risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 13/09/2018
Aceitação/Acceptance: 11/11/2018

Un estudio comparativo de técnicas de minería de datos y aprendizaje máquina para la estimación del esfuerzo utilizando puntos de función

Christian Quesada López[1], Juan Murillo-Morera[1,2], Marcelo Jenkins[1]

{cristian.quesadalopez;marcelo.jenkins}@ucr.ac.cr, juan.murillo.morera@una.cr

[1] Universidad de Costa Rica, San Pedro, Costa Rica.

[2] Universidad Nacional de Costa Rica, Heredia, Costa Rica.

Pages: 595–609

Resumen: En los últimos años, una gran cantidad de técnicas de minería de datos y de aprendizaje máquina han sido utilizadas para la construcción de modelos de estimación del esfuerzo de desarrollo del software. La literatura ha reportado resultados inconsistentes acerca de la efectividad de los modelos dependiendo de los conjuntos de datos. En este estudio utilizamos un procedimiento automatizado para la comparación exhaustiva de modelos de estimación de esfuerzo y presentamos los resultados del análisis comparativo a partir de la combinación de un conjunto de técnicas de pre-procesamiento de datos, selección de atributos y algoritmos de aprendizaje aplicado a distintos sub conjuntos de datos del repositorio ISBSG. Los resultados indican que las técnicas con mejores resultados para los modelos basados en los puntos de función IFPUG-FPA fueron *LeastMedSq*, *SMOreg* y *GaussianProcesses* y para COSMIC-FFP fueron *SMOreg*, *MP5* y *AdditiveRegression*. Las técnicas que incorporan estrategias de regresión son las que mejores resultados obtienen. Asimismo, la combinación de técnicas de pre procesamiento y selección de atributos mejoran los resultados de exactitud. Los modelos de estimación evaluados alcanzaron valores en la exactitud estandarizada entre el 49.94% y 64.05% para IFPUG-FPA y entre el 80.30% y el 67.31% para COSMIC-FFP. Con nuestro procedimiento de evaluación es posible analizar la exactitud de distintos modelos de estimación, cuáles técnicas obtienen los mejores resultados de exactitud a partir de cada conjunto de datos y la combinación de técnicas que puede mejorar el desempeño de los modelos.

Palabras-clave: Estimación de esfuerzo de desarrollo del software; puntos de función, COSMIC-FFP IFPUG-FPA; minería de datos; aprendizaje máquina; estudio empírico.

A comparative study of data mining and machine learning techniques for software effort estimation using function points

Abstract: In recent years, a large number of data mining and machine learning techniques have been used for the construction of software development effort estimation models. The literature has reported inconsistent results about the

effectiveness of the models because of their dependence on the data sets. In this study, we use an automated procedure for the exhaustive comparison of effort estimation models, and present the results of a comparative analysis derived from the combination of a set of data pre-processing, attribute selection techniques and learning algorithms applied to different sub sets of data from the ISBSG repository. The results indicate that the techniques with the best results for models based on IFPUG-FPA function points were LeastMedSq, SMOreg, and GaussianProcesses, but for COSMIC-FFP were SMOreg, MP5, and AdditiveRegression. The techniques that incorporate regression strategies yield the best results obtained for these data sets. In addition, the combination of pre-processing techniques and attributes selection improve the accuracy results. The estimation models reached a standardized accuracy between 49.94% and 64.05% for IFPUG FPA and between 80.30% and 67.31% for COSMIC FFP. With our evaluation procedure, it is possible to analyze the accuracy of different estimation models, which techniques obtain the best accuracy results from each data set, as well as the combination of techniques that can improve the performance of the models.

***Keywords**:* Software effort estimation; function points COSMIC-FFP IFPUG-FPA; data mining; machine learning; empirical study.

1. Introducción

La necesidad de modelos de estimación de esfuerzo y productividad precisos es uno de los aspectos más importantes en la ingeniería del software (Abran, 2015). En los últimos años, la evaluación de estos modelos de estimación ha sido un área de investigación activa en la comunidad de la ingeniería del software (Dejaeger, Verbeke, Martens & Baesens, 2012). Sin embargo, la estimación del esfuerzo utilizando el tamaño funcional como predictor es aún un reto para los investigadores y profesionales (Jorgensen & Shepperd, 2007; Gencel, 2008). Una gran cantidad de técnicas de minería de datos y de aprendizaje máquina han sido utilizadas para la construcción de los modelos de estimación (Wen, Li, Lin, Hu & Huang, 2012; Dejaeger et al, 2012; Huang, Li & Xie, 2015), pero los resultados reportados acerca de la efectividad de las técnicas utilizadas para la construcción de los modelos han sido inconsistentes dada su dependencia con los conjuntos de datos utilizados como entrada (Dejaeger et al, 2012).Por lo tanto, son necesarios estudios empíricos que permitan realizar comparaciones sin sesgo (Kitchenham & Mendes, 2009).

La evaluación de combinaciones de técnicas para distintos conjuntos de datos es un campo de investigación abierto (Keung, Kocaguneli & Menzies, 2013). Uno de los principales desafíos en la evaluación de los modelos de estimación de esfuerzo se relaciona con la configuración de las técnicas, la selección de los datos y las métricas de evaluación, dado que una validación incompleta o inapropiada puede presentar sobreajuste o sub ajuste de los resultados. Un procedimiento de evaluación de modelos debe implementar las recomendaciones de los marcos de trabajo de estimación propuestos en la literatura para lograr un proceso sistemático y sin sesgo que permita la comparación de los resultados de la exactitud de las técnicas utilizadas para la estimación de esfuerzo (Song, Jia, Shepperd, Ying & Liu, 2011; Menzies & Shepperd, 2012; Shepperd & MacDonell, 2012; Song, Minku & Yao, 2013; Keung et al. 2013; Huang et al, 2015; Langdon, Dolado, Sarro & Harman, 2016). Las evaluaciones comparativas que permitan una selección

acertada de los modelos de estimación en un contexto determinado es clave para tomar decisiones acerca de su efectividad (Shepperd, 2007).

