

Comportamiento de Métricas para Proyectos de Explotación de Información en PyMEs

Diego Basso^{1,2}, Darío Rodríguez², Ramón García-Martínez²

1. Departamento de Ingeniería e Investigaciones Tecnológicas.
Universidad Nacional de La Matanza

2. Laboratorio de Investigación y Desarrollo en Ingeniería de Explotación de Información.
Grupo de Investigación en Sistemas de Información. Universidad Nacional de Lanús.
diebasso@yahoo.com.ar, drodrigu@unla.edu.ar, rgm1960@yahoo.com

Resumen. En este trabajo se presenta un estudio del comportamiento de las métricas propuestas para proyectos de explotación de información, la cual considera características y parámetros identificados para proyectos pequeños, de aplicación en las empresas PyMEs. Para ello, se realiza una introducción sobre la categorización definida para las métricas consideradas y el modelo de proceso de desarrollo utilizado como referencia, se delimita el problema presentando el diseño experimental y los resultados obtenidos para finalizar con la puntualización de algunas conclusiones.

Palabras claves: Explotación de Información. Métricas. Comportamiento de Métricas. Ingeniería de Proyectos de Explotación de Información.

1. Introducción

El uso de métricas es una característica importante de todas las disciplinas de ingeniería. Dentro de un marco de trabajo ingenieril, las métricas permiten cuantificar aspectos específicos de un proceso, de un producto o de un proyecto. En este sentido, recolectar métricas constituye el primer paso para saber cómo controlar y mejorar el proceso de desarrollo de software [Fenton y Bieman, 2014].

Los proyectos de explotación de información deben considerar la aplicación de una metodología de desarrollo [Chapman et al., 1999; Pyle, 2003; García-Martínez et al., 2011] que incluya entre sus actividades el registro de métricas, que permitan medir y controlar el avance del proyecto y evaluar su calidad. En esta disciplina de la Ingeniería, los proyectos presentan características diferentes respecto a los proyectos de desarrollo de software tradicional y esa diferencia está en la naturaleza del producto resultante [Marbán, 2003; Marbán et al., 2008]. Esto implica que las etapas habituales y métricas definidas para un proyecto de desarrollo clásico no sean totalmente apropiadas para un proyecto de explotación información [Basso et al., 2013; Pytel et al., 2015]. Dentro de este campo de conocimiento, se han propuesto y desarrollado distintas herramientas, entre las que se incluyen un modelo de proceso [Martins et al., 2014], modelos de estimación de esfuerzo [Pytel et al., 2015], y una propuesta de métricas para proyectos de explotación de información [Basso, 2014].

Esta comunicación presenta el estudio del comportamiento de algunas de las métricas propuestas en [Basso, 2014]. Se indica la categorización definida para las métricas (sección 2); luego se delimita el problema (sección 3), se detallan los materiales y métodos usados en el estudio (sección 4) y se presentan los resultados obtenidos (sección 5). El trabajo finaliza con la exposición de conclusiones preliminares obtenidas (sección 6).

2. Métricas para Proyectos de Explotación de Información

Se establece la clasificación que se indica en la Figura 1, para las métricas propuestas en [Basso et al. 2013; Basso, 2014]. Esta clasificación se formula a partir de: (a) las características y consideraciones presentadas en [Marbán, 2003; Marbán et al., 2008] para el modelo de estimación DMCoMo, (b) los parámetros establecidos en [Pytel, 2014] para Proyectos Pequeños, (c) las indicadas en [Pytel et al., 2015] para el Método de Estimación de Esfuerzo propuesto para PyMEs, y (d) los subprocesos y tareas propuestas por [Martins et al., 2014] para el Modelo de Proceso de Desarrollo para Proyectos de Explotación de Información.

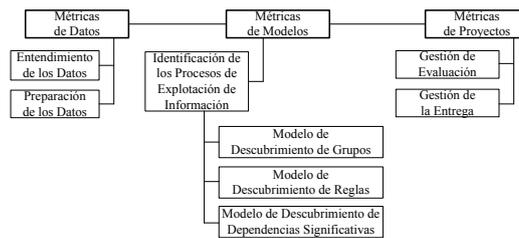


Fig. 1. Clasificación de las Métricas para Proyectos de Explotación de Información

Asimismo, se menciona que las métricas del subproceso de Identificación de los Procesos de Explotación de Información se agrupan según su tarea de descubrimiento y aplicación en un proyecto, contemplando los procesos de explotación de información definidos en [Britos, 2008; García-Martínez et al., 2013], los cuales se basan en la utilización de sistemas inteligentes [García-Martínez et al., 2003].

