Ciencias Matemáticas, Vol. 34, No. 1, 2020, Pag. 123-127
Recibido 06-2020, Aceptado 07-2020

Predicción a corto plazo del comportamiento de la COVID-19 en Cuba: Un análisis desde la perspectiva del Aprendizaje Automático.
Short-term prediction of COVID-19 behavior in Cuba: An analysis from the perspective of Machine Learning.

Héctor González[1], Carlos Morell[2], Yanet Rodríguez[3]*

Resumen La COVID-19 es una enfermedad infecciosa que se propaga rápidamente por todo el mundo y que ha representado un reto importante para los sistemas sanitarios nacionales. Cuba también se ha visto afectada por esta pandemia y la capacidad de predecir el comportamiento a corto plazo de la cantidad de casos infectados acumulados en un día es una herramienta muy útil que necesitan las autoridades sanitarias nacionales. El presente manuscrito aborda el problema de la predicción a corto plazo de la variable de interés mediante el uso de técnicas clásicas del Aprendizaje Automatizado. Para ello se propone un método para el pre-procesamiento de los datos originales que permita la creación de un conjunto de aprendizaje válido. Posteriormente se selecciona y entrena un modelo predictivo basado en la regresión lineal con penalización que permite hacer predicciones acertadas y robustas de la variable bajo estudio.

Abstract COVID-19 is an infectious disease that spreads rapidly throughout the world and has represented a major challenge for national health systems. Cuba has also been affected by this pandemic and the ability to predict the short-term behavior of this disease is a very useful tool that the national health authorities need. This manuscript addresses the problem of short-term prediction of the variable of interest through the use of classical machine learning techniques. For this, a method is proposed for the pre-processing of the original data that allows the creation of a valid learning set. Subsequently, a predictive model based on linear regression with regularization is selected and trained to allow accurate and robust predictions of the variable under study.

[1] *Facultad 2, Universidad de las Ciencias Informáticas, La Habana, Cuba, hglez@uci.cu*
[2] *Centro de Investigaciones de la informática, Universidad Central "Marta Abreu" de Las Villas, Villa Clara, Cuba, cmorellp@uclv.edu.cu*
[3] *Departamento de Ciencias de la computación, Universidad Central Marta Abreu de Las Villas, Villa Clara, Cuba, yrsarabia@uclv.edu.cu*
***Autor para Correspondencia, Corresponding Author**

Introducción

El pronóstico de aparición de nuevos casos acumulados de infectados de COVID-19 cada día en el país es un elemento clave en la toma de decisiones de las autoridades gubernamentales y sanitarias. Este número depende de varios factores, la mayoría de los cuáles son muy difíciles de cuantificar de forma precisa. En tal sentido, estudios revelan que la aparición de nuevos casos está relacionada con la tasa de transmisión del virus, la duración de la enfermedad, la movilidad de las personas y las medidas higiénicas y de aislamiento social que ellas tomen [9, 16]. Excepto la duración de la enfermedad, el resto de los factores resulta imposible cuantificarlos con precisión. Es por ello que en este trabajo se intenta construir un modelo predictivo a partir de los datos disponibles de nuevos casos aparecidos cada día, desde el inicio de la epidemia. Este tipo de datos, coleccionado de forma periódica de una variable de interés, se conoce como serie temporal univariada. Tradicionalmente el pronóstico en series de tiempo se realiza utilizando herramientas estadísticas bien establecidas como por ejemplo los modelos auto-regresivos de Box-Jenkins o el enfoque de Holt-Winters al suavizado exponencial [3, 6]. Sin embargo, existe un marcado interés en la comunidad científica por abordar esta problemática utilizando técnicas de Aprendizaje Automático. Esta disciplina, dedicada al estudio de algoritmos capaces de aprender un comportamiento a partir de la supervisión que aportan los datos, ha tomado un auge impresionante en los últimos años y ha cosechado éxitos impactantes en casi todas las áreas de aplicación [2, 1, 5, 10, 8]. En el caso específico de las series temporales, no existe un consenso aun de cuándo sería mejor un enfoque u otro [7].

