Impacto de la estandarización y escalado: factor para predicción de costos en proyectos a través de una red neuronal artificial

Impact of Standardization and Scaling: factor to predicting costs in projects through artificial neural network

Joselyn Rodríguez González[1]* Edgar Ugalde Saborio[2]

Recibido 15 de mayo de 2019, aceptado 13 de marzo de 2020
Received: May 15, 2019 Accepted: March 13, 2020

RESUMEN

Este artículo presenta una comparación de los métodos de estandarización y escalado en la predicción de costos. Se utilizaron cuatro métodos de estandarización y escalado para el procesamiento previo de datos; después de eso, los datos se procesaron a través de la red neuronal artificial (RNA). El primer paso fue crear variables comunes en proyectos de información basados en las opiniones de algunos gerentes de proyecto. El segundo paso fue simular un conjunto de datos basado en la información proporcionada por la empresa colaboradora: CRConsulting. La tercera etapa fue procesar los datos con algoritmos de aprendizaje automático de acuerdo con los cuatro métodos propuestos, los algoritmos de aprendizaje automático fueron los mismos en los cuatro casos. Por último, los resultados de la comparación se presentaron mediante modelos de ajuste según el método aplicado. El proceso anterior permitió determinar que los métodos de escalado y rango, aportan un mejor ajuste para la predicción de los costos, además de poseer un error medio cuadrático y un error cuadrático inferior en comparación con los datos no escalados y con los datos que fueron procesados por otros algoritmos de estandarización.

Palabras clave: Estandarización, escalado, predicción de costos, redes neuronales artificiales.

ABSTRACT

This paper presents comparison of Standardization and Scaling methods in cost predicting. Four methods were used for pre-processing dataset; after that, data was processed through artificial neural network (RNA). The first step was to build common variables in information projects based on opinions of some project managers. Second step was to simulate dataset based on information provided by CRConsulting. Third one was process data with machine learning according to the four methods proposed, RNA algorithms were the same in four cases. Last, the comparison results were presented through adjustment models according to the applied method.

Keywords: Standardization, scaling, cost predicting, artificial neural network.

[1] Universidad Latina de Costa Rica. Escuela de Ingeniería en Sistemas. San José, Costa Rica. E-mail: joselyn.rodriguez@ulatina.net
[2] Universidad Latina de Costa Rica. Escuela de Ingeniería en Sistemas. San José, Costa Rica. E-mail: edgar.ugalde@ulatina.net
* Autor de correspondencia: joselyn.rodriguez@ulatina.net

INTRODUCCIÓN

En la actualidad la literatura se enfoca en el análisis de la predicción de costos en proyectos, sin embargo, los modelos explicativos se orientan a exponer los resultados del procesamiento de los datos ejecutados por técnicas de aprendizaje automático, tales como las redes neuronales artificiales (RNA), sin profundizar en los detalles del manejo o depuración de la información.

Un factor que puede influir para mejorar el rendimiento en los resultados mostrados por una RNA en la predicción de costos en proyectos es la depuración del set de datos antes de ser procesados por la RNA.

Los modelos de aprendizaje estadístico son ampliamente utilizados para hacer regresión, clasificación y minería de datos. Se proponen cada vez más nuevos modelos de aprendizaje para tratar diferentes tipos de conjuntos de datos. La función de escala es un paso necesario para el preprocesamiento de datos y se usa ampliamente en aplicaciones [7].

La depuración se realiza a través de técnicas conocidas como normalización estadística y escalado, las cuales aportan valor a la predicción y a las técnicas de exploración de análisis, reduciendo ruidos y generando predicciones y modelos más certeros.

Los análisis que manejan datos masivos requieren mucho tiempo y esfuerzo para la preparación de datos [1] lo que incluye un preprocesamiento de estos. La minimización del impacto de diferencias, integración, transformación y reducción es parte del alcance principal de la preparación de set de datos, de forma que se pueda certificar la calidad antes de alimentar cualquier algoritmo de aprendizaje [2].

El presente estudio se enfoca en los métodos de transformación de datos, especialmente la normalización, y escalado de características, basado en variables para obtener costos de proyectos y en la utilización de algoritmos para la predicción de costo, con el objetivo de minimizar el impacto de los errores en la predicción.

El conocer los mecanismos adecuados para la reducción de errores permite el desarrollo de modelos con un desempeño apropiado para la predicción de costos en proyectos.

La estimación precisa de los costos de los proyectos de software hace que los proyectos se lleven a cabo en formato del tiempo y los costos identificados [13]. Sin una estimación adecuada de los costos para el desarrollo de los proyectos de software, el gerente de proyectos no podría diagnosticar cuánto tiempo o costo se necesita para el mismo y si ocurre algún error, el proyecto sería derrotado o enfrentaría riesgos [13]. La estimación precisa del costo y la confiabilidad, especialmente al comienzo del proyecto, son un factor importante para el éxito del proyecto [13].

No obstante, con frecuencia falla, debido a la complejidad en la gestión de la cartera de proyectos causada por muchos factores, como incertidumbre, interrelaciones entre proyectos, cambios en el tiempo y factores de éxito que son difíciles de medir [12].