3. Delimitación del Problema

En [García-Martínez et al., 2011] se señala la necesidad de disponer de métricas asociadas al proceso de desarrollo de un proyecto de explotación de información para PyMEs, que permitiera suministrar información relevante a tiempo y establecer objetivos de mejora en los procesos y productos, con el fin de garantizar la calidad de estos proyectos. En este contexto, el objetivo de este trabajo es analizar el comportamiento de las métricas propuestas de Datos, Modelos y Proyectos, con especial focalización en proyectos de explotación de información de tamaño pequeño, que son los que usualmente requieren las PyMEs [García-Martínez et al., 2011].

4. Materiales y Métodos

Para construir el conocimiento asociado al presente proyecto de investigación, se siguió un enfoque de investigación clásico [Creswell, 2002; Rosas y Riveros, 1985] con énfasis en la producción de tecnologías [Sábato y Mackenzie, 1982], identificando métodos y materiales necesarios para desarrollar el mismo.

El método utilizado para generar los estudios del comportamiento de métricas es el de generación de valores de variables independientes del proyecto por el Método de Monte Carlo [Cohen, 1988].

Los materiales usados para la investigación que se reporta son las métricas que se proponen en [Basso, 2014] y se derivan de los trabajos [Marbán, 2003; Marbán et al., 2008; Martins et al., 2014; Pytel, 2014; Pytel et al., 2015].

El protocolo utilizado para la experimentación fue el siguiente:

Paso 1: Desarrollo de un banco de pruebas donde se generan los datos de distintos proyectos y modelos de explotación de información con los valores de las variables experimentales independientes (métricas básicas) y se aplican las variables experimentales dependientes (métricas derivadas).

Paso 2: Integrar estadísticamente la información obtenida generando los gráficos y tablas auxiliares que se consideren necesarias.

Paso 3: Interpretar los resultados experimentales obtenidos y formular, como conclusión, una regla de comportamiento general de la métrica.

5. Resultados Obtenidos

En esta sección se presentan los resultados obtenidos del estudio experimental de comportamiento de las métricas de Datos (sección 5.1), métricas de Modelos (sección 5.2) y métricas de Proyectos (sección 5.3).

5.1. Métricas de Datos

En esta sección se presentan los resultados experimentales del comportamiento estudiado para las métricas de Densidad de Valores Nulos, Nivel de Compleción de la Tabla, Grado de Corrección de Datos de la Tabla y Grado de Utilidad de Atributos.

Densidad de Valores Nulos y Nivel de Compleción de la Tabla:

Métricas	Referencia
$DVN(T) = \frac{NVN(T)}{NR(T) * NA(T)}$ $NCT(T) = 1 - DVN(T)$	- Nro. de valores nulos o faltantes - NVN (T) - Nro. de registros de la tabla - NR (T) - Nro. de atributos de la tabla - NA (T)

Del estudio del comportamiento de las métricas analizadas se concluyó que: al aumentar el número de valores nulos en las tablas, la densidad de valores nulos tiende a 1 y el nivel de compleción a 0. Lo contrario ocurre al disminuir el número de valores nulos. En base a este análisis y a lo establecido en el trabajo de [Pytel, 2014], se sugiere que el valor de la métrica DVN (T) sea inferior al 15% de los datos,

mientras que la métrica NCT (T) sea superior al 85%. Esto contribuye a considerar datos con una calidad aceptable y facilitar el análisis posterior de los mismos, permitiendo de esta manera una adecuada aplicación de los tipos de modelado para descubrimiento de conocimiento. La misma conclusión es aplicable si se analiza la densidad de valores nulos por atributo o registro de la tabla.

Grado de Corrección de Datos de la Tabla:

Métrica	Referencia
$GCD(T) = 1 - \frac{NVE(T) + NVN(T)}{NR(T) * NA(T)}$	<ul style="list-style-type: none"> - Nro. de valores erróneos - NVE (T) - Nro. de valores nulos o faltantes - NVN (T) - Nro. de registros de la tabla - NR (T) - Nro. de atributos de la tabla - NA (T)

Se consideró que cuando la medida de la métrica toma un valor ALTO (un valor superior a 0.70) significa que los datos de la tabla son útiles para el proyecto, mientras que un valor por debajo, considerado BAJO, indica que deben analizarse los datos con mayor precisión y evaluar la relevancia o representatividad de la tabla para el proyecto.

