresión es encontrar la mejor aproximación a una ley $f : \Re^N \to \Re$, teniendo como datos un conjunto de pares (x^μ, y_ϵ^μ) $(\mu = 1, ..., P)$ para los cuales $y_\epsilon^\mu = f(x^\mu) + \epsilon^\mu$. En esta expresión x^μ es un vector de N componentes e y_ϵ^μ es un escalar que incluye un cierto error aleatorio ϵ^μ, usualmente supuesto de distribución normal $\mathcal{N}(0, \sigma_\epsilon)$. Consideraremos en lo que sigue que los pares (x^μ, y_ϵ^μ) corresponden a registros a diferentes tiempos discretos $t(\mu) = \mu \delta t$.

El problema de regresión arriba planteado corresponde, en términos del campo de la Inteligencia Artificial (IA) conocido como Aprendizaje de Máquina, al aprendizaje supervisado de una ley f a partir de datos o ejemplos[1]. En particular, dado que f no cambia en el tiempo, dicha ley describe el comportamiento intrínseco de un sistema estacionario. En el caso más general de sistemas no estacionarios, f tiene una cierta dependencia temporal f^μ, asociada a acoplamientos débiles del sistema con un medioambiente cambiante o alteraciones en los parámetros internos del mismo (correspondiente en IA al aprendizaje de un comportamiento que muta en el tiempo). Proponemos aquí un método para tratar este caso, suponiendo que f^μ tiene una variación suave en escalas temporales $\Delta t \gg \delta t$. Presentaremos el mismo ejemplificando su aplicación en el caso más sencillo, el modelado (aprendizaje) de sistemas no estacionarios lineales mediante perceptrones adaptativos. Es decir, supondremos que la ley dinámica[1]

[1] Note que por simplicidad obviamos incluir un término aditivo v_0^μ en esta ley lineal; dicho término no genera ninguna dificultad adicional y puede ser tratado en forma completamente análoga definiendo una variable auxiliar $x_0^\mu \equiv 1$.

$f^\mu(x^\mu) = v^\mu \cdot x^\mu/\sqrt{N}$, donde el vector maestro $v^\mu \in \Re^N$ tiene una evolución lenta, y trataremos de reconstruir esta ley utilizando como aprendices a perceptrones lineales cuyos pesos w^μ varían con el tiempo. Un problema similar para el caso de clasificación en lugar de regresión (correspondiente en IA al aprendizaje de conceptos[1]) puede hallarse en [2]. Para estudios sobre aprendizaje de perceptrones en ambientes estacionarios el lector puede referirse a [3]. No obstante, en ambos casos se usan métodos estadísticos diferentes a los considerados en este trabajo.

Es de destacar que si bien los resultados obtenidos corresponden estrictamente al caso sencillo de una ley f^μ lineal, los mismos son de interés por las siguientes razones: i) en muchas aplicaciones prácticas el comportamiento del sistema cerca de su punto de operación óptimo se modela en forma lineal (por ejemplo, para control del mismo), ii) si bien con una mayor dificultad técnica, el tratamiento de leyes dinámicas f^μ no lineales puede realizarse exactamente de la misma forma remplazando los perceptrones lineales adaptativos por redes neuronales artificiales[4], y iii) el estudio de regresión vía hiperplanos locales es un primer paso necesario en el desarrollo de métodos más potentes tales como máquinas regresoras de vectores soporte[5] adaptativas. Trabajos en las direcciones indicadas en ii) y iii) se hallan en progreso.

2. Descripción del Método

Como fuera planteado precedentemente, consideremos el problema de modelar (aprender) la ley dinámica lineal $f^\mu(x^\mu) = v^\mu \cdot x^\mu/\sqrt{N}$ de un sistema no estacionario a partir de registros (ejemplos) (x^μ, y^μ_ϵ), con $\mu = 1, \ldots P$, que verifican

$$y^\mu_\epsilon = f(x^\mu) + \epsilon^\mu.$$

Para ello utilizaremos perceptrones lineales cuyos pesos w^μ_i ($i = 1, \ldots, N$) dependen del tiempo μ:

$$\hat{y}^\mu = \frac{1}{\sqrt{N}} w^\mu \cdot x^\mu.$$

Proponemos determinar dichos pesos minimizando la función error

$$\begin{aligned} E(\boldsymbol{w}) &= \frac{1}{2P} \sum_\mu \frac{1}{|V|} \sum_{\eta \in V} (y^{\mu+\eta} + \epsilon^{\mu+\eta} - \frac{1}{\sqrt{N}} w^\mu \cdot x^{\mu+\eta})^2 \\ &\quad + \frac{\gamma}{2PN} \sum_\mu |w^{\mu+1} - 2w^\mu + w^{\mu-1}|^2 + \frac{\lambda}{2PN} \sum_\mu |w^\mu|^2 \\ &= E_{Aprendizaje} + \gamma E_{Suavizado} + \lambda E_{WD} \ , \end{aligned}$$

donde el conjunto V incluye al punto μ y sus vecinos temporales $\mu + \eta$ ($\eta = 0, \pm 1, \ldots, \pm n$) y $|V| = 2n + 1$ denota la cantidad de puntos en V.

El primer término, $E_{Aprendizaje}$, es el error cuadrático medio sobre el conjunto de aprendizaje. El término $E_{Suavizado}$ penaliza, en el espíritu de la regularización de Tikhonov[6], valores grandes de la derivada segunda de f^μ respecto al

tiempo –acorde con la hipótesis de suavidad de dicha ley en escalas temporales $\delta t \ll \Delta t$. Note que el mismo efecto se podría obtener con un error de suavizado basado en la derivada primera de la función,

$$E_{Suavizado} = \frac{1}{2PN} \sum_{\mu} |w^{\mu+1} - w^{\mu}|^2.$$

Ambos casos serán considerados más abajo. Finalmente, el error E_{WD} ("Weight Decay") es usado en la teoría de redes neuronales artificiales[4] para moderar posibles sobreajustes de los datos. Se incluye aquí por completitud, pero su efecto no será considerado en este trabajo. Los coeficientes γ y λ determinan la importancia relativa que se les asigna a los errores $E_{Suavizado}$ y E_{WD} frente a $E_{Aprendizaje}$.

La función error puede expresarse matricialmente en la forma

$$E(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}A\boldsymbol{w} - \boldsymbol{h}\cdot\boldsymbol{w} + B,$$

donde $A = T + X$, y

$$X_{ij}^{\mu\nu} = \frac{1}{NP|V|}\left(\sum_{\eta\in V} x_i^{\mu+\eta} x_j^{\mu+\eta}\right)\delta^{\mu\nu}$$

$$h_i^{\mu} = \frac{1}{P\sqrt{N}|V|}\sum_{\eta\in V}(y^{\mu+\eta} + \epsilon^{\mu+\eta})x_i^{\mu+\eta}$$

$$B = \frac{1}{2P|V|}\sum_{\mu}\sum_{\eta\in V}(y^{\mu+\eta} + \epsilon^{\mu+\eta})^2.$$

La matriz T a su vez viene dada alternativamente por

$$T1_{ij}^{\mu\nu} = \frac{\lambda}{NP}\delta_{ij}^{\mu\nu} + \frac{\gamma}{NP}(2\delta_{ij}^{\mu\nu} - \delta_{ij}^{\mu+1,\nu} - \delta_{ij}^{\mu-1,\nu})$$

$$T2_{ij}^{\mu\nu} = \frac{\lambda}{NP}\delta_{ij}^{\mu\nu} + \frac{\gamma}{NP}(6\delta_{ij}^{\mu\nu} + \delta_{ij}^{\mu-1,\nu+1} + \delta_{ij}^{\mu+1,\nu-1} - 4\delta_{ij}^{\mu+1,\nu} - 4\delta_{ij}^{\mu-1,\nu}),$$

dependiendo de que se considere la derivada primera ($T1$) o la segunda ($T2$) para suavizar la depependencia temporal de f^{μ}. De acuerdo a lo expresado más arriba, en lo sucesivo tomaremos $\lambda = 0$.