El presente trabajo tiene como objetivo la utilización de técnicas convencionales del Aprendizaje Automático para el pronóstico de nuevos casos acumulados de infectados en un día del COVID-19, a partir del comportamiento de los días anteriores. Entre las contribuciones del presente trabajo se propone:

- Realizar la transformación de la serie de datos univariada en un conjunto de muestras de aprendizaje. Los datos disponibles se enriquecen con la creación de nuevos atributos que capturen el posible comportamiento no lineal y la relación temporal de la variable de interés.
- Utilizar un algoritmo de Regresión Lineal con Penalización (LASSO) [12, 13, 14] que permite conformar un modelo predictivo para:
  - Obtener predicciones rápidas y con elevada precisión.
  - Seleccionar de manera automática las variables explicativas.
  - Evitar el sobre-ajuste del modelo para el conjunto de datos con que fue entrenado.

La validación de los resultados muestra que el modelo propuesto obtiene buenos resultados en la predicción a corto plazo de los nuevos casos acumulados infectados de COVID-19 en Cuba. También se han creado modelos similares para las provincias de Villa Clara y Pinar del Río con resultados satisfactorios.

Transformación de los datos

Los datos fueron colectados desde el servicio oficial COVID - 19 CUBA DATA disponible en [4] [1]. Se tomaron como datos de referencia para el estudio el comportamiento de la aparición de nuevos casos para el país (data-cu), las provincias Villa Clara (data-vc) y Pinar del Río (data-pr) por ser representativos de los diferentes comportamientos que ha tenido la propagación del virus en el territorio. El registro de los primeros casos confirmados en el país datan del 11 de marzo de 2020 con lo cual el muestreo de la data inicia en este rango de fecha. Finalmente, quedaron conformados 3 conjuntos de datos de series de tiempo univariadas para estudiar los modelos de pronósticos en la propagación de la pandemia. Para el estudio se empleo el valor acumulado de nuevos casos y los valores de la variable fueron normalizando a media cero y varianza uno.

Para la modelación del problema de estimación del comportamiento de nuevos casos positivo en Cuba se emplea un modelo que toma en cuenta, de manera local, la aparición de casos positivos en un intervalo de tiempo de días anteriores que denotaremos por t. De este modo la estimación de la aparición de nuevos casos y_i dependerá de como se ha comportado los días anteriores $y_{i-1}, y_{i-2}, \ldots, y_{i-|t|}$

[1] https://covid19cubadata.github.io/#cuba

Ademas, se utilizaron una familia de funciones con forma polinomial (En nuestro problema hasta de grado 5), exponenciales positivas y negativas y una función trigonométrica senosoidal que permite modelar el ruido. Finalmente, se usa un banco de $m = 8$ funciones las cuales al ser aplicadas sobre la ventana de datos nos permite obtener $p = m|t|$ descriptores.

Cada valor de la ventana aplicado a la familia de funciones genera un conjunto de datos para cada día de estimación representado de la forma:

$$x_i \in \mathbb{R}^p = \{f_j(y_{i-t})\} \quad \begin{matrix} j = 1, \ldots, m \\ t = 1, \ldots, |t| \end{matrix} \tag{1}$$

Luego de las primeras N observaciones del comportamiento de la aparición de nuevos casos del virus, se puede estimar la importancia w_i de cada función al ser aplicado sobre cada elemento de la ventana de tiempo.

Para la transformación de los datos usaremos un modelo Box-Cox con el banco de funciones que se especifican en la siguiente tabla las cuales fueron enunciadas con anterioridad:

Tabla 1. Funciones Box-Cox empleadas en el modelo.

$f_1(x) = x$	$f_2(x) = x^2$
$f_3(x) = x^3$	$f_4(x) = x^4$
$f_5(x) = x^5$	$f_6(x) = 1 - e^{-\alpha x}$
$f_7(x) = e^{\alpha x} - 1$	$f_8(x) = \sin(\beta x)$

Los valores de los parámetros α y β serán considerados en correspondencia con la escala de medición de los datos.