Del estudio del comportamiento de la métrica analizada se concluyó que: al aumentar el número de valores erróneos y nulos en las tablas por encima del 30% de los datos totales, el grado de corrección de los datos de la tabla GCD (T) tiende a ser BAJO. Al igual que con el número de valores nulos NVN (T), se sugiere que el número de valores erróneos NVE (T) en una tabla sea inferior al 15% de los datos disponibles, para lograr que la misma posea un ALTO grado de corrección (mayoría de datos útiles). La misma conclusión es aplicable por cada atributo o registro de la tabla, permitiendo identificar los atributos y registros con más valores erróneos.

Grado de Utilidad de Atributos:

Métrica	Referencia
$GUA(T) = \frac{NA(T) - (NO_UTILES(T) + 0,5 * NAUD(T))}{NA(T)}$ $NA(T) = NASE(T) + NAUD(T) + NANC(T) + NANS(T)$ $NO_UTILES(T) = NANC(T) + NANS(T)$	<ul style="list-style-type: none"> - Nro. de atributos útiles sin errores – NASE (T) - Nro. de atributos útiles con defectos – NAUD (T) - Nro. de atributos no correctos – NANC (T) - Nro. de atributos no significativos – NANS (T)

Se consideró que cuando la medida de la métrica toma un valor ALTO (un valor superior a 0.70) significa que la mayoría de los atributos son de utilidad y significativos para el proyecto y no se requieren demasiadas correcciones; cuando el valor es MEDIO (un valor entre 0.30 y 0.70) implica que los atributos son aceptablemente usables en el proyecto pero requieren correcciones en los datos, mientras que cuando toma un valor BAJO (un valor inferior a 0.30) significa que los atributos de la tabla no son útiles y representativos para el proyecto, pudiéndose descartar la misma. En la Figura 2 se muestra la proporción por tipo de atributos necesarios en una tabla, para alcanzar los distintos grados de utilidad de la métrica en un proyecto de explotación de información para PyMEs.

La métrica y los parámetros propuestos alcanzan una precisión promedio del 70% en la valoración de la utilidad de una tabla. Del estudio del comportamiento de la métrica analizada, se concluyó que la misma alcanza sus diferentes umbrales de referencia cuando se producen los siguientes comportamientos en las métricas básicas:

- El grado de utilidad de los atributos es ALTO, cuando existe un mínimo de 80% de atributos útiles. Este indicador representa una cantidad de atributos útiles sin

errores, asociado a la métrica NASE (T), mayor o igual al 70% y una cantidad de atributos útiles con defectos, asociado a la métrica NAUD (T), menor al 30%. A su vez, la cantidad de atributos no útiles, asociado a las métricas NANC (T) y NANS (T), no debe superar el 20% del total de atributos. Las conclusiones obtenidas indican que es preciso disponer de muchos atributos útiles y pocos o muy pocos atributos no útiles para que la tabla sea representativa en el proyecto.

- El grado de utilidad de los atributos es MEDIO, cuando la cantidad de atributos útiles es del 40 al 80%. Este resultado implica una cantidad de atributos útiles sin errores, asociado a la métrica NASE (T) entre el 30 y 70%, y una cantidad de atributos útiles con defectos, asociado a la métrica NAUD (T), menor al 50%. A su vez, la cantidad de atributos no útiles, asociado a las métricas NANC (T) y NANS (T), es del 20 al 60% del total de atributos. Las conclusiones obtenidas indican que si la cantidad de atributos útiles en la tabla es de normal a muchos y los atributos no útiles se incrementan en el rango de pocos a bastantes, la utilidad de dichos atributos cae a un grado medio, requiriéndose algunas correcciones en los datos.
- El grado de utilidad de los atributos es BAJO, cuando existe un máximo de 40% de atributos útiles. Este indicador constituye una cantidad de atributos útiles sin errores, asociado a la métrica NASE (T), menor al 30% y una cantidad de atributos útiles con defectos, asociado a la métrica NAUD (T), menor o igual al 20%. A su vez, la cantidad de atributos no útiles, asociado a las métricas NANC (T) y NANS (T), es al menos el 60% del total de atributos. Las conclusiones obtenidas indican que si se tienen muy pocos o una cantidad normal de atributos útiles, pero bastantes o muchos atributos no útiles, la tabla deja de ser relevante para el proyecto.