El error $E(\boldsymbol{w})$ es cuadrático en $\boldsymbol{w}$ y tiene un mínimo para $\boldsymbol{w}_0$ tal que $A\boldsymbol{w}_0 - \boldsymbol{h} = 0$. Note que esta ecuación vectorial constituye un sistema de NP ecuaciones lineales con igual número de incógnitas $w_{0,i}^{\mu}$ ($\mu = 1, P\ ; i = 1, N$). Sin embargo, si consideramos sólo el error de aprendizaje ($\gamma = 0$) y pretendemos un ajuste perfecto de los ejemplos ($E_{Aprendizaje} = 0$), obtenemos P subsistemas independientes de $|V| = 2n+1$ ecuaciones lineales y N incógnitas. De esta forma, según sea $N > |V|$, $N = |V|$ ó $N < |V|$ tendremos, respectivamente, un conjunto de sistemas compatibles indeterminados, determinados o sobredeterminados

para obtener $\boldsymbol{w}_0$. Es decir, para $N \geq |V|$ y $\gamma \approx 0$ es esperable un sobreajuste total de los datos. Por el contrario, cuando $N < |V|$ el método es capaz de filtrar el ruido aún para valores pequeños de γ.[2] Finalmente, en el límite opuesto $\gamma \gg 1$, si se utiliza la matriz $T = T1$ el método tiende a una regresion lineal multivariada convencional (estacionaria), mientras que con $T = T2$ tiende a una regresión multivariada cuyos pesos tienen la posibilidad de variar linealmente con el índice temporal μ.

De acuerdo a la discusión previa, variando γ y $|V|$ el método puede interpolar de manera contínua entre un ajuste total (sobreajuste) de los ejemplos de entrenamiento en ventanas independientes para $\gamma = 0$ y una simple regresión multivariada (convencional o permitiendo un cambio lineal con el tiempo en los parámetros de la regresión) para $\gamma \gg 1$. Es de esperar que para valores intermedios de estos hiperparámetros se obtengan resultados que permitan reconstruir adecuadamente la variación temporal del vector maestro v^μ. En la siguiente sección discutiremos estas cuestiones analizando los resultados sobre un ejemplo simple.

3. Experimentos

En los experimentos utilizamos como perceptrón maestro a

$$v_i^\mu = 0{,}5 + \sin\left(\frac{2\pi}{P}\mu + \phi_i\right),$$

donde ϕ_i $(i = 1, N)$ es una fase dependiente de la componente del vector. Se generaron $P = 100$ pares de ejemplos para aprendizaje, tomando los vectores de entrada x^μ con distribución $\mathcal{N}(0, \sigma_x = 1)$ e $y_\epsilon^\mu = (1/\sqrt{N})v^\mu \cdot x^\mu + \epsilon^\mu$. Si bien se experimentó con diferentes relaciones ruido-señal $\sigma_\epsilon^2/\sigma_y^2$, debido a que no se observaron cambios cualitativos en los comportamientos en lo que sigue se presentarán resultados para un valor fijo igual a 0.04 de dicha relación. De igual manera, se generó un conjunto con 100 datos para test del modelo aprendido. Con respecto a la dimensionalidad del espacio de entrada, se consideraron los casos $N = 1, \ldots, 5$ y cada perceptrón local se ajustó con $|V| = 1, 3, 5$ puntos. En lo sucesivo se muestran sólo ejemplos típicos de los resultados obtenidos, que resumen el comportamiento del método para los distintos valores de parámetros investigados.

En la figura 1 se grafican los diferentes errores en función de γ para $N = 5$ y $|V| = 1$, con $E_{Suavizado}$ con derivada primera (T_1). La figura 2, en cambio, es con $E_{Suavizado}$ con derivada segunda (T_2). Como puede apreciarse, los resultados son muy similares en ambos casos, aunque el mínimo del error de generalización E_{Test} es levemente menor a la vez que "más ancho" en el segundo caso.

[2] Note que para γ estrictamente nulo el resultado corresponde a un modelado usando ventanas deslizables de ancho $2n + 1$ datos. En consecuencia, la comparación de este valor con los resultados obtenidos para $\gamma > 0$ indica la ganancia del método propuesto frente al modelado estándar en ventanas.

Como era esperable, del análisis de las figuras surge que cuando γ es muy pequeño el método prioriza la minimización de $E_{Aprendizaje}$. En este caso, como $N = 5$ y $|V| = 1$ existen suficientes parámetros para hacer este error nulo. De hecho, existe un subespacio de soluciones $\boldsymbol{w}_0$ para $E_{Aprendizaje} = 0$; de este subespacio el método elige el vector que minimiza $E_{Suavizado}$, dado que γ no es estrictamente nulo. Se encuentra entonces que, para $0 < \gamma \ll 1$, $E_{Aprendizaje} \to 0$ pero ésta no es la situación óptima desde el punto de vista de la generalización, ya que el sistema está ajustando los errores ϵ^μ en lugar de la dinámica intrínseca f^μ.