El objetivo del estudio es realizar una comparación exhaustiva de modelos de estimación de esfuerzo a partir de la combinación de un conjunto de técnicas de pre-procesamiento de datos, selección de atributos y algoritmos de aprendizaje aplicado a distintos sub conjuntos de datos del repositorio de proyectos del *International Software Benchmarking Standards Group* (ISBSG). El repositorio de proyectos ISBSG proporciona información de proyectos de software de múltiples compañías, países, industrias y ambientes de desarrollo. En su versión (R12), el repositorio contiene 6 mil proyectos con 105 características entre las cuales se encuentra el tamaño funcional, el esfuerzo de desarrollo y los factores de contexto (ISBSG, 2013).

Para realizar la evaluación, utilizamos un procedimiento automatizado que recibe como entrada las métricas de tamaño funcional y un conjunto de factores de contexto de proyectos de desarrollo de aplicaciones de negocios y calcula las métricas de exactitud. En total, se evalúan 600 distintos modelos de estimación producto de la combinación de las técnicas seleccionadas que se aplican a treinta grupos de datos escogidos a partir del repositorio ISBSG R12. Este estudio compara algoritmos de las familias de funciones y regresiones, bayesianos, árboles, reglas y redes. Los proyectos seleccionados son aplicaciones medidas bajo los métodos de medición de tamaño funcional IFPUG FPA y COSMIC FFP, ya que el método IFPUG FPA es el más utilizado en la industria (Fingerman, 2011) y el método COSMIC FFP es el de mayor crecimiento en adopción en los últimos años (Dumke & Abran, 2011).

El resto del artículo se estructura de la siguiente manera: la Sección 2 presenta los trabajos relacionados, la Sección 3 describe el procedimiento para la comparación de los modelos de estimación, la Sección 4 describe el estudio empírico, la Sección 5 discute los resultados y finalmente, la Sección 6 presenta las conclusiones y trabajo futuro.

2. Trabajo Relacionado

En el área de la estimación del esfuerzo de desarrollo del software se ha estudiado como mejorar la estabilidad de los resultados de los modelos de estimación, la selección y configuración de las técnicas utilizadas, y la selección de los conjuntos de datos para obtener resultados confiables en los estudios comparativos. La investigación sobre la estimación plantea distintos enfoques para la evaluación comparativa de modelos o combinación de estos, estrategias de línea base y una correcta selección de métricas de evaluación que permitan realizar comparaciones con el menor sesgo posible (Whigham, Owen & Macdonell, 2015; Saeed, Butt, Kazmi & Arif, 2018). Estos se dirigen, principalmente, a lograr la estabilidad de los resultados de evaluación a través de diferentes conjuntos de datos. Dejaeger et al. (2012) establecen que la selección y combinación de diferentes técnicas de minería de datos e inteligencia artificial permiten mejorar los resultados de la evaluación de los métodos de estimación para conjuntos de datos con características específicas. La motivación detrás de la evaluación de estas combinaciones es depender lo mínimo posible de las características de los datos cuando se realiza el entrenamiento de los modelos de estimación. Los enfoques de ingeniería de software basados en búsquedas (SBSE) pueden apoyar los procesos de selección de

técnicas y parámetros (Ferrucci, Harman & Sarro, 2014; Xia, Mathew, Shen & Menzies, 2018) dado que estos modelos dependen de las características de los datos y no existen reglas únicas para decidir cuál es la mejor combinación y configuración de las técnicas.

Oliveira, Braga, Lima & Cornélio (2010) realizaron investigaciones comparativas para evaluar el efecto de la configuración de parámetros en el desempeño de los modelos de estimación de esfuerzo construidos mediante técnicas de regresión. Del mismo modo, Song et al. (2013) realizaron comparaciones para evaluar el efecto de los de parámetros en el desempeño de los modelos construidos a partir de técnicas de aprendizaje máquina. Por su parte, Dejaegeret al. (2012) realizaron una comparación empírica a gran escala usando diferentes tipos de técnicas de inteligencia artificial y minería de datos para analizar los aspectos relacionados con la selección de las características de los conjuntos de datos, al igual que Liu, Xiao & Zhu. (2018) quienes desarrollaron una propuesta utilizando como principales técnicas la selección de atributos y el razonamiento basado en casos para proyectos con características semejantes. Keung et al. (2013) evaluaron noventa predictores con veinte conjuntos de datos y siete métricas de evaluación para determinar la estabilidad según la selección de los predictores. Finalmente, Huang et al. (2015) estudiaron un conjunto de técnicas de pre procesamiento y de aprendizaje máquina y analizaron las características de los conjuntos de datos y algoritmos de acuerdo a su efectividad.