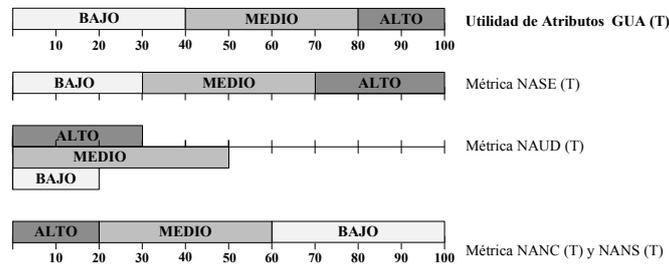


Fig. 2. Proporción de atributos útiles y no útiles según rango en métrica GUA (T)

5.2. Métricas de Modelos

En esta sección se presentan los resultados experimentales del comportamiento estudiado para las métricas de Precisión del Modelo, Cobertura de una Regla y Grado de Incidencia de un Atributo.

Precisión del Modelo:

Métrica	Referencia
$PDM(C_i) = \frac{NCVC_i}{NCVC_i + NCFC_i}$	<ul style="list-style-type: none"> - Nro. de casos pertenecientes a la clase C_i (o grupo) correctamente clasificados por el modelo en esa misma clase - $NCVC_i$. - Nro. de casos incorrectamente clasificados por el modelo como clase C_i y que pertenecen a otras clases - $NCFC_i$.

Se consideró que cuando la medida de la métrica toma un valor ALTO (un valor superior 0.70) significa que el modelo posee un buen nivel de efectividad para clasificar casos en su clase o grupo; cuando el valor es MEDIO (un valor entre 0.50 y 0.70) implica que el modelo puede confundir los casos clasificados a cada clase, convirtiéndose en un modelo inestable y poco eficiente, mientras que cuando toma un valor BAJO (un valor inferior a 0.50) significa que el modelo confunde las clases a las que pertenecen los casos clasificados, resultando un modelo poco efectivo.

La métrica y los parámetros propuestos alcanzan una exactitud promedio del 70% en la valorización del grado de precisión de los modelos en un proyecto. Asimismo, el comportamiento de la métrica comprende a los modelos de explotación de información cuya exactitud de clasificación es superior al 70%. Del estudio del comportamiento de la métrica analizada, se concluyó que alcanza sus diferentes umbrales cuando se producen los siguientes comportamientos en las métricas básicas:

- El grado de precisión de un modelo es ALTO, cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos, asociado a la métrica NVC_i , está entre el 11 y 70% del total de casos (representado del 20 al 70% de casos), el número de casos clasificados incorrectamente en esa misma clase, asociado a la métrica NFC_i , es menor al 10% y la exactitud del modelo para generalizar, predecir y clasificar correctamente nuevos casos es superior al 80%.
- El grado de precisión de un modelo es MEDIO, cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos está entre el 1 y 40% del total de casos (representado del 10 al 40% de casos), el número de casos clasificados incorrectamente en esa misma clase está entre el 1 y 20% (representado del 10 al 20% de casos) y la exactitud del modelo está entre 70 y 90%.
- El grado de precisión de un modelo es BAJO, cuando el número de casos clasificados correctamente en cualquiera de las clases o grupos es menor al 10% del total de casos, el número de casos clasificados incorrectamente en esa misma clase está entre el 11 y 30% (representado del 20 al 30% de casos) y la exactitud del modelo está entre 70 y 80%.

Las conclusiones obtenidas muestran que si la precisión de un modelo de explotación de información es aceptable (con valores ALTOS), el mismo resulta efectivo y puede ser utilizado para predecir y clasificar nuevos casos del mismo dominio del proyecto, de los que desconoce su clase o grupo. En cambio, si la precisión alcanza niveles MEDIOS o BAJOS, aumenta el riesgo que el modelo se vuelva inestable y comience a confundir las clases o grupos durante la clasificación de los casos, convirtiéndose en un modelo poco efectivo y confiable.