Modelo autoregresivo con penalización tipo LASSO

Variante de estimación para un día

Para expresar el modelo de estimación, emplearemos un enfoque en el que consideraremos todo el conjunto de funciones $\{f_j(y_{i-t})\}$ y sus respectivos pesos asociados w_{jt}. Luego, es posible mediante un algoritmo de aprendizaje obtener un predictor que tome la forma definida en la siguiente expresión:

$$\hat{y}_i = \sum_{t=1}^{|t|} \sum_{j=1}^{m} w_{jt} f_j(y_{i-t}); \quad i = |t|, \ldots, N \tag{2}$$

El modelo de aprendizaje consiste en aprender la combinación lineal, por medio de los pesos w_{jt}, de las funciones de base expresadas en la tabla 1. En principio se asume que todas las funciones de base están presentes y por tanto, la magnitud de estos pesos en relación con el resto definirá la prevalencia de estas funciones respecto a las demás. Un punto que distingue la propuesta, es el uso de un regularizador tipo LASSO que permite anular aquellos pesos irrelevantes del modelo y por ende, seleccionar el subconjunto de funciones cuya combinación lineal ponderada se ajuste a los datos reales.

Para resolver el problemas de aprendizaje se ha utilizado como función objetivo el error medio cuadrático combinado con la función de regularización, como se expresa en la ecuación 3.

$$w_{jt}{}^{*} = \underset{w_{jt}}{\operatorname{argmin}} \sum_{i=|t|}^{N} \left(y_i - \sum_{t=1}^{|t|} \sum_{j=1}^{m} w_{jt} f_j(y_{i-t}) \right)^2 + \lambda \sum_{t=1}^{|t|} \sum_{j=1}^{m} |w_{jt}| \tag{3}$$

En el modelo, el valor de λ controla el peso de la regularización y por tanto el sobre ajuste del mismo. Esta función de regularización discrimina aquellos atributos irrelevantes del problema, o lo que es equivalente, permite seleccionar del banco de funciones cox-box, a las que se corresponden al modelo univariado.

Variante de estimación para q días

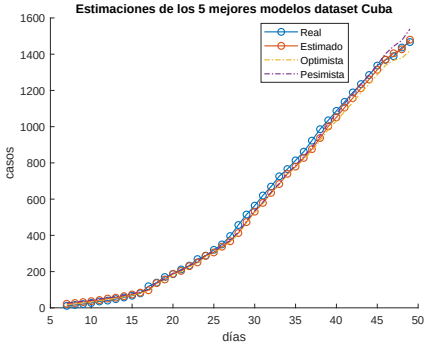
Para formalizar el problema de la estimación de múltiples días de manera secuencial, se emplea un esquema similar a los códigos de cadenas, en un escenario de predicción con salidas múltiples [11]. Para la estimación de los q días se propone un modelo de predicción que tome en cuenta las estimaciones de días anterior para obtener los nuevos valores de la serie, como queda expresado en el segundo termino de la expresión 4. Este enfoque no modifica el modelo de aprendizaje sin embargo el deslizamiento de la ventana hace que las nuevas estimaciones formen parte de las variables predictoras.

$$\hat{y}_{i+l} = \sum_{t=1}^{|t|-l} \sum_{j=1}^{m} w_{jt} f_j(y_{i-t}) + \sum_{t=1}^{l-1} \sum_{j=1}^{m} w_{tj} f_j(\hat{y}_t); \qquad \begin{matrix} i = |t|, \ldots, N \\ l = 1, \ldots, q \end{matrix} \tag{4}$$

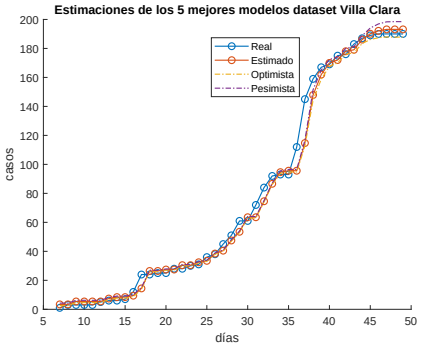
En este modelo el primer día se estima con las variables predictoras definidas, en tanto para el segundo día se desecha el último día de la ventana y se agrega la estimación del primer día. Este proceso se realiza de forma secuencial mientras se deseen estimar nuevos días. En la medida en que la ventana de variable de entrada se encuentre en el mismo orden de los días a estimar los resultados alcanzados se deben ajustar al modelo, en tanto el número de días aumente los resultados esperados se alejaran de la curva real.