Con el fin de establecer marcos de trabajo para la evaluación comparativa de modelos de estimación de esfuerzo, múltiples estudios en el área han propuesto enfoques para mejorar las comparaciones. Shepperd & MacDonell (2012) propusieron un marco de trabajo para evaluación de modelos de estimación que reduce la inconsistencia, contempla el uso de métricas que disminuyen el sesgo, y permite la agregación de resultados entre estudios y futuros meta análisis. Menzies & Shepperd (2012) y Keung et al. (2013) discuten el problema de la inestabilidad de los sistemas de estimación y analizan las fuentes de inestabilidad. Azhar, Riddle, Mendes, Mittas & Angelis (2013) proponen un marco de trabajo estadístico basado en un algoritmo de múltiples comparaciones para clasificar la efectividad de los modelos de estimación de esfuerzo. Whigham et al. (2015) presentaron un método para establecer una línea base de comparación para los modelos de estimación de esfuerzo basado en regresiones lineales múltiples. Dolado, Rodriguez, Harman, Langdon & Sarro (2016) proponen una medida para comparar modelos de estimación basada en un enfoque de pruebas del mínimo intervalo de equivalencia para el error absoluto y una estimación aleatoria como línea base de referencia. Langdon et al. (2016) proponen un método para establecer líneas base de selección aleatoria (*random guessing*) para la comparación de los modelos de estimación de esfuerzo. Finalmente, Lavazza & Morasca (2017) proponen un marco de trabajo para la construcción de indicadores de exactitud sobre modelos de estimación de esfuerzo. Estos indicadores son basados en los cuadrados y los valores absolutos de los residuos de los resultados obtenidos por los modelos.

3. Procedimiento automatizado para la comparación de modelos de estimación de esfuerzo

En esta sección se presenta el procedimiento automatizado para la comparación exhaustiva de modelos de estimación de esfuerzo. Este presenta una estrategia

sistemática de evaluación de modelos que permite realizar una comparación reduciendo el sesgo entre los distintos modelos basados en los métodos IFPUG FPA y COSMIC FFP. Este procedimiento presenta una extensión de los trabajos previos realizados en Quesada López & Jenkins (2016) y Murillo-Morera, Quesada-López, Castro-Herrera & Jenkins (2017) donde se han refinado las estrategias de construcción y evaluación de los modelos de estimación, el procesamiento de los conjuntos de datos, y se han agregado las funcionalidades de parametrización. La estrategia de evaluación se basa principalmente en las recomendaciones del marco de trabajo propuesto por Song et al. (2011) y trabajos en el área de modelos de estimación de esfuerzo tales como (Dejaeger et al., 2012; Langdon et al., 2016; Menzies, Yang, Mathew, Boehm & Hihn, 2017; Lavazza & Morasca, 2017). El procedimiento para la evaluación construye los modelos de estimación de esfuerzo y los evalúa para determinar cuál combinación de técnicas genera los mejores resultados a partir de las características de los conjuntos de datos.

A continuación, describimos la estrategia de evaluación que utiliza las mediciones de tamaño funcional y los factores de contexto como predictores, y obtiene como salida la estimación del esfuerzo de desarrollo de un proyecto de software. Con este trabajo definimos, implementamos y evaluamos un marco de trabajo exhaustivo de comparación de modelos de estimación de esfuerzo el cual tiene la flexibilidad para incorporar nuevas técnicas de pre procesamiento, selección de atributos, algoritmos de aprendizaje y su configuración de parámetros, que permite determinar los mejores modelos para un conjunto de datos específico.

La Figura 1 presenta los componentes de la estrategia de evaluación de modelos. Primero, el componente de evaluación permite analizar el desempeño de los modelos a partir de los datos históricos de proyectos y segundo, el componente de estimación de esfuerzo de desarrollo construye los modelos a partir de las técnicas que mejor resultado presentaron en el contexto y realiza la estimación de esfuerzo a partir de los nuevos datos. En el proceso de construcción de los modelos se combinan distintas técnicas para el pre procesamiento de datos, la selección de atributos o características y los algoritmos de aprendizaje. Cada combinación de los tres tipos de técnicas representa un esquema de aprendizaje que es comparado durante la aplicación de la estrategia de evaluación. A continuación describimos las fases de la estrategia:

Construcción de los modelos de estimación de esfuerzo: en esta fase se analizan los modelos de estimación construidos a partir de cada uno de los esquemas de aprendizaje para determinar su desempeño. Los modelos se construyen para cada uno de los esquemas de aprendizaje y los datos históricos de los proyectos desarrollados que funcionan como los conjuntos de entrenamiento. Cada esquema de aprendizaje se construye a partir de la selección de una técnica de pre procesamiento de datos (DP), una de selección de atributos (AS) y un algoritmo de aprendizaje (LA). Los parámetros de las técnicas o algoritmos de aprendizaje son configurados para determinar cuáles ofrecen un mejor resultado en las métricas de desempeño. El proceso de construcción contempla los siguientes aspectos:

a. Los datos históricos se dividen en dos conjuntos de datos, un conjunto de entrenamiento (90%) que permite la construcción de los modelos a partir de los distintos esquemas de aprendizaje y un conjunto de prueba (10%) que permite evaluar el desempeño de los modelos.