Un aspecto que afecta en cualquier proyecto se refiere a las estimaciones de la variabilidad potencial en relación con las medidas de rendimiento, como el costo, la duración o la calidad vinculados con las actividades planificadas particulares. En el caso de proyectos, son afectados por la vaguedad, ambigüedad y contradicciones asociadas con la falta de claridad por datos ausentes, detalles incompletos e inexactos [14]. En consecuencia, se obtiene una esquematización y programación imprecisa por parte del administrador [3].

Cuando se trata de gestión de proyectos, la evaluación de costos se considera una de las tareas más desafiantes [4]. La estimación del costo es una actividad compleja que requiere el conocimiento de parámetros o variables sobre el proyecto para el cual se está construyendo el estimado [4]. Los variables son fluctuantes al no denotar los valores exactos y pueden diferenciarse o alejarse uno del otro fácilmente. Estos variables reflejan que las entradas para los modelos predictivos pueden verse afectadas por la variabilidad de estos.

Los modelos predictivos, como una regresión lineal, requieren un conjunto de entradas conocidas para predecir un resultado o valor de destino [5], sin embargo, en muchas aplicaciones del mundo real, los valores de las entradas son inciertos. La simulación permite explicar la incertidumbre de las entradas en modelos predictivos y evaluar la posibilidad de varios resultados del modelo en presencia de esa incertidumbre [5]. Por ejemplo, se tiene un modelo de beneficio que incluye el costo de los materiales

como una entrada, pero hay incertidumbre en ese costo por la volatilidad del mercado. Puede utilizar la simulación para modelar esa incertidumbre y determinar el efecto que tiene en los beneficios [5].

Por otra parte, muchos modelos estadísticos, de regresión y técnicas clásicas están disponibles para la estimación del costo del proyecto [6]. Los modelos clásicos, pese a su limitado alcance, tienen una gran demanda en el mercado. En los últimos años, varios modelos de aprendizaje automático han recibido atención de investigadores que trabajan en el área de la tarea de estimación de costos [6]. Entre estos enfoques, los modelos basados en redes neuronales han demostrado una habilidad distinguida en la predicción a partir de la experiencia [6], no obstante, las investigaciones centran su atención en los algoritmos de aprendizaje de la red neuronal artificial y no así en el preprocesamiento o transformación de los datos.

Los variables de los proyectos suelen caracterizarse por ser multidimensionales, estos tipos de datos con frecuencia tienen unidades diferentes [7]. El escalado y normalización son pasos para preprocesar la información que provenga de los variables del proyecto, en los casos que el algoritmo de procesamiento lo necesite. Por ejemplo, número de participantes de proyectos, prioridad del proyecto, complejidad o duración en meses. Todos estos datos difieren en unidad y en escalado, al respecto las medidas afectan el resultado final cuando difieren en muchas unidades unas de otras.

Por lo tanto, es necesario aplicar los métodos de escalado y normalización para procesar todos los datos, además las categorías de los datos pueden presentarse en numéricas, categóricas u ordinarias, por lo que se deben escalar en la misma ruta [7].

Los métodos de normalización que permiten la transformación de datos incluyen el min-max, las desviaciones y la escala, por mencionar algunos [2]. Los métodos para la normalización y escalado se describen en el siguiente apartado, al igual que el desarrollo matemático.

MÉTODO DE ESCALADO Y NORMALIZACIÓN

Los métodos que a continuación se presentan han sido reconocidos por afectar el resultado y desempeño en las RNA.

Escalado de características/rango

Los métodos de escalado de características se utilizan con objetivos tanto para aumentar el rendimiento de predicción del algoritmo empleado como para disminuir el coste computacional del algoritmo utilizado. En general, estos métodos se usan en clasificación, predicción, reconocimiento de patrones y procesamiento de señales [15].

La función para realizar el escalado de características es la siguiente, según la ecuación (1):

$$\overline{X_{ij}} = \frac{X_{ij} - X_{i_{\min}}}{\left(X_{i_{\max}} - X_{i_{\min}}\right)} \tag{1}$$

En una escala de rango, después de centrar cada columna, los valores se dividen por el rango mínimo y máximo del indicador. Como muchas tecnologías utilizadas en los estudios de escalado producen resultados de valor cero, y se reconoce que estos valores son límites inferiores naturales, solo los valores máximos parecen reflejarse como el divisor esencial [9]. Este método es sensible a valores atípicos donde el rango es muy grande[9, 10].