Resultados y Discusión.

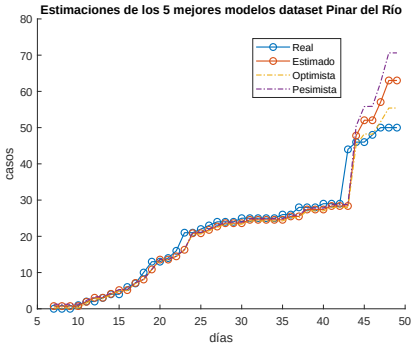
Se construyeron tres conjuntos de datos a partir de la información de nuevos casos de COVID-19 reportados en cuba durante los primeros 49 días de la enfermedad. Los últimos 7 días fueron separados para evaluar el modelo mientras los restantes días se emplearon para entrenar el modelo. El punto de partida para que los modelos de aprendizaje funcionaran adecuadamente fueron los primeros 30 días de la enfermedad, tiempo suficiente para que se minimicen el sesgo estadístico que se introduce cuando hay pocos casos para entrenar el modelo. Se repitieron 20 corridas, en una escala logarítmica, con un procedimiento de 5-folds CV interno para escoger el mejor valor del parámetro λ. Se empleo una ventana $|t| = 3$ de modo que el problema tiene dimensión $p = 24$.



(a) Comportamiento del conjunto de datos Cuba



(b) Comportamiento del conjunto de datos Villa Clara



(c) Comportamiento del conjunto de datos Pinar del Rio

Figura 1. Modelos Pronóstico sobre los conjuntos de datos estudiados para la variante de estimación del día siguiente.

Los resultados de las estimaciones y las desviaciones de los casos favorables y no favorables se muestran en la figura 1, para cada conjunto de datos que se ha estudiado. El modelo de pronóstico favorable y desfavorable se consideró en base a los resultados del valor medio y las desviaciones en las ejecuciones realizadas en el problema de aprendizaje. Como se aprecia en las curvas que esbozan el comportamiento real y estimado de nuevos casos acumulados existe muy poca

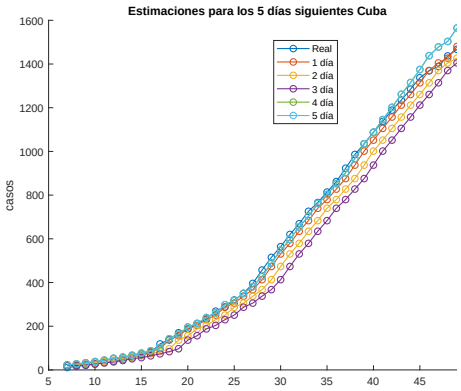
variabilidad en los resultados. De igual manera el cono cerrado de casos favorables y desfavorable da muestra de la estabilidad en las ejecuciones de los modelos de aprendizaje. Se debe notar que el uso de ventanas de tiempo permitió capturar cambios bruscos como los que se aprecian en la provincia Villa Clara entre los días 35 y 40 o en Pinar del Río entre los días 44 y 47. Para la mayoría de los valores de las series, el valor real de casos acumulados se encuentra en el rango de las curvas que hemos denominado optimista y pesimista, a excepción de los últimos días de muestreo del conjunto Pinar del Río donde ha ocurrido un cambio brusco de la serie en el final de la misma. Las estimaciones que se muestran de los primeros 30 días, se corresponden con las estimaciones de los propios datos de entrenamiento evaluados en el modelo, de ahí que en la tabla 2 se establezca una diferenciación en los resultados.