b. Para la separación de los conjuntos de datos se realiza un proceso de una ronda de validación cruzada MxN. Se realizan M repeticiones donde se aplica una aleatorización al conjunto de datos históricos y se generan N particiones de las cuales se asigna un 10% de las instancias para el conjunto de datos de prueba y un 90% de las instancias para el conjunto de datos de entrenamiento. La selección se realiza aleatoriamente donde se hace la separación de N partes, el modelo es construido a partir de las N-1 partes y la parte restante es utilizada para las pruebas. Esto se repite para las N rondas lo que permite que cada parte sea utilizada para el entrenamiento y las pruebas lo que minimiza el sesgo de la selección. El proceso se repite M veces aplicando una aleatorización para reducir el efecto del ordenamiento. El conjunto de datos de prueba es independiente del proceso de construcción de los modelos y es utilizado para evaluación.
c. El proceso de validación cruzada permite la construcción y evaluación de MxN modelos por cada conjunto de datos y para cada uno de los esquemas de aprendizaje. La estrategia de múltiples rondas de validación cruzada para cada uno de los conjuntos de datos se realiza para reducir la variabilidad de los resultados del desempeño de cada uno de los modelos construidos. Los resultados de cada una de las rondas son promediados para obtener el valor total del desempeño.
d. Los modelos se construyen y evalúan para cada uno de los esquemas de aprendizaje a partir del conjunto de datos de entrenamiento realizando tres pasos: Primero, se aplica la técnica de pre procesamiento de datos para la transformación, la creación de variables ficticias en escala [0, 1], eliminación de valores atípicos y el tratamiento de los valores faltantes. Segundo, se realiza la selección de atributos para la reducción de la dimensión del conjunto de datos aplicando métodos de filtrado o envoltura y evaluando el desempeño de cada uno de los subconjuntos seleccionados mediante una validación cruzada que permite determinar el subconjunto de atributos que muestra mejor desempeño. Tercero, se realiza la construcción del modelo utilizando las técnicas de aprendizaje de máquina y de regresión a partir del conjunto de datos reducido. En este caso se evalúa el desempeño de distintas configuraciones de parámetros mediante una validación cruzada que permite determinar la configuración con mejor desempeño.
e. Para realizar el proceso de validación, el conjunto de datos de prueba es tratado del mismo modo que el conjunto de entrenamiento y los valores reales y de estimación son calculados a partir de las métricas de desempeño. Cada uno de los resultados de evaluación se almacena para calcular los promedios y varianzas durante cada uno de los procesos de validación cruzada de MxN rondas.

El proceso de evaluación de esquemas de aprendizaje permite construir los modelos y determinar los mejores esquemas de aprendizaje que pueden ser utilizados para la estimación del esfuerzo de desarrollo para los nuevos datos.

Evaluación de modelos de estimación de esfuerzo*:* en la fase de evaluación y a partir de los resultados de la fase de construcción de modelos se selecciona el o los esquemas de aprendizaje para construir los modelos que permiten realizar las estimaciones de esfuerzo a partir de los nuevos datos y evaluar su desempeño. El

proceso obtiene la estimación de esfuerzo para los nuevos datos y reporta los resultados de estimación y desempeño. El proceso de evaluación contempla:

a. Para realizar la estimación de esfuerzo se seleccionan los mejores esquemas de aprendizaje y se construyen los modelos a partir dell conjunto de datos históricos.
b. El proceso de creación de los modelos de estimación se realiza de acuerdo a la estrategia descrita en la primera fase. Se pre procesan los datos, se realiza la selección de los mejores atributos para los nuevos datos y se selecciona la mejor configuración de parámetros de acuerdo a los resultados de los modelos analizados en la fase anterior.

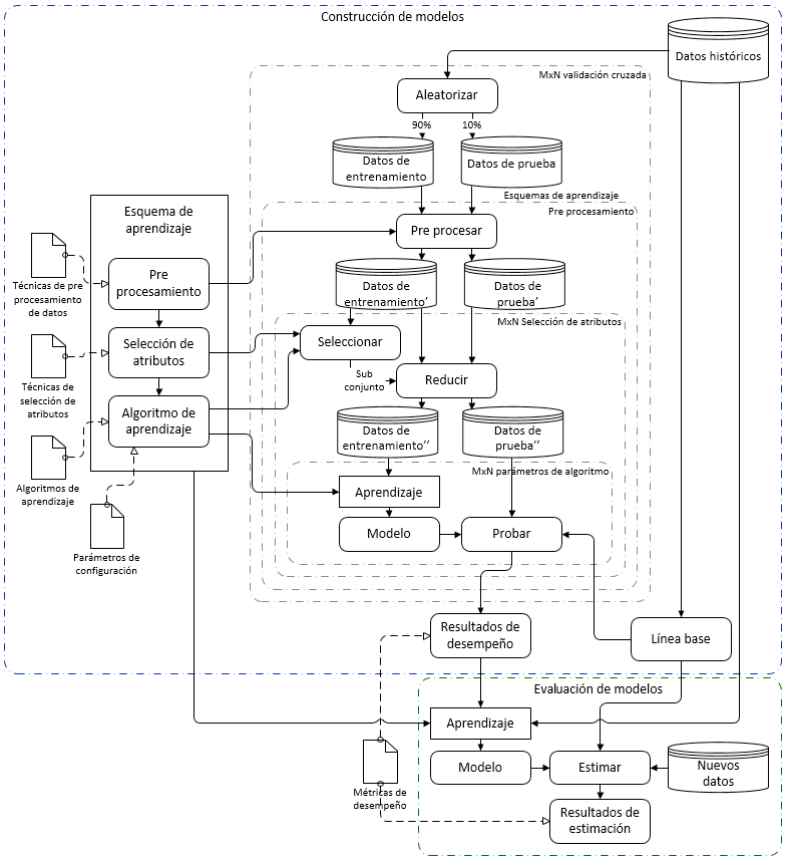

Figura 1 – Componentes de la estrategia de evaluación de modelos de estimación.