Cobertura de una Regla:

Métrica	Referencia
$COBER(R) = \frac{NCCNS(R)}{NCE(C)}$	<ul style="list-style-type: none"> - Nro. de casos que satisfacen la aplicación de la regla de comportamiento o de pertenencia de una clase o grupo - NCCNS (R) - Nro. de casos por clase para entrenar el modelo - NCE (C).
Valor Mín. Umbral: $COBER(R) > \frac{100}{NRGL(C)} [\%]$	<ul style="list-style-type: none"> - Nro. de reglas descubiertas por el modelo por cada clase o grupo - NRGL (C)

Se propuso un valor mínimo de umbral de cobertura de una regla, a partir del número de reglas de pertenencia obtenidas para una clase o grupo. La utilidad de esta métrica

en un proyecto que aplica un modelo de Descubrimiento de Reglas [García-Martínez et al., 2013], está en que aquellas reglas con mayor cobertura son las más representativas y útiles del modelo para definir el comportamiento de una clase o grupo. Por consiguiente, representan las reglas que mayor credibilidad e interés le otorgan al modelo construido para clasificar nuevos casos a una clase o grupo.

Del estudio del comportamiento de la métrica analizada se concluyó que: cuando se incrementa el número de casos a los que se le puede aplicar una regla de pertenencia a una clase o grupo, asociada a la métrica NCCNS (R), respecto del total de casos de esa misma clase, la métrica COBER (R) aumenta su valor. Mientras que, cuando el número de casos a los que se puede aplicar la regla disminuye, también lo hace la métrica de cobertura de la regla. Por otra parte, obtener un bajo número de reglas para una clase o grupo, permite identificar más fácilmente las características representativas del comportamiento de una clase o grupo del modelo de explotación de información. A su vez, provoca un aumento en el valor de umbral de cobertura, la necesidad de una mayor cantidad de casos a cubrir con la regla, dándole credibilidad e interés a ésta para clasificar nuevos casos a una clase o grupo. Si el número de reglas es alto, el valor de umbral para la métrica de cobertura es bajo, menor es la cantidad de casos cubiertos con una regla y más difícil y complejo resulta identificar el comportamiento asociado a una clase o grupo del modelo. En consecuencia, es deseable que la cantidad de reglas que definen el comportamiento de una clase o grupo sea siempre la más baja posible.

Grado de Incidencia de Atributos:

Métrica	Referencia
$\text{GINC}(A_{V_i}) = \frac{\text{NCAT}(A_{V_i})}{\text{NCE}(C)}$	<ul style="list-style-type: none"> - Nro. de casos cubiertos por una clase o grupo del atributo clase, cuando un atributo significativo A toma el valor V_i - NCAT (A_{V_i}) - Nro. de casos por clase para entrenar el modelo - NCE (C).

Del estudio del comportamiento de la métrica analizada se concluyó que: al aumentar el número de casos cubiertos por alguno de los valores V_i de un atributo significativo, para un valor específico del atributo clase, la métrica GINC (A_{V_i}) aumenta para ese valor V_i , mientras que para los demás valores del atributo significativo disminuye. Asimismo, todo valor V_i , asociado a un atributo significativo de un modelo, cuyo grado de incidencia sea máximo en una clase o grupo de un atributo clase, representa las características o factores que más influencia tienen sobre un determinado resultado para el proyecto, según el problema de explotación de información planteado. Cabe mencionar que, cuando esta métrica toma valores muy similares en dos o más valores del atributo significativo, implica que no hay una única característica (en ese mismo atributo) que sea más relevante sobre el resultado del atributo clase.

5.3. Métricas de Proyectos

En esta sección se presentan los resultados experimentales del estudio del comportamiento realizado para la métrica de Éxito de Resultado del Proceso de Desarrollo de un proyecto de explotación de información.

Éxito de Resultado del Proceso de Desarrollo:

Métrica	Referencia
$EPD(P) = \frac{\sum NMI_i * Peso(i)}{NMOD}$	- Nro. de modelos de explotación de información cuyo interés <i>i</i> y utilidad de resultados es Alto, Medio o Bajo - NMI_i - Valor de ponderación asignado a cada modelo de acuerdo al interés de los resultados obtenidos - $Peso(i)$ - Nro. de modelos construidos para el proyecto - $NMOD$