Normalización

La autoescala o normalización, también llamada escala de varianza de unidad o unidad, se aplica comúnmente y usa la desviación estándar como factor de escala [11, 12]. Después del autoescalado, todos los parámetros tienen una desviación estándar de uno y, por lo tanto, los datos se analizan sobre la base de correlaciones en lugar de covarianzas [11]. La ecuación (2) muestra la forma de calcular el método de normalización:

$$\overline{X}_{ij} = \frac{X_{ij} - \overline{X}_i}{S_i} \tag{2}$$

Escalado Pareto

Las familias de distribuciones de Pareto proveen ajustes razonables y permiten que los datos encajen para muchas distribuciones empíricas, por ejemplo, a distribuciones de ingresos y de valores de propiedad [8]. En la mayoría de estos casos, la información auxiliar presente podría utilizarse si se dispusiera de una distribución de Pareto multivariada adecuada [8]. La ecuación (3) se utiliza para calcular el método de escalado de Pareto.

$$\overline{X}_{ij} = \frac{X_{ij} - \overline{X}_i}{\sqrt{S_i}} \quad (3)$$

La escala de Pareto es similar a la normalización, pero después del centrado medio, cada columna se divide por la raíz cuadrada de la desviación estándar. No obstante, esto tiene la desventaja de que es muy sensible a los grandes cambios en los datos [16].

Escalado Vast

La estabilidad Vast (que también se conoce como gran escala) es un método bastante nuevo y sólido que podría considerarse como otra extensión de la autoescala [9]. La escala VAST puede mejorar el análisis de cualquier conjunto de datos multivariados donde las diferencias de clase se vean oscurecidas significativamente por la variación espuria [17]. La función para realizar el método de escalado de VAST es la siguiente, según la ecuación (4):

$$\overline{X}_{ij} = \frac{\left(X_{ij} - \overline{X}_i\right)}{S_i} \cdot \frac{\overline{X}_i}{S_i} \quad (4)$$

Esto se debe al hecho de que el valor final de la normalización se multiplica por un factor de escala dividido por la desviación estándar [9]. Por consiguiente, este método puede ser útil en la identificación de costos con pequeñas fluctuaciones [9].

DISEÑO

Este apartado detalla el proceso de diseño de la investigación, desde la adquisición del set datos principales, el pre-procesamiento de los datos para verificar cuál de los métodos expuestos anteriormente tienen buen rendimiento en el preprocesamiento y, por último, el procesamiento.

La Figura 1 describe el modelo de desarrollo del estudio y la secuencia de las actividades realizadas para la investigación, que incluye la definición de las variables para los proyectos. El objetivo de la investigación era corroborar si existe un mejor desempeño en la red neuronal cuando los datos son depurados por métodos de escalado y/normalización. Por lo tanto, para probar los métodos se solicitó a un gestor de proyectos su opinión de experto, además el mismo contribuyo con la información de 10 proyectos de la empresa para la que labora. El primer proceso que se llevó a cabo fue la elección de variables, lo cual consistió en identificar los atributos convenientes para la simulación de proyectos en general.

Con ayuda de un PMP (Project Manager Professional), se hizo un análisis de las encuestas para determinar cuáles eran los variables que se presentaban en común cuando los gestores debían realizar una estimación del costo de un proyecto, de manera que se eliminaron aquellos atributos que parecían redundantes. Cada destacar que estas variables son sugeridas por el experto considerando que el estudio lo que pretende es verificar si los métodos de escalado y normalización denotan un impacto en escalares que varían significativamente en su magnitud.

Por último, se establecieron 5 variables que son elementales para todos los proyectos, pero además se resguardaron ciertas características que pueden contribuir a la complejidad de los datos. Por ejemplo, unidades diferentes como dimensión del equipo de trabajo (las unidades eran sujetos) y duración planificada en meses (unidad meses), o unidades muy lejanas en su medida, como es el caso de complejidad que su ubica en una escala de 1 a 3 y costo planificado que no tiene escala definida y puede

Fuente: Propia.

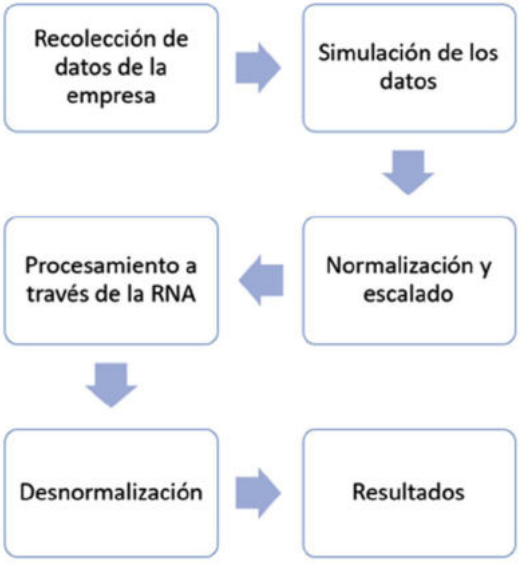

Figura 1. Modelo de desarrollo.

presentarse en unidades de millar. Las características tomadas en consideración fueron: Complejidad, duración planificada meses, duración real, tamaño del equipo, costo planificado y costo Real.