La estrategia busca la estabilización de los modelos y reducir el riesgo de que los mismos presenten problemas de sobre ajuste o sub ajuste que puede producir resultados no confiables sobre el desempeño. Asimismo, mediante la implementación de las estrategias de modelos de línea base, bajo el mismo conjunto de datos, es posible

realizar comparaciones más objetivas sobre el desempeño de los distintos modelos de estimación. Esto es posible porque la estrategia de evaluación plantea la comparación de los resultados de los modelos construidos para cada uno de los esquemas de aprendizaje contra los modelos construidos a partir de la línea base, ambos a partir de los mismos datos en el contexto evaluado. En este caso, se comparan los modelos generados por la estrategia exhaustiva contra los modelos basados en una estrategia de selección aleatoria de acuerdo al marco de trabajo de línea base propuesto por Langdon et al. (2016). Las técnicas específicas para cada uno de los componentes de los esquemas de aprendizaje se listan en La Sección 4.

La Figura 2 muestra los compontes de la herramienta prototipo para la evaluación de los modelos de estimación de esfuerzo. Esta incluye los siguientes componentes: (1) El componente de construcción de modelos se encarga de construir y evaluar el desempeño de los modelos creados a partir de cada uno de los esquemas de aprendizaje y los datos históricos. Este componente incluye el pre procesamiento de datos, la selección de atributos y la configuración de los parámetros de los algoritmos de aprendizaje. El proceso de evaluación construye los modelos y determina los mejores esquemas de aprendizaje que son utilizados para la estimación de los nuevos datos. (2) El componente de evaluación permite aplicar los mejores esquemas de aprendizaje para la construcción de los modelos utilizados en la estimación del esfuerzo de desarrollo del software. Este componente incluye las mejores técnicas de pre procesamiento, selección de atributos, configuración de atributos y algoritmo de aprendizaje que son aplicados en la estimación del esfuerzo de los nuevos datos. (3) El componente de reporte se encarga de realizar el proceso de presentación de los resultados de evaluación de los modelos y de la estimación del esfuerzo de desarrollo. Este componente incluye el reporte de los resultados de desempeño para cada uno de los esquemas de aprendizaje y el resultado del proceso de estimación.

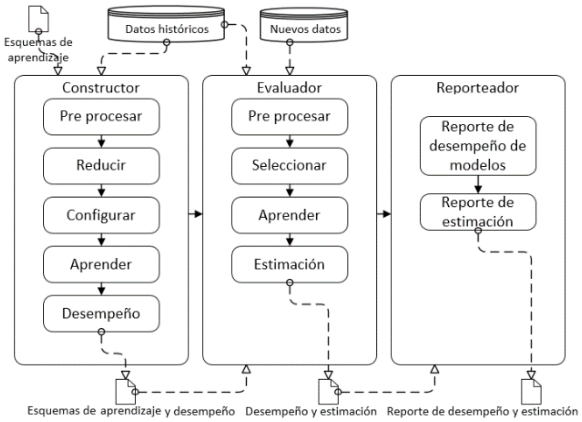

Figura 2 – Herramienta prototipo de evaluación de modelos.

La herramienta que implementa la estrategia de evaluación se desarrolla en Java. Esta utiliza como base las técnicas implementadas en las librerías de *Waikato Environment for Knowledge Analysis* (Witten, Frank, Hall & Pal, 2016).

4. Descripción del estudio empírico

La estructura del estudio empírico se realiza de acuerdo a los lineamientos por Wohlin, Runeson, Höst, Ohlsson, Regnell & Wesslén (2012). El estudio realiza un análisis empírico exhaustivo donde (a) se valida la estabilidad de los resultados del procedimiento para la construcción de modelos de estimación de esfuerzo basados en el tamaño funcional, (b) se estudian modelos simplificados de estimación de esfuerzo y (c) se reportan las técnicas que presentan los mejores resultados para los conjuntos de datos estudiados. Las preguntas de investigación analizadas son:

RQ1. ¿Cuál es la consistencia de los resultados de la exactitud obtenidas por el procedimiento de evaluación?

RQ2. ¿Cuál es la exactitud de los modelos construidos a partir de la combinación de técnicas para los conjuntos de datos?

El proceso general del estudio se lleva a cabo en dos fases. Primero, realizamos la selección de datos a partir del repositorio de proyectos ISBSG R12 para obtener los sub conjuntos de proyectos y variables utilizadas en el experimento. Segundo, se evalúan los modelos de estimación de esfuerzo y se determina el nivel de exactitud de las estimaciones. Para la recolección de datos se aplica la estrategia de evaluación descrita en la Sección 3.

Selección del conjunto de datos. Los conjuntos de datos (DS) utilizados en este estudio son aplicaciones de negocio del dominio de sistemas de información administrativo (MIS) medidas bajo los métodos de medición IFPUG FPA y COSMIC FFP. La Tabla 1 muestra los parámetros de selección utilizados. Se seleccionan los proyectos con calidad A o B, del método de medición IFPUG FPA y COSMIC FFP y del año 2005 o superior. Este estudio incorpora las variables de contexto de los proyectos tales como el tipo de desarrollo (TD), el tipo de lenguaje (TL), el lenguaje programación (LP), la plataforma de desarrollo (PD), la arquitectura (AR) y el tamaño del equipo de desarrollo (TE).

DS	Método	Calidad datos	Calidad conteo	Periodo	N	Grupo	Nivel de recursos	Dominio	Tipo
DS202	IFPUG FPA 4+	A	A|B	2005-2011	202	BA	1	BA	Multi
DS053	COSMIC FFP 2+	A|B	A|B	2008-2012	53	BA	1	BA	Multi

Tabla 1 – Conjuntos de datos seleccionados del ISBSG R12.