A continuación se plantean las expresiones obtenidas de cada modelo de estimación asociado a cada conjunto de datos. En los mismos prevalece las funciones exponenciales positivas o crecientes. De igual manera, se observan el conjunto de funciones seleccionadas luego de la resolución del problema de optimización. Nos llama la atención que en el proceso de aprendizaje prevalecen solo aquellas dependencia temporales con elementos del día anterior de la serie. Este hecho lo atribuimos a la forma de la serie temporal de los tres conjuntos estudiados, con cambios crecientes relativamente suaves, por lo que en próximos trabajos debemos seleccionar conjuntos de datos que presenten una naturaleza cambiante bruscamente y contrastar este hallazgo experimental. En tal sentido, si es importante resaltar que en el proceso de convergencia del algoritmo de optimización basado en el descenso por coordenada en bloques [12][15], las relaciones temporales son consideradas entre una iteración y otra en la interacción entre bloque de variables.

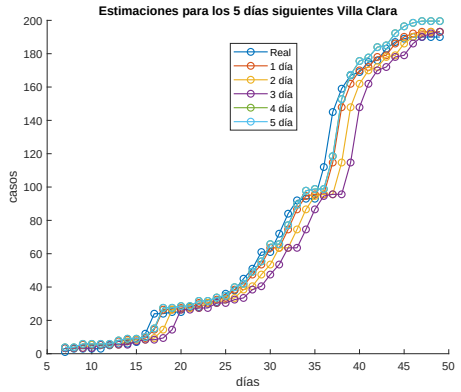
$$y_t = 525{,}55 + 479{,}07y_{t-1} + 5{,}87y_{t-1}^3 + 27644{,}18(e^{\alpha y_{t-1}} - 1)$$
$$y_t = 72{,}31 + 68{,}8y_{t-1} + 59{,}87(e^{\alpha y_{t-1}} - 1)$$
$$y_t = 19{,}19 + 0{,}62y_{t-1}^5 + 293{,}49(e^{\alpha y_{t-1}} - 1)$$

Por otra parte, los modelos de estimación para q días fueron ejecutado para 5 días. Los resultados obtenidos muestran muy poca variabilidad en los modelos de pronóstico respecto al comportamiento real, como se muestra en la figura 2. La ventaja fundamental de este enfoque es que obtiene, con un único modelo de aprendizaje, no solo el elemento de la serie siguiente sino los $q-1$ elementos sucesivos. El problema a controlar en este enfoque es el sesgo estadístico que se introduce en el proceso de propagación, donde para un valor de q muy grande las curvas de pronóstico se alejan del comportamiento real. En los resultados alcanzados sobre los conjuntos de datos estudiados, se evidencia un comportamiento estable para un valor de $q = 5$. Esta estabilidad en la estimación de los siguientes q días en los conjuntos de datos estudiados queda evidenciado en las gráficas de la figura 2.

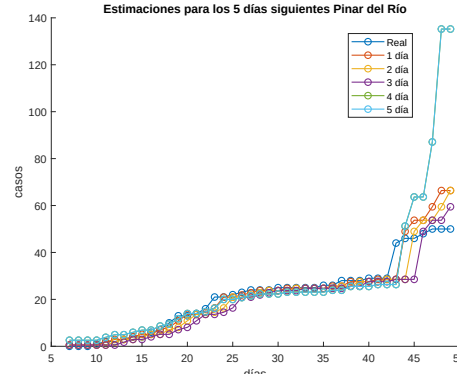
La tabla 2 resume las medidas de variabilidad de los errores cuadrático medio (RRMSE) así como el error relativo (RMSE) para los conjunto de datos estudiados. Los resultados reportados en la tabla se han dividido en resultados para entrenamiento, prueba y la serie completa en la estimación del siguiente día. La variabilidad máxima relativa en las curvas de pronóstico sobre los conjuntos de datos estudiados en el modelo de 1 día y el del 5 día no supera el $9{,}4\,\%$. Este resultado indica una estabilidad en el modelo de propagación de cadenas combinado con los mecanismos de aprendizaje empleados en este estudio.



(a) Comportamiento del conjunto de datos Cuba



(b) Comportamiento del conjunto de datos Villa Clara



(c) Comportamiento del conjunto de datos Pinar del Rio

Figura 2. Modelos Pronóstico sobre los conjuntos de datos estudiados para la variante de estimación de los 5 días siguientes.

Tabla 2. Medidas de error para los conjuntos de datos estudiados.