Se consideró que cuando la medida de la métrica toma un valor ALTO (un valor superior a 0.80) significa que el proceso de desarrollo aplicado es correcto y los resultados alcanzados cumplen con los criterios de éxito del proyecto; cuando el valor es MEDIO (un valor entre 0.61 y 0.80) implica que los resultados cumplen parcialmente con los criterios de éxito y las tareas del proceso deben ser revisadas y ajustadas, mientras que cuando toma un valor BAJO (un valor inferior a 0.60) significa que los resultados obtenidos no responden a los criterios de éxito del proyecto. Por otra parte, por cada modelo se estableció un criterio de valoración (muy bajo, bajo, medio y alto) y un valor de ponderación (de 0 a 1), por nivel de satisfacción de resultados e interés del conocimiento descubierto para el proyecto. En la Figura 3 se muestra el número estimado de modelos necesarios, junto al grado de interés esperado de resultados, para alcanzar los diferentes niveles de éxito de la métrica en un proyecto de explotación de información para PyMEs.

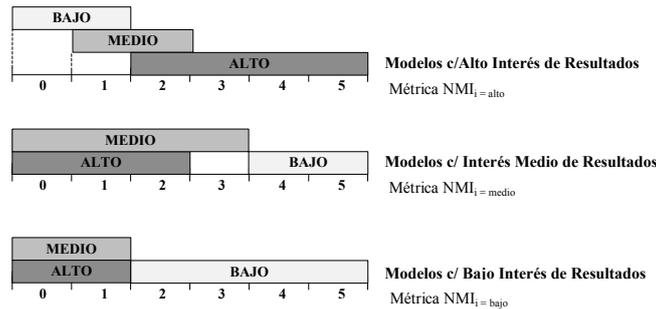


Fig. 3. Número de modelos por nivel de interés de resultados s/rango de la métrica EPD (P)

La métrica y los parámetros propuestos alcanzan una exactitud promedio del 70% en la valorización el éxito del proceso de desarrollo en un proyecto. Del estudio del comportamiento de la métrica analizada alcanza sus diferentes umbrales cuando se producen los siguientes comportamientos en las métricas básicas:

- El éxito de resultados del proceso de desarrollo es ALTO, cuando el número de modelos que presentan un alto interés para el usuario, por el conocimiento descubierto y su utilidad, asociado a la métrica NMI_i , es mayor o igual a 2. A su vez, el número de modelos que presentan un interés medio es menor o igual a 2, mientras que el número de modelos con bajo interés es a lo sumo 1. Estas conclusiones indican que el proceso de desarrollo aplicado es correcto y se cumplen satisfactoriamente los criterios de éxito definidos para el proyecto. Cuanto mayor sea el número de modelos con este nivel de interés, más se acerca la medida de la métrica a su valor óptimo.
- El éxito de resultados del proceso de desarrollo es MEDIO, cuando el número de modelos que presentan un alto interés para el usuario, asociado a la métrica NMI_i ,

es 1 ó 2. A su vez, el número de modelos que presentan un interés medio es menor o igual a 3, mientras que el número de modelos con bajo interés para el usuario es a lo sumo 1. Estas conclusiones significan que se cumplen parcialmente con los criterios de éxito definidos para el proyecto. Esto se debe a que si bien se obtienen patrones de conocimiento interesante de los modelos, algunos expresan resultados que no son totalmente útiles como se presentan. En consecuencia, se necesita revisar el proceso de explotación de información utilizado y ajustar algunas tareas del proceso de desarrollo.

- El éxito de resultados del proceso de desarrollo es BAJO, cuando el número de modelos que presentan un alto interés para el usuario, asociado a la métrica NMI, es a lo sumo 1. Mientras que el número de modelos con bajo interés es mayor o igual a 2. Estas conclusiones indican que no se alcanzaron los criterios de éxito. En consecuencia, resulta necesario revisar las tareas del proceso de desarrollo, especialmente los subprocesos de: (a) entendimiento de los datos, (b) preparación de los datos y (c) identificación de los procesos de explotación de información.

6. Conclusiones

Se definieron tres categorías generales de métricas aplicables a proyectos de explotación de información: *Métricas de Datos*, *Métricas de Modelos* y *Métricas de Proyectos*. En cada una de estas categorías se incluyeron algunas de las métricas propuestas, que cubren el Modelo de Proceso de Desarrollo con las tareas definidas en [Martins et al., 2014] para este tipo de proyectos. Además, se estudió de manera analítica, mediante simulación por Monte Carlo, el comportamiento de las métricas, con particular énfasis en su utilización en empresas PyMEs y proyectos pequeños.