Cada uno de los conjuntos de datos (DS202 y DS53) es dividido manualmente para obtener distintos grupos que contengan subconjuntos de los componentes funcionales básicos (BFC) del tamaño funcional y variables de contexto que permitan realizar

comparaciones de efectividad de los modelos. Por ejemplo, conjuntos de datos que solo contengan el tamaño funcional total, sub conjuntos de BFC y/o variables de contexto. Los conjuntos de datos del método de medición IFPUG FPA (DS202) se dividen en 21 grupos y los conjuntos de datos del método de medición COSMIC FFP (DS53) se dividen en 9 grupos, todos con distintas variables predictoras de tamaño funcional y contexto.

Las variables de contexto (atributos nominales) son pre procesadas para codificarlas con valores entre 0 y 1 (*dummy coding*). No se aplica ningún tratamiento para los valores faltantes de las variables de contexto. Se aplican técnicas de distancia para la identificación de los valores atípicos.

Técnicas utilizadas para la construcción de modelos. En total se utilizan 2 técnicas de pre procesamiento de datos (DP), 5 técnicas de selección de atributos (AS) y 15 métodos de algoritmos de aprendizaje (LA) que son combinadas para la construcción de los modelos de estimación (DP×AS×LA). En el caso de la técnica de transformación de BoxCox se utilizan 6 parámetros para *lambda*. Las técnicas de DP permiten limpiar el ruido, eliminar datos faltantes y valores atípicos y transformar las variables. Las técnicas de AS permiten seleccionar los atributos predictores más apropiados que explican las variaciones para predecir la variable de salida. Finalmente, los LA son los modelos que permiten realizar las estimaciones. Las principales categorías de algoritmos utilizados son las funciones y regresiones, bayesianos, árboles, reglas y redes. Los algoritmos utilizados en la evaluación son los implementados por las librerías de WEKA para valores continuos (Witten, Frank, Hall & Pal, 2016). Se aplica el enfoque de *filter* para la reducción de las variables de los conjuntos de datos. Se selecciona este enfoque dado que, aunque no ofrece resultados de exactitud tan buenos como un enfoque de *wrapper*, es menos costoso computacionalmente.

Las técnicas de DP son la transformación logaritmo (LG) y *BoxCox* (BC), las técnicas de DS son *GeneticSearch* (GS), *BestFirst* (BF), *LinearForwardSelection* (LFS), *BackwardElimination* (BE), *ForwardSelection* (FS), y finalmente, los algoritmos LA son *GaussianProcesses* (GP), *LeastMedSq* (LMS), *LinearRegression* (LR), *MultilayerPerceptron* (MP), *RBFNetwork* (RBFN), *SMOreg* (SMO), *AdditiveRegression* (AR), *Bagging* (BGN), *ConjunctiveRule* (CR), *DecisionTable* (DT), *M5Rules* (M5R), *ZeroR* (ZR), *DecisionStump* (DS), *M5P* (M5P) y *REPTree* (RT). Las configuraciones de parámetros específicas utilizadas para cada una de las técnicas implementadas por los LA se seleccionan a partir de las recomendaciones presentadas por la documentación de las librerías WEKA (Witten et al., 2016).

Por cada uno de los sub conjuntos de datos se construye un modelo para cada uno de los esquemas de aprendizaje (DP×AS×LA). En esta evaluación, para cada uno de los sub conjuntos de datos se ejecuta un N_PASS=10 donde para cada pasada de la evaluación se calculan y promedian las métricas de exactitud. En total se realizan mil ejecuciones (N_PASS=10, M×N donde M=10 y N=10) por esquema de aprendizaje (DP×AS×LA).

Métricas de evaluación de exactitud. La exactitud de los modelos de estimación de esfuerzo se analiza basada en las métricas de exactitud: magnitud del error relativo (MRE), coeficiente de correlación de Spearman (SP), exactitud estandarizada (SA) y número de predicciones de los valores reales (Pred25) de acuerdo a las recomendaciones de estudios en el área.

Amenazas a la validez. El tamaño y las características limitadas de los conjuntos de datos utilizados en los análisis puede ser una amenaza a la validez interna. Los datos fueron filtrados para asegurar que solo se utilizó información de alta calidad. Fueron utilizadas técnicas, estrategias y métricas de evaluación recomendados en estudios previos del área. Los conjuntos de datos son del dominio de aplicaciones de negocios del ISBSG por lo que los resultados solo pueden generalizarse a este contexto y de acuerdo a los paramentos de selección de proyectos aplicados en el estudio.

5. Análisis de Resultados

En esta sección se reportan los resultados de la evaluación.

Consistencia de los resultados de la exactitud (RQ1). Para verificar la consistencia de la estrategia de evaluación se recolectan los resultados de las métricas de exactitud para cada uno de los modelos de estimación de esfuerzo mediante dos corridas independientes. Para cada uno de los grupos de los conjuntos de datos DS202 y DS053 se construyen los 600 modelos a partir de la combinación de las técnicas de DP (n=8), las técnicas de AS (n=5) y los LA (n=15) y se comparan los resultados de las métricas MRE, SP, SA y Pred25. Se aplica la prueba estadística *Wilcoxon signed Rank* para determinar la consistencia entre los resultados obtenidos por cada modelo de cada uno de los grupos de datos. Los resultados de las métricas de evaluación no pertenecen a una distribución normal, por lo que se aplica esta prueba no paramétrica pareada entre los resultados de las dos corridas.