	RMSE			RRMSE		
Dataset	Train	Test	All	Train	Test	All
Cuba	26.13	17.68	24.95	0.51	0.27	5.37
Villa Clara	6.41	7.03	6.51	0.48	1.09	0.36
Pinar del Río	1.21	11.76	4.87	2.03	1.26	1.65

Conclusiones y Trabajo futuro

Los resultados alcanzados en la modelación del comportamiento de aparición de nuevos casos acumulados de infectados de COVID-19 en nuestro país, mediante la aplicación de métodos de aprendizaje automático, muestran resultados relevantes sobre los conjuntos de datos estudiados. El uso de mecanismos de penalización tipo LASSO favorecieron la selección automática de las funciones, cuya combinación lineal ponderada, ajustan el comportamiento de los valores reales. La introducción de mecanismos basados en códigos de cadenas combinados con los modelos de aprendizajes propuesto obtienen resultados estables en la estimación a corto plazo. Como continuidad de la presente investigación se considera la extensión del análisis en problemas univariados a considerar el problema multivariado combinando diferentes variables que estén relacionadas con la propagación del virus. De igual manera se sugiere contar con una mayor variedad de conjuntos de entrenamientos que permitan estudiar el problema de la relación temporal de la serie univariada y multivariada.

Referencias

[1] Benıtez-Pena, Sandra, Emilio Carrizosa, Vanesa Guerrero y Marıa Dolores: *Short-Term Predictions of the Evolution of COVID-19 in Andalusia. An Ensemble Method*. 2020.

[2] Chakraborty, Tanujit y Indrajit Ghosh: *Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis*. arXiv preprint arXiv:2004.09996, 2020.

[3] Collins, Sean: *Prediction techniques for Box–Cox regression models*. Journal of Business & Economic Statistics, 9(3):267–277, 1991.

[4] COVID-19, CUBA DATA: *CUBA DATA COVID-19*, 2020. `https://covid19cubadata.github.io/#cuba`, visitado el 2020-05-30.

[5] Flaxman, Seth, Swapnil Mishra, Axel Gandy, H Unwin, H Coupland, T Mellan, H Zhu, T Berah, J Eaton, P Perez Guzman *y cols.*: *Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries*. 2020.

[6] Granger, Clive WJ y Paul Newbold: *Forecasting transformed series*. Journal of the Royal Statistical Society: Series B (Methodological), 38(2):189–203, 1976.

[7] Makridakis, Spyros, Evangelos Spiliotis y Vassilios Assimakopoulos: *The M4 Competition: 100,000 time series and 61 forecasting methods*. International Journal of Forecasting, 36(1):54–74, 2020.

[8] Papacharalampous, Georgia, Hristos Tyralis y Demetris Koutsoyiannis: *Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece*. Water resources management, 32(15):5207–5239, 2018.

[9] Sanahuja, José Antonio: *COVID-19: riesgo, pandemia y crisis de gobernanza global*. Anuario CEIPAZ 2019-2020. Riesgos globales y multilateralismo: el impacto de la COVID-19, páginas 27–54, 2020.

[10] Siami-Namini, Sima y Akbar Siami Namin: *Forecasting economics and financial time series: ARIMA vs. LSTM*. arXiv preprint arXiv:1803.06386, 2018.

[11] Spyromitros-Xioufis, Eleftherios, Grigorios Tsoumakas, William Groves y Ioannis Vlahavas: *Multi-target regression via input space expansion: treating targets as inputs*. Machine Learning, 104(1):55–98, 2016.

[12] Tibshirani, Robert: *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

[13] Tibshirani, Robert: *The lasso method for variable selection in the Cox model*. Statistics in medicine, 16(4):385–395, 1997.

[14] Tibshirani, Robert: *Regression shrinkage and selection via the lasso: a retrospective*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3):273–282, 2011.

[15] Tseng, Paul: *Convergence of a block coordinate descent method for nondifferentiable minimization*. Journal of optimization theory and applications, 109(3):475–494, 2001.

[16] (WHO), World Health Organization *y cols.*: *Protocolo de investigación de los primeros casos y sus contactos directos (FFX) de la enfermedad por Coronavirus 2019 (COVID-19)*.