El segundo proceso realizado fue la simulación de datos. Para recrear los costos de los proyectos, se contó con la colaboración de la empresa CRConsulting, la cual se orienta a ofrecer servicios en tecnología. El proceso se llevó a cabo gracias a la guía del PMP, quien colaboró con la simulación de datos, basado en los proyectos que ha dirigido en la empresa y su trayectoria por más de 15 años en la gestión de proyectos. El gestor aportó la información de 10 proyectos de la empresa y la simulación se realizó basada en estos datos, esta simulación es relevante ya que permite que los siguientes datos tengan un comportamiento similar a los que se esperan como target en la red neuronal.

Para la estandarización, se aplicaron los métodos de escalado y normalización antes descritos. Los cuales se compararon en el rendimiento que cada uno ellos mostraban. Cada método fue aplicado al mismo set de datos y se verificó cual facilito que la red tuviera un mejor ajuste. Por último, cada set depurado se procesó por medio de una red neuronal artificial, la cual tenía objetivo poder procesar el costo final de los proyectos reales, a través del entrenamiento con los proyectos simulados.

MÉTODO

En este estudio se utilizó la información de 10 proyectos provenientes de CRConsulting y se simularon 190 proyectos informáticos, a través del método de Montecarlo, el cual permite generar valores aleatorios basados en la probabilidad de ocurrencia y a la vez simula el comportamiento similar de los proyectos. Los proyectos simulados se utilizaron para entrenar a la red neuronal y los 10 proyectos reales son el objetivo de la red. Los datos fueron almacenados en Excel ™, versión 2016.

Al efectuar la simulación, se utilizaron las funciones de distribución (media, desviación estándar, máximos y mínimos, distribución normal) y los datos se almacenaron en hojas de Excel™. Los variables para la simulación fueron: complejidad del proyecto, tiempo en meses estimado, tiempo en meses real, costo estimado y costo real. Estas variables suele impactar en los proyectos, sin embargo, no necesariamente deben ser las mismas. Las redes neuronales pueden contar con las variables que un usuario necesite. Este artículo verifica el rendimiento al utilizar métodos de normalización y escalado, cuando las variables para la red neuronal tienen escalares con magnitudes lejanas, por ejemplo, meses estimados (números que no sobrepasan miles, es un valor discreto, unidad de tiempo) y costo estimado (valor de gran magnitud, es un valor continuo, unidad moneda), estás diferencias escalares suelen impactar el rendimiento de una red neuronal. Las variables de esta investigación son los métodos de escalado, las variables de los proyectos propuestas en este estudio no han sido sometidas a pruebas y ni se propone que sean estas las únicas consideras en posteriores estudios.

Preprocesamiento de los datos

A los datos simulados se les aplicó un proceso de depuración, por lo cual se manipularon en las hojas de cálculo que se encontraban almacenadas, donde se empleó cada uno de los métodos de estandarización y escalado. Los métodos aplicados fueron escalado de rango, normalización, escalado Pareto y escalado de Vast. En los cálculos se utilizaron por aparte los mínimos, máximos y desviación estándar del set de datos. Los datos preprocesados se almacenaron en una hoja de Excel™.

Para el procesamiento se empleó el software Matlab™ y Neural Network ToolBox™. Los datos fueron procesados por un tipo de red neuronal artificial prealimentada, retropropagación. La función de aprendizaje fue gradiente descendiente. Las RNA diseñadas fueron de 2 capas y contaron con 1000 repeticiones. El vector de entrada está definido por los valores normalizados: complejidad del proyecto, tiempo en meses estimado, tiempo en meses real, costo estimado. El objetivo u objetivo de la red eran los costos reales de los 10 proyectos que se conocían con certeza.

En cuanto al diseño de la red, se alimentó con los 190 proyectos, estas son las entradas de entrenamiento que se utilizaron, y los 10 proyectos analizados por CRConsulting como objetivo. Del entrenamiento, se obtuvo el error medio cuadrático, error cuadrático, desviaciones medio y salidas de la red.

La red neuronal empleada en este estudio posibilita mapear entre un conjunto de datos de entradas

numéricas y un conjunto de objetivos numéricos. La aplicación de la red neuronal selecciona los datos, crea y entrena una red, y evalúa su rendimiento utilizando el error cuadrático medio y el análisis de regresión (Matlab).

Como se observa en la Figura 2, la red utilizada es alimentación de dos capas con neuronas sigmoideas ocultas y neuronas de salida lineal, que encaja en los problemas de mapeo multidimensionales, apoyados por datos consistentes y suficientes neuronas en su capa oculta. La red se capacita con el algoritmo de retropropagación de Levenberg-Marquardt (Matlab).

Posterior a las salidas del entrenamiento, se invirtieron las funciones de normalización y escalado para desnormalizar los datos de salida de la RNA, después de haber sido evaluadas por funciones estadísticas para cálculos del error en desempeño y pronóstico. Se evaluó el algoritmo de clasificación en combinación con cuatro escalas diferentes, así como el uso de no escalar. En consecuencia, se registraron 5 predicciones producto de la red neuronal (uno por cada método) en conjunto con su intervalo de confianza del 95 %, correspondiente para cada uno de los datos procesados con las distintas escalas.