Las conclusiones obtenidas para algunas métricas se basaron en valores numéricos asignados a los términos lingüísticos “Alto”, “Medio” y “Bajo”, a partir de la propia experiencia. En este contexto surge como futura línea de trabajo estudiar si ajustando estos valores, se obtiene un comportamiento más estable de las métricas cuando caen en estos rangos. Por otra parte, con algunas métricas se observaron comportamientos un poco imprecisos, ya que dependían de los valores utilizados en la simulación.

Se prevé validar las métricas propuestas de manera empírica, en el marco de los proyectos de explotación de información recopilados en [Pytel, 2014]. A su vez, se estima validar y ajustar los parámetros de las métricas de manera empírica, utilizando proyectos de explotación de información reales, a fin de evaluar su viabilidad.

7. Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Proyecto de Investigación 33A205 de la Secretaria de Ciencia y Técnica de la Universidad Nacional de Lanús (Argentina).

8. Referencias

- Basso, D., 2014. *Propuesta de Métricas para Proyectos de Explotación de Información*. Revista Latinoamericana de Ingeniería de Software, 2(4): 157-218, ISSN 2314-2642.
- Basso, D., Rodríguez, D., García-Martínez, R., 2013. *Propuesta de Métricas para Proyectos de Explotación de Información*. Workshop de Bases de Datos y Minería de Datos. Proceedings XIX Congreso Argentino de Ciencias de la Computación. Pag. 983-992. ISBN 978-987-23963-1-2.
- Britos, P., 2008. *Procesos de Explotación de Información basados en Sistemas Inteligentes*. Tesis Doctoral. Universidad Nacional de La Plata. Facultad de Informática. Argentina.
- Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. 1999, *CRISP-DM 1.0 Step by step BGuide*. Edited by SPSS.
- Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2^{da} edición). Lawrence Earlbaum Associates Publishers. ISBN 0-8058-0283-5.
- Creswell, J. 2002. *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall.
- Fenton, N., & Bieman, J. 2014. *Software metrics: a rigorous and practical approach*. CRC Press.
- García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo-Cattaneo, F., Rodríguez, D., Pytel, P., Vanrell, J., 2011. *Towards an Information Mining Engineering*. En *Software Engineering, Methods, Modeling and Teaching*. Sello Editorial Universidad de Medellín. ISBN 978-958-8692-32-6. Páginas 83-99.
- García-Martínez, R., Britos, P., Rodríguez, D. 2013. *Information Mining Processes Based on Intelligent Systems*. Lecture Notes on Artificial Intelligence, 7906: 402-410. ISBN 978-3-642-38576-6.
- García-Martínez, R., Servente, M., Pasquini, D., 2003. *Sistemas Inteligentes*. Editorial Nueva Librería. Buenos Aires.
- Marbán, O., 2003. *Modelo Matemático Paramétrico de Estimación para Proyectos de Data Mining (DMCOMO)*. Tesis Doctoral. Departamento de Lenguajes y Sistemas e Ingeniería Software. Facultad de Informática. Universidad Politécnica de Madrid (UPM). España.
- Marbán, O., Menesalvas E., Fernández-Baizán, C., 2008. *A cost model to estimate the effort of datamining projects (DMCoMo)*. Elsevier. Science Direct. Information System. Volume 33, Issue 1 (March 2008). Pp. 133–150.
- Martins, S., Pesado, P., García-Martínez, R. 2014. *Propuesta de Modelo de Procesos para una Ingeniería de Explotación de Información: MoProPEI*. Revista Latinoamericana de Ingeniería de Software, 2(5): 313-332, ISSN 2314-2642.
- Pyle, D. 2003. *Business Modeling and Business intelligence*. Morgan Kaufmann Publishers.
- Pytel, P. 2014. *Viabilidad y Estimación de Proyectos de Explotación de Información*. Tesis Doctoral en Ciencias informáticas. Facultad de Informática. Universidad Nacional de La Plata.
- Pytel, P., Hossian, A., Britos, P., García-Martínez, R. 2015. *Feasibility and Effort Estimation Models for Medium and Small Size Information Mining Projects*. Information Systems Journal, 47: 01-14. Elsevier. ISSN 0306-4379.
- Rosas, L., Riveros, H. 1985. *Iniciación al método científico experimental*. Trillas.