En el otro límite, $\gamma \gg 1$, se encuentra que tanto $E_{Aprendizaje}$ como E_{Test} crecen, dado que la solución satisface $E_{Suavizado} = 0$ (corresponde a una regresión multivariada convencional). Para un rango de valores intermedios de γ se encuentra una situación óptima donde E_{Test} alcanza un mínimo. En esta situación $\boldsymbol{w}_0$ ya no pertenece al subespacio para el cual $E_{Aprendizaje} = 0$, sino que se ha apartado de dicho valor para disminuir en parte $E_{Suavizado}$.

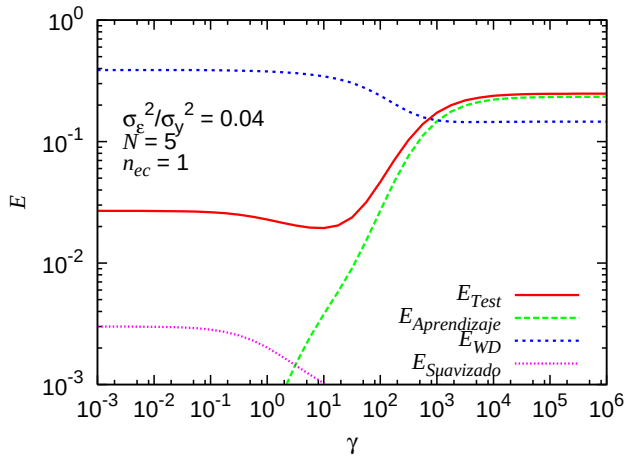

Figura 1. Diferentes errores como función de γ para $|V| \equiv n_{ec} = 1$ y $T = T1$.

En la figura 3 se muestra una de las componentes de w_0^μ comparada con el valor de v^μ con el que se generaron los datos, para el caso de $N = 2$ y tres valores destintos de γ: el valor óptimo y dos valores límites. Para $0 < \gamma \ll 1$ se observa que w_0^μ presenta fuertes fluctuaciones. Resulta evidente que en este caso se está produciendo un ajuste de los errores ϵ^μ. Con $\gamma = 10$ (valor óptimo para este caso) comienza a pesar el término de suavizado, produciendo un buen ajuste de la curva en función de μ. Finalmente, en el límite $\gamma \gg 1$, w_0^μ se vuelve constante para satisfacer $E_{Suavizado} = 0$. Dentro de los posibles valores constantes a tomar, tiende a aquel que minimiza $E_{Aprendizaje}$, es decir, el valor medio de $v^\mu (= 0{,}5)$.

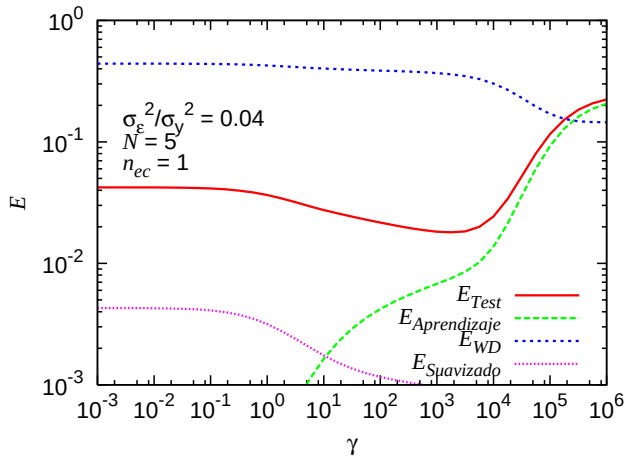

Figura 2. Idem Fig. 1 para $T = T2$.