RESULTADOS

Durante el entrenamiento de cada RNA aplicada a cada set de datos, se obtuvo una salida que predecía el costo de los proyectos asignados como objetivo. Los resultados del entrenamiento se compararon, ponderando los errores que los mismos presentaban. Para cada salida de la red neuronal artificial, se calcularon los criterios estadísticos: error medio cuadrático (MSE) y raíz cuadrada del error cuadrático medio (RMSE), con el fin de evaluar el desempeño del promedio de las salidas de la RNA.

La Tabla 1 indica los errores de acuerdo con la predicción en las redes neuronales. Notablemente la predicción presenta un mejor desempeño cuando se aplicaron los métodos de depuración de datos, que sin ellos. El rendimiento de la red se incrementa cuando se emplean los métodos de normalización y escalado, por ende, los errores sufren un descenso. Sin embargo, algunos de los métodos de normalización no sensibilizan suficiente los datos para alcanzar un error mínimo como es el caso particular del escalado de Pareto, el cual tiene un error cuadrático medio elevado. En el caso del análisis, el escalado de rango favorece el ajuste de los datos para la predicción, lo cual permite que el modelo que se ajuste tenga un desempeño óptimo.

Fuente: Matlab.

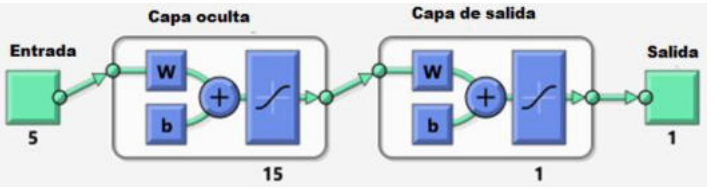

Figura 2. Modelo de la red neuronal utilizada.

Tabla 1. MSE Y RMSE según el método de normalización o escalado.

Método	MSE	RMSE
No normalizado	289420000000	537980
Escalado de rango	0.0016955	0.0399
Normalización	0.0532	0.2307
Escalado Pareto	594650	771.1336
Escalado Vast	2.9341	1.7129

Fuente: Propia.

En la Figura 3 se observa el desempeño según la red neuronal artificial, el método de escalado de rango presenta un ajuste mayor que los otros modelos utilizados. Por otra parte, el método de Pareto tiene un mayor MSE y RMSE, además se caracteriza por mantener los datos cercanos a la medida original.

La Figura 4 muestra la dispersión del costo total del proyecto y los modelos de ajuste de las regresiones realizadas por la RNA. Los gráficos exponen cómo la aplicación de métodos de normalización y escalado posibilitan que los datos se distribuyan mejor, encontrándose cada caso próximo al siguiente.

La Tabla 2 presenta las salidas obtenidas de la RNA luego de los entrenamientos. En la primera columna se aprecia el objetivo original y las siguientes son cada una de las salidas según los métodos utilizados para reducir la incertidumbre y el ruido.

La Figura 5 denota cómo los desajustes se manifiestan de acuerdo con las salidas de la RNA. El modelo que tiene un mayor desajuste es el de Vast; una de las características de este método es que no puede utilizarse cuando los datos poseen gran varianza. Particularmente, el set de datos para esta investigación que la desviación estándar de algunas

Tabla 2. Costo real versus salidas de la red según el método.

Costo real	Rango	Normalización	Pareto	Vast
1.E + 06	1.E + 06	1.E + 06	6.E + 05	9.E + 05
9.E + 05	8.E + 05	8.E + 05	8.E + 05	6.E + 05
1.E + 06	1.E + 06	1.E + 06	6.E + 05	2.E + 06
7.E + 05	7.E + 05	7.E + 05	1.E + 06	1.E + 06
1.E + 06	1.E + 06	1.E + 06	8.E + 05	2.E + 06
1.E + 06	1.E + 06	1.E + 06	8.E + 05	4.E + 05
1.E + 06	1.E + 06	1.E + 06	8.E + 05	7.E + 05
8.E + 05	7.E + 05	8.E + 05	9.E + 05	8.E + 05
8.E + 05	8.E + 05	8.E + 05	8.E + 05	1.E + 06
1.E + 06	1.E + 06	1.E + 06	6.E + 05	3.E + 06

Fuente: Propia.

de sus características era lo suficiente grande, de forma que el método de Vast colaboro para que la predicción de la red neuronal artificial se desviará del costo real.

En la Figura 6 se observa que el método de ajuste Pareto no contribuye a la predicción de los costos, por el contrario, tiende a desajustarse en gran medida. Cumple con ser muy sensible cuando los cambios son grandes en los datos.

Fuente: Propia.