Los resultados de las pruebas para cada grupo de datos y las métricas indican que no se encuentra una diferencia significativa entre los resultados de exactitud entre las dos corridas realizadas con la estrategia de evaluación. Esto implica que la estrategia de evaluación utilizada permite la obtención de resultados consistentes. En la práctica, la consistencia de la estrategia de evaluación permite realizar comparaciones exhaustivas y sin sesgo entre los modelos de estimación de esfuerzo para determinar el modelo que mejor desempeño produce en el contexto de evaluación. Del mismo modo, al permitir aplicar una estrategia sin sesgo de selección de proyectos es posible mejorar la confiablidad en cuanto al sobre ajuste en los resultados de los modelos. Por tanto, a partir de los resultados obtenidos en la evaluación de los modelos, es posible analizar la exactitud de los modelos de estimación de esfuerzo, las técnicas que obtienen los mejores resultados de exactitud y los aspectos que pueden impactar el desempeño.

Exactitud de los modelos de estimación de esfuerzo (RQ2). Para determinar los mejores modelos se utilizan los resultados de la métrica de exactitud estandarizada (SA). Los resultados de los 30 grupos de los dos conjuntos de datos analizados, 21 para los basados en el método IFPUG FPA (DS202) y 9 para los basados en el método COSMIC FFP (DS053) son seleccionados a partir de los mejores resultados de exactitud. Los modelos de estimación evaluados alcanzaron valores entre el 49.94% y el 64.05% para el método de medición IFPUG FPA y entre el 67.31% y el 80.30% para el método de medición COSMIC FFP con base en la exactitud estandarizada (SA). Esto significa que existen modelos basados en los proyectos del método IFPUG FPA que podrían superar la línea base de selección aleatoria hasta por un 54.05% y hasta un 80.30% en el caso del método COSMIC FFP. Los resultados obtenidos son similares a los reportados en

estudios previos en el área que han alcanzado valores que oscilan entre el 30% y 60% con valores de hasta el 80% (Whigham et al., 2015; Dolado et al., 2016).

Considerando los modelos con los mejores resultados (1% superior del total de 600 modelos por grupo), la mayor frecuencia de técnicas (LA) para el conjunto de datos del método IFPUG FPA son las de la regresión *LeastMedSq* que representa un 54.0% (n=68) del total y *SMOreg* que representa un 41.3% (n=52) del total. En el caso de los conjuntos de datos del método COSMIC FFP son las de *SMOreg* que representa un 29.6% (n=16) del total, *MP5* que representa un 29.6% (n=16) del total y *AdditiveRegression* que representa un 24.1% (n=13) del total. Todas estas técnicas (LA) incorporan estrategias de regresión para su implementación.

Las mejores combinaciones de técnicas de DP×AS×LA para los grupos del conjunto del método IFPUG FPA son las de BC×BF×LMS que representa un 10.3% (n=13) del total, BC×BE×SMO que representa un 9.5% (n=12), BC×BE×LMS que representa un 8.7% (n=11) y BC×LFS×LMS que representa un 7.9% (n=10). En el caso del método de medición COSMIC FFP fueron las de BC×LFS×SMO que representa un 7.4% (n=4) del total, BC×FS×AR que representa un 7.4% (n=4) y None×BF×M5P que representa un 5.6% (n=3). En general, los mejores resultados las obtuvieron las técnicas *LeastMedSq*, *SMOreg* y *GaussianProcesses* para los conjuntos de datos del método IFPUG FPA y *SMOreg*, *MP5* y *AdditiveRegression* para COSMIC FFP. Asimismo, los resultados indicaron que la combinación con técnicas de pre procesamiento y selección de atributos mejoraron los resultados de exactitud. Las combinaciones que mejores resultados obtuvieron para IFPUG FPA fueron las de BC×BF×LMS, BC×BE×SMO, BC×BE×LMS, y BC×LFS×LMS. En el caso de COSMIC FFP fueron las combinaciones de BC×LFS×SMO, BC×FS×AR y BF×M5P.

6. Conclusiones

Los resultados del estudio indican que el procedimiento de evaluación presentó resultados consistentes a través de múltiples corridas evitando el sobreajuste o sub ajuste de los modelos. Los modelos de estimación evaluados alcanzaron valores entre el 49.94% y el 64.05% para el método de medición IFPUG FPA y entre el 67.31% y el 80.30% para el método de medición COSMIC FFP con base en la exactitud estandarizada (SA).

Las técnicas de construcción de modelos (LA) que mejores resultados obtuvieron para los conjuntos de datos del método IFPUG FPA fueron: *LeastMedSq*, *SMOreg* y *GaussianProcesses*. En el caso de los conjuntos de datos del método COSMIC FFP los mejores resultados los obtuvieron las técnicas: *SMOreg*, *MP5* y *AdditiveRegression*. Los resultados mostraron que las técnicas que incorporaron estrategias de regresión para su implementación son las que mejores resultados obtuvieron para los conjuntos de datos. Asimismo, la combinación con técnicas de pre procesamiento y selección de atributos mejoran los resultados de exactitud.

La selección de las técnicas de construcción de modelos de estimación de esfuerzo depende de los conjuntos de datos para los cuales son creadas, por lo que el procedimiento de evaluación podría permitir a los profesionales identificar los modelos que mejor se ajustan a sus proyectos particulares.