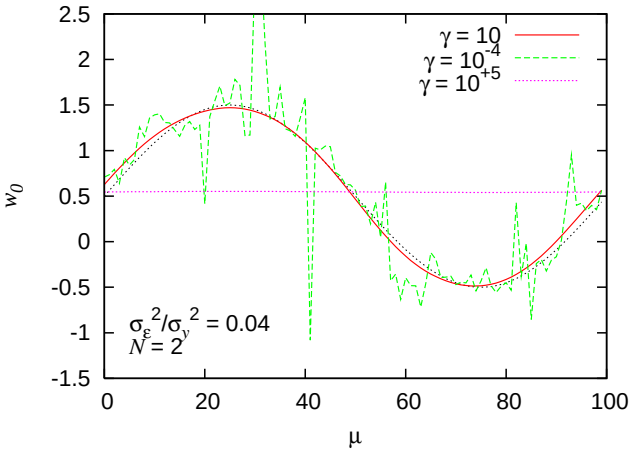

Figura 3. Reconstrucción de la variación temporal del vector maestro v^μ para distintos valores de γ.

En las figuras 4 y 5 se grafica el error NMSE ("Normalized Mean Squared Error") de generalización ($= 2E_{Test}/\sigma_y^2$) para $N = 3$. Se consideraron los casos con $|V| = 1, 3, 5$. En la figura 4 se muestran resultados correspondientes a un perceptrón maestro que varía con frecuencia $\omega = 2\pi/P$, tal como en los estudios anteriores. En la figura 5, en cambio, la frecuencia utilizada es 4 veces mayor. El $E_{Suavizado}$ está basado en la derivada primera.

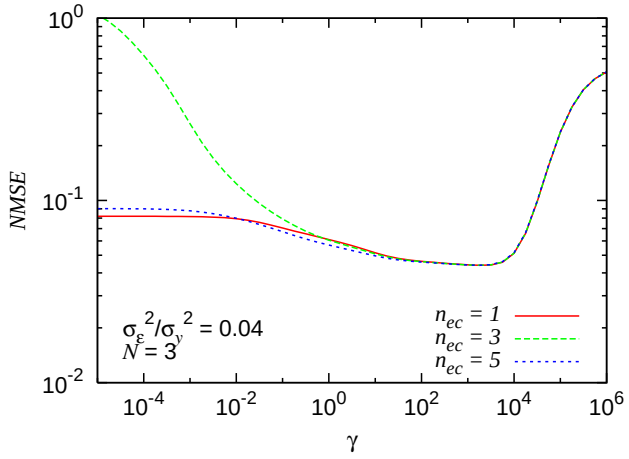

Figura 4. NMSE en test como función de γ, para diferentes valores de $|V| \equiv n_{ec}$. La variación temporal de v^μ ocurre con frecuencia $2\pi/P$.

En ambas figuras se observa el mismo comportamiento para γ muy pequeño. Cuando $|V| = 1 < N = 3$ se tiene un subespacio de soluciones de $E_{Aprendizaje} = 0$ con dos grado de libertad para minimizar $E_{Suavizado}$. En el caso $|V| = 5 > N = 3$ no es posible hacer $E_{Aprendizaje} = 0$ y se obtiene un valor no nulo de este error. Dado que $\gamma E_{Suavizado}$ resulta despreciable frente a $E_{Aprendizaje}$ no es de esperar ningun efecto de suavizado (más allá de la reducción de ruido producida al considerar los puntos vecinos para ajustar los perceptrones). En esta situación en que $|V| > N$, el límite $\gamma \to 0$ corresponde a un modelo de ajuste en ventanas locales sin un suavizado global. Puede verse que en ningún caso se obtiene en este límite un error NMSE de test que sea menor que el que se obtiene con un γ óptimo (mínimo de NMSE), lo cual muestra la importancia de la estrategia adoptada en nuestro método. En particular, note que el error de test para el valor ótimo de γ es siempre menor al 50 % del error para $\gamma = 0$, correspondiente al modelado estándar. Finalmente, en el caso crítico $|V| = 3 = N$, para γ pequeño el error crece a valores grandes, ya que el número de parámetros permite hacer $E_{Aprendizaje} = 0$ pero no queda libertad para suavizar el modelo. En este caso es mucho más evidente la ganancia de nuestro método frente al modelado en ventanas temporales.