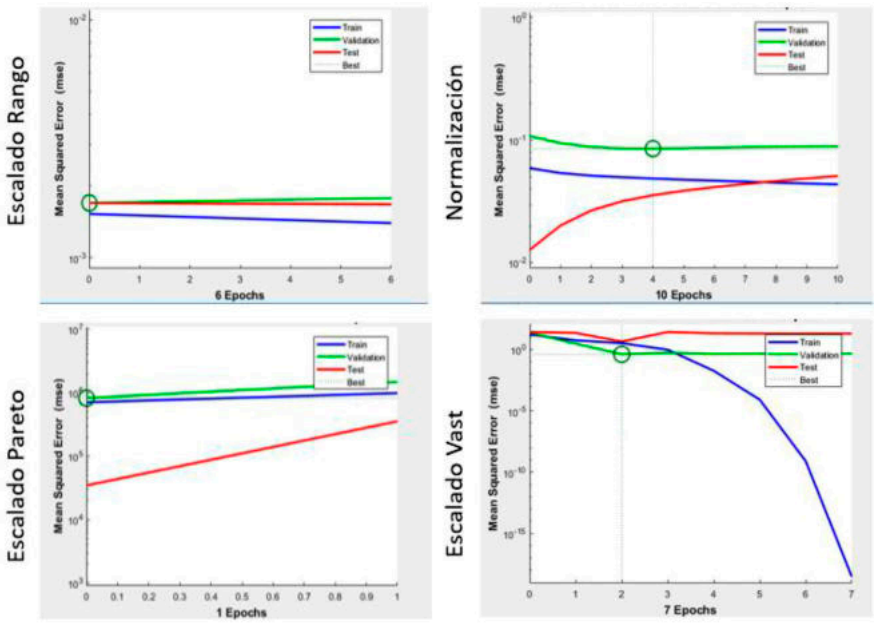

Figura 3. Desempeño según el método de escalado.

Fuente: Propia.

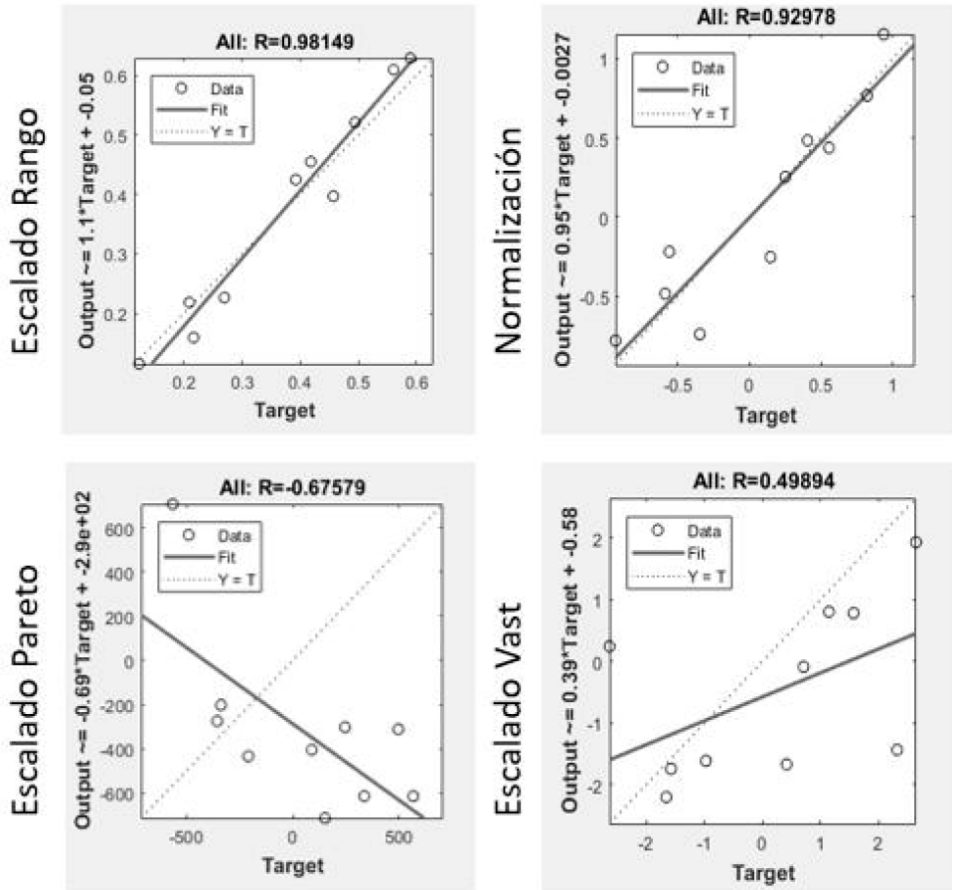

Figura 4. Modelo de ajuste según el escalado.

Fuente: Propia.