Como trabajo futuro, se desea determinar el impacto de las distintas técnicas de pre procesamiento de datos, selección de atributos y algoritmos de aprendizaje en los resultados de exactitud de los modelos a partir de los parámetros de configuración. Asimismo, es de interés replicar este estudio utilizando el repositorio ISBSG R16.

Agradecimientos. Este estudio fue apoyado por la Universidad de Costa Rica No. 834-B8-A27. Nuestro agradecimiento al *International Software Benchmarking Standards Group*.

Referencias

Abran, A. (2015). Software project estimation: the fundamentals for providing high quality information to decision makers. John Wiley & Sons.

Azhar, D., Riddle, P., Mendes, E., Mittas, N., & Angelis, L. (2013). Using ensembles for web effort estimation. In 2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (pp. 173–182). IEEE.

Dejaeger, K., Verbeke, W., Martens, D., & Baesens, B. (2012). Data mining techniques for software effort estimation: a comparative study. IEEE transactions on software engineering, 38(2), 375–397. doi: 10.1109/TSE.2011.55

Dolado, J., Rodriguez, D., Harman, M., Langdon, W., & Sarro, F. (2016). Evaluation of estimation models using Minimum Interval of Equivalence. App Soft Comp, 49, 956–967.

Dumke, R., & Abran, A. (2011). COSMIC Function Points: Theory and Advanced Practices. CRC Press.

Ferrucci, F., Harman, M., & Sarro, F. (2014). Search-based software project management. In Software Project Management in a Changing World (pp. 373–399). Springer, Berlin.

Fingerman, S. (2011). Practical software project estimation; a toolkit for estimating software development effort & duration. Sci-Tech News (Vol. 65). McGraw Hill Professional.

Gencel, C. (2008). How to Use COSMIC Functional Size in Effort Estimation Models?. In Software Process and Product Measurement (pp. 196–207). Springer, Berlin, Heidelberg.

Huang, J., Li, Y. F., & Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. Information and soft. Technology, 67, 108–127.

ISBSG. (2013). The ISBSG Development & Enhancement project data. ISBSG, Ed., R 12.

Jorgensen, M., & Shepperd, M. (2007). A Systematic Review of Software Development Cost Estimation Studies. IEEE Transactions on Software Engineering, 33(1), 33–53.

Keung, J., Kocaguneli, E., & Menzies, T. (2013). Finding conclusion stability for selecting the best effort predictor in software effort estimation. Automated Soft Eng, 20(4), 543–567.

Kitchenham, B., & Mendes, E. (2009). Why comparative effort prediction studies may be invalid. In Proceedings of the 5th international Conference on Predictor Models in Software Engineering (p. 4). ACM. doi: 10.1145/1540438.1540444

Langdon, W. B., Dolado, J., Sarro, F., & Harman, M. (2016). Exact mean absolute error of baseline predictor, MARP0. Information and Software Technology, 73, 16–18.

Lavazza, L., & Morasca, S. (2017, June). On the evaluation of effort estimation models. In Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering (pp. 41–50). ACM.

Liu, Q., Xiao, J., & Zhu, H. (2018). Feature selection for software effort estimation with localized neighborhood mutual information. *Cluster Computing*, 1-9.

Menzies, T., & Shepperd, M. (2012). Special issue on repeatable results in software engineering prediction. doi: 10.1007/s10664-011-9193-5

Menzies, T., Yang, Y., Mathew, G., Boehm, B., & Hihn, J. (2017). Negative results for software effort estimation. Empirical Software Engineering, 22(5), 2658–2683.

Murillo-Morera, J., Quesada-López, C., Castro-Herrera, C., & Jenkins, M. (2017). A genetic algorithm based framework for software effort prediction. Journal of Software Engineering Research and Development, 5(1), 4.

Oliveira, A. L., Braga, P. L., Lima, R. M., & Cornélio, M. L. (2010). GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation. Information and Software Technology, 52(11), 1155–1166.

Quesada López, C., & Jenkins Coronas, M. (2016). Function point structure and applicability: A replicated study. Journal of Object Technology (15).

Saeed, A., Butt, W. H., Kazmi, F., & Arif, M. (2018, February). Survey of Software Development Effort Estimation Techniques. In Proceedings of the 2018 7th International Conference on Software and Computer Applications (pp. 82–86). ACM.

Shepperd, M., & MacDonell, S. (2012). Evaluating prediction systems in software project estimation. Information and Software Technology, 54(8), 820–827.

Song, Q., Jia, Z., Shepperd, M., Ying, S., & Liu, J. (2011). A general software defect-proneness prediction framework. IEEE Transactions on Software Engineering, 37(3), 356–370.

Song, L., Minku, L. L., & Yao, X. (2013). The impact of parameter tuning on software effort estimation using learning machines. In Proceedings of the 9th international conference on predictive models in software engineering (p. 9). ACM.

Wen, J., Li, S., Lin, Z., Hu, Y., & Huang, C. (2012). Systematic literature review of machine learning based software development effort estimation models. Information and Software Technology, 54(1), 41-59. doi: 10.1016/j.infsof.2011.09.002

Whigham, P. A., Owen, C. A., & Macdonell, S. G. (2015). A baseline model for software effort estimation. ACM Transactions on Software Engineering and Methodology, 24(3), 20.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). Experimentation in software engineering. Springer Science & Business Media.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Xia, T., Chen, J., Mathew, G., Shen, X., & Menzies, T. (2018). Why Software Effort Estimation Needs SBSE. arXiv preprint arXiv:1804.00626.