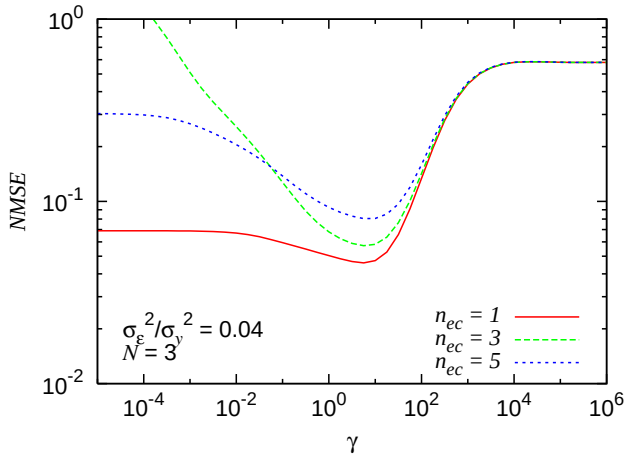

Figura 5. Idem Fig. 4 para un cambio temporal de v^μ más rápido, con frecuencia $8\pi/P$.

La diferencia entre las figuras 4 y 5 es la velocidad del cambio no estacionario de f^μ. Dado que la hipótesis básica es que dicha función varía lentamente con μ, como es de esperar, el error de test es levemente menor en la figura 4. En realidad, la diferencia observada es mucho menor aún si se suaviza utilizando la derivada segunda en lugar de la derivada primera. Otro aspecto a destacar es que en el primer caso el valor mínimo en E_{Test} es prácticamente el mismo independientemente de la cantidad de vecinos considerados. En el segundo caso, en cambio, el menor valor se obtiene para $|V| = 1$. Esta diferencia se debe a que, dado el rápido cambio de v^μ, los vecinos de un punto comienzan a ser generados por leyes dinámicas sensiblemente distintas, por lo cual no resulta conveniente utilizarlos para ajustar el perceptrón local (aún cuando ello ayude filtrando parcialmente el ruido). Note además que en la figura 5 el suavizado óptimo se obtiene para valores bastante menores de γ que los requeridos en la figura 4, lo cual está reflejando nuevamente el efecto del cambio más rápido de v^μ.

4. Conclusiones

En este trabajo, de carácter preliminar, presentamos un nuevo método de modelado de sistemas no estacionarios, correspondiente al problema de aprendizaje de leyes que mutan en el tiempo en IA. Estos comportamientos están usualmente asociados a acoplamientos débiles de los sistemas en estudio con medioambientes cambiantes, o a alteraciones en parámetros internos del mismo sistema. El método propuesto se ejemplificó en su aplicación al caso más sencillo, el modelado de sistemas no estacionarios lineales mediante perceptrones adaptativos.

Un análisis detallado de los resultados obtenidos muestra la eficiencia de nuestro método para recuperar los cambios en las leyes dinámicas perturbadas. Más aún, los mismos indican que la estrategia propuesta es superior al método usual de modelado local del sistema en ventanas temporales, obteniéndose errores de generalización en todos los casos menores al 50 % de los que genera este método.

Como fuera mencionado previamente, si bien los resultados obtenidos corresponden estrictamente al caso de una ley dinámica lineal, los mismos son de interés dado que en muchas aplicaciones prácticas el comportamiento del sistema cerca de su punto de operación óptimo se modela linealmente. Por otro lado, el tratamiento de leyes dinámicas no lineales puede realizarse exactamente de la misma forma, remplazando los perceptrones lineales adaptativos por redes neuronales artificiales u otros métodos modernos tales como máquinas regresoras de vectores soporte. Actualmente se hallan en ejecución extensiones del presente trabajo en estas direcciones.

Referencias

1. Tom Mitchell, "Machine Learning", McGraw-Hill, Boston (1997)
2. Renato Vicente, Osame Kinouchi, and Nestor Caticha, "Statistical Mechanics of Online Learning of Drifting Concepts: A Variational Approach", Machine Learning **32**, 179-201 (1998)
3. Peter Sollich, "Finite-size effects in learning and generalization in linear perceptrons", Journal of Physics A: Mathematical and General **27**, 7771-7784 (1994)
4. Christopher Bishop, "Neural Networks for Pattern Recognition", Oxford University Press, (1995)
5. N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines", Cambridge University Press, Cambridge (2000)
6. A. N. Tikhonov and V. A. Arsenin, "Solution of Ill-posed Problems", Winston & Sons, Washington (1977)