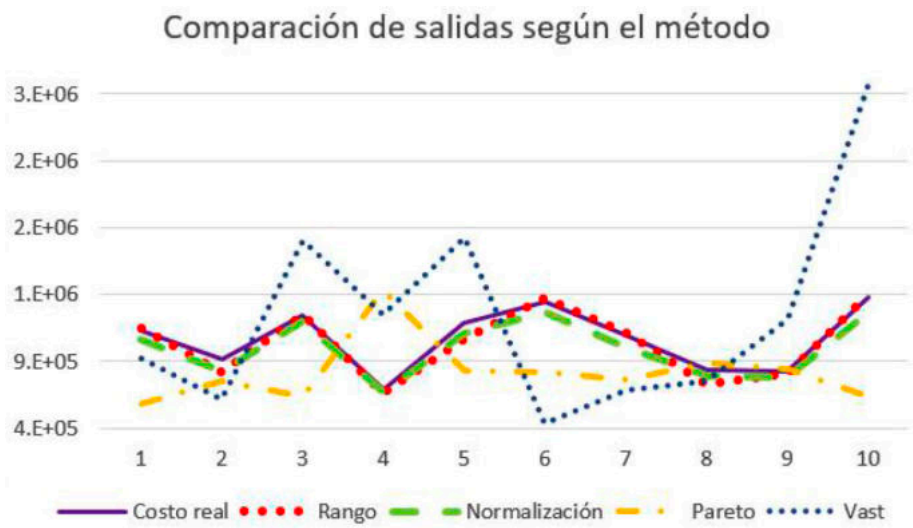

Figura 5. Comparación de salidas según el método de normalización.

Fuente: Propia.

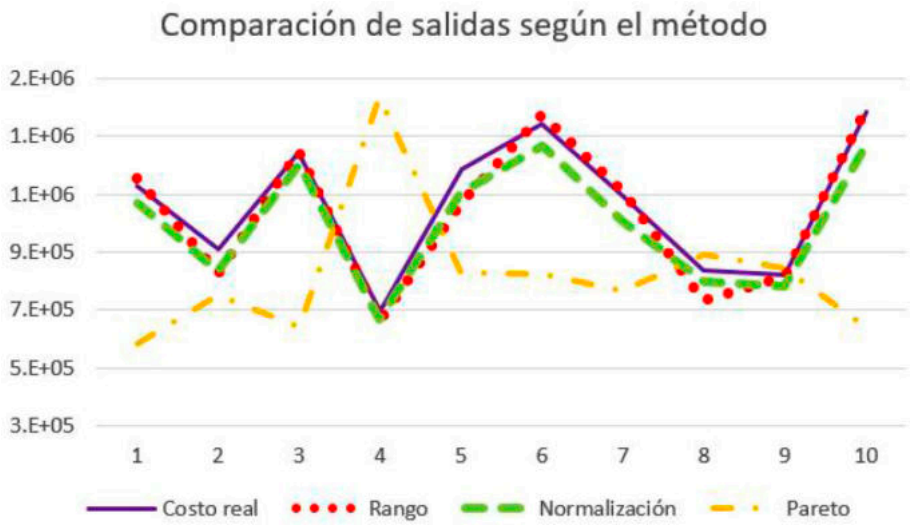

Figura 6. Comparación de salidas según el método, sin el método de Vast.

CONCLUSIONES

Los resultados presentes en este artículo denotan como los métodos de normalización y escalado colaboran en mejorar el rendimiento de la red para alcanzar la predicción de los costos.

Este artículo expone la relevancia de aplicar las propiedades de escalado y normalización para mejorar el rendimiento, los ajustes de las RNA y los modelos estadísticos utilizados orientados a la predicción de costos. No obstante, antes de efectuar el preprocesamiento, se debe considerar cuáles son las características del set de datos. Por ejemplo, para el estudio en cuestión, el set de datos contaba con dos problemas básicos; el primero, la diferencia escalar y, por otra parte, la diferencia de las unidades de medida.

En cuanto al método de Pareto y de Vast, el error cuadrático medio es elevado. Al comparar el método de Pareto con el escalado de Vast, el escalado Pareto tiene un índice mayor de error, pero al graficar los resultados del entrenamiento, el desajuste de costos afecta la predicción del escalado de Vast; en consecuencia, cuando el método tiene un error menos marcado, pero los resultados no lo reflejan, el problema se encuentra en las características de escalado. El método de Vast es un método sensible que no se ajusta a las características del set de datos estudiado, debido a que los datos sufren grandes fluctuaciones en su comportamiento. Respecto al método, la predicción se sale de los límites del costo original, mostrando un modelo muy desajustado a la realidad y, por ende, una deplorable predicción. Este modelo es sensible a las variaciones muy grandes. Por lo tanto, no es conveniente cuando las predicciones de los costos implican grandes diferencias escalares.

En la predicción de los costos de proyectos, así como en cualquier proyecto, los problemas del set de datos para ser procesados son similares a los que se encuentran en esta simulación, lo cual implica que los métodos que deben utilizarse son aquellos que mejoran las diferencias escalares y de unidades de medida. Es decir, para la selección de alguno de los métodos de normalización o escalado, debe considerarse cuales son las características del conjunto de datos, si estos son espurios, pero no fluctuantes, VAST probablemente daría un mejor ajuste, mientras que los datos como diferencias escalares significativas y que convienen magnitudes muy grandes con magnitudes de una o dos unidades.

En los resultados, se observa que el escalado de rango es el método con mayor incidencia en el ajuste y permite que la red neuronal tenga predicciones más cercanas a los datos reales.

En el caso particular de la normalización o estandarización, posibilita un desempeño de la RNA adecuado, funcionando mejor que los otros métodos como Vast y Pareto. Por lo tanto, la importancia de los métodos de escalado y normalización, radica en ajuste que realizan para que la red neuronal muestre un mejor desempeño y pueda realizar predicciones cercanas a los valores reales.

Realizar un preprocesamiento que no responde o sana las brechas que tienen los datos o no calza con los problemas específicos de la muestra, puede generar resultados índices altos de erros y con ello interpretaciones incorrectas sobre el costo que se predijo.

El método de Pareto es sensible a valores atípicos de forma que, al enfrentarse a diferencias muy radicales en los costos, se reduce el éxito en la predicción, generando que el comportamiento y desempeño de la RNA sea confuso e incongruente.

En la predicción de costos de proyectos, una práctica adecuada antes de procesar los datos es primero caracterizarlos, posteriormente aplicar el método que se ajuste a esas características y, por último, procesarlos con el método de aprendizaje automático seleccionado.

Los métodos de normalización y escalado son un factor crítico para mejorar el desempeño de una red neuronal artificial en la predicción de costos.

REFERENCIAS

[1] P. Tan, M. Steinbach and V Kumar. "Introduction to Data Mining" Person. 2° Edición. New York, Estados Unidos. 2006. ISBN: 9780133128901

[2] E. Ogasawara, L. Martínez, D. De Oliveira, G. Zimbrão, G. Pappa and M. Mattoso. "Adaptive Normalization: A novel data normalization approach for non-stationary time series". International Joint Conference on Neural Networks (IJCNN). Barcelona, España. July 2010.

[3] F. Gharehchopog and A. Maroufi. "Approach of software cost estimation with hybrid of imperialist competitive and artificial neural network algorithms". Journal of Scientific Research and Development. Vol. 1 Nº 1, pp. 50-57. 2014. ISSN: 1115-7569.

[4] L. Patil, R. Waghmode, S. Joshi and V. Khanna. "Generic model of software cost estimation: A hybrid approach". IEEE International Advance Computing Conference (IACC). Gurgaon, India. 2014.

[5] IBM Knowledge Center. "Simulation". Date of visit: December 20, 2017. URL: www.ibm.com/support/knowledgecenter/es/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/base/simulation.htm

[6] D. Vachik and D. Kamlesh. "Neural network-based models for software effort estimation: a review. Artificial Intelligence Review. Vol 42 Nº 2, pp. 295–307. 2014. ISSN: 1573-7462.

[7] T. Li, B. Jing, N. Ying and X. Yu. "Adaptive Scaling". Annals of Applied Statistics. Date of visit: December 20, 2017. URL: https://arxiv.org/pdf/1709.00566.pdf

[8] V. Mardia. "Multivariate pareto distributions". The Annals of Mathematical Statistics. Vol. 33 Nº 3, pp. 1008-1015. 1962. ISSN: 00034851.

[9] P. Gromski, P. Xu, K. Hollywood, M. Turner and R. Goodacre. "The influence of scaling metabolomics data on model classification accuracy". Metabolomics. Vol. 11 Nº 3, pp. 684-695. 2015. ISSN: 1573-3890.

[10] R. Wehrens. "Chemometrics with R - multivariate data analysis in the natural sciences and life sciences". Springer. 1º Edición. Berlin, Alemania. ISBN: 978-3-642-17840-5.

[11] J. Jackson. "A user's guide to principal components". John Wiley & Sons, INC. 1º Edición. New York, Estados Unidos. 1991. ISBN: 0-471-62267-2.

[12] F. Costantino, G. Gravio and F. Nonino. "Project selection in project portfolio management: An artificial neural network model based on critical success factors". International Journal of Project Management. Vol. 33 Nº 9, pp. 1744-1754. 2015. ISSN: 0263-7863.

[13] I. Maleki, A. Ghaffari and M. Masdari. "A New Approach for Software Cost Estimation with Hybrid Genetic Algorithm and Ant Colony Optimization". International Journal of Innovation and Applied Studies. Vol. 5 Nº 1, pp. 72-81. 2014. ISSN: 2028-9324.

[14] R. Atkinson, L. Crawford and S. Ward "Fundamental uncertainties in projects and the

scope of project management". International Journal of Project Management. Vol. 24 Nº 8, pp. 687-698. 2006. ISSN: 0263-7863.

[15] K. Polat. "A novel data preprocessing method to estimate the air pollution (SO2): neighbor-based feature scaling (NBFS)". Neural Computing and Applications. Vol. 21 Nº 8, pp. 1987-1994 2011. ISBN: 1433-3058.

[16] P. Gromski, Y. Xu, K. Hollywood, M. Turner and R. Goodacre. "The influence of scaling metabolomics data on model classification accuracy". Metabolomics. Vol. 11 Nº 3, pp. 684-695. 2014. ISBN: 1573-3890.

[17] H. Keun, T. Ebbels, H. Antti, M. Bollard, O. Beckonert, E. Holmes, *et al.* "Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling". Analytica Chimica Acta. Vol. 490 Nº 1-2, pp. 265-276. 2003. ISSN: 0003-2670.