Una herramienta para la evaluación y comparación de metodologías de minería de datos

Juan Miguel Moine[1], Ana Silvia Haedo[2]

1. Universidad Tecnológica Nacional, Facultad Regional Rosario
2. Facultad de Ciencias Exactas, Universidad Nacional de Buenos Aires

[1]juanmiguelmoine@gmail.com, [2]ahaedo@dc.uba.ar

Resumen. Para llevar a cabo en forma sistemática el proceso de descubrimiento de conocimiento en bases de datos, conocido como minería de datos, es necesaria la implementación de una metodología. Actualmente las metodologías en éste área se encuentran en etapas tempranas de madurez, aunque algunas como CRISP-DM ya están siendo utilizadas exitosamente por los equipos de trabajo para la gestión de sus proyectos. En el presente trabajo se propone un marco comparativo como herramienta para la evaluación y confrontación de metodologías de minería de datos. A partir del mismo se han comparado las metodologías más difundidas en la actualidad.

Palabras clave: Minería de Datos, Gestión de Proyectos, Knowledge Discovery in Databases, Explotación de Información, Metodologías de Minería de Datos, CRISP-DM, Catalyst, P3TQ, KDD.

1 Introducción

Los esfuerzos en el área de la minería de datos o explotación de información se han centrado en su gran mayoría en la investigación de técnicas para la explotación de información y extracción de patrones (tales como árboles de decisión, análisis de conglomerados y reglas de asociación). Sin embargo, se ha profundizado en menor medida el hecho de cómo ejecutar este proceso hasta obtener el "nuevo conocimiento", es decir, en las metodologías.

Una metodología permite llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Ayuda a las organizaciones a entender el proceso de descubrimiento de conocimiento y provee una guía para la planificación y ejecución de los proyectos. Define además de las fases del proceso, las tareas específicas que deben realizarse y cómo llevarlas a cabo.

En los inicios del año 1996, KDD (Knowledge Discovery in Databases) [1] constituyó el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de explotación de información.

A partir del año 2000, con el gran crecimiento en el área de la minería de datos, surgen tres nuevos modelos que plantean un enfoque sistemático para llevar a cabo el proceso: SEMMA [2], CRISP-DM [3] y Catalyst (conocida como P3TQ) [4].

SEMMA fue desarrollada por el SAS Institute y su nombre es el acrónimo correspondiente a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Evaluación).

CRISP–DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implementación. Cada fase se descompone en varias tareas generales de segundo nivel.

La metodología Catalyst, conocida como P3TQ (Product, Place, Price, Time, Quantity), fue propuesta por Dorian Pyle en el año 2003. Esta metodología plantea la formulación de dos modelos: el Modelo de Negocio y el Modelo de Explotación de Información. El Modelo de Negocio, proporciona una guía de pasos para identificar un problema (o la oportunidad del mismo) y los requerimientos reales de la organización. El Modelo de Explotación de Información, proporciona una guía de pasos para la construcción y ejecución de modelos de minería de datos a partir del Modelo de Negocio.

Según [5] KDD y SEMMA no llegan a ser una metodología propiamente dicha, ya que dejan a criterio del equipo de trabajo la definición de las actividades específicas a realizar en cada etapa del proyecto. En cambio, CRISP-DM y Catalyst podrían ser considerados una metodología, porque además de especificar detalladamente las tareas en cada fase proporcionan guías sobre cómo ejecutarlas.

En la actualidad son escasos los estudios que comparan las metodologías mencionadas, los cuales están enfocados en establecer un paralelismo entre las fases del proceso [6, 7, 8] y no un análisis comparativo confrontando las características de cada una.

A medida que la minería de datos madure con el tiempo, irán surgiendo nuevas metodologías y cada vez serán más las alternativas que estarán a disposición del equipo de trabajo. Surge entonces la necesidad de contar con una herramienta que permita evaluar y confrontar diferentes metodologías para esta emergente disciplina, es decir, un marco comparativo.

Objetivos. El objetivo principal de este trabajo es la construcción de un marco comparativo que sirva como herramienta para poder evaluar y confrontar diferentes metodologías de minería de datos. Como objetivo secundario se pretende utilizar esta herramienta para evaluar y comprar CRISP-DM y Catalyst, las cuales constituyen las dos metodologías más difundidas en la actualidad para la gestión de proyectos de minería de datos.

2 Metodología

El marco comparativo que se propone en este trabajo está formado por cuatro aspectos (Fig. 1):

- Nivel de detalle en la descripción de las actividades de cada fase
- Escenarios de aplicación
- Actividades específicas de cada fase
- Actividades destinadas a la dirección del proyecto.

Fig. 1. Aspectos del marco comparativo

Para cada uno de estos aspectos se propone la evaluación de una serie de características, las cuales deberían estar presentes en una metodología de minería de datos bien definida.

Entre los cuatro aspectos se evalúan un total de 52 características, que se describen y explican en forma completa en [9].

Aspecto 1: Nivel de detalle en la descripción de las actividades de cada fase

Una metodología está formada por un conjunto de actividades, las cuales por lo general son agrupadas en un mayor nivel de abstracción denominado fase. La cantidad de niveles de abstracción puede variar entre una metodología y otra. Lo importante es que para cada fase, se definan un conjunto de actividades específicas que secuencien el trabajo a realizar. En una metodología completa, se espera que no sólo se describan las actividades, sino también se especifique la forma en la que deben llevarse a cabo. El resultado tangible de una actividad se denomina entregable.

Este aspecto del marco comparativo está compuesto por 5 características [9] enumeradas en la Tabla 1, y su objetivo es evaluar el nivel de detalle con el que una metodología define las actividades que conforman al proceso.

Tabla 1. Características del primer aspecto del marco comparativo

Características a evaluar
• ¿Se definen actividades específicas para cada fase del proceso? • ¿Se explicitan los pasos a seguir para llevar a cabo cada actividad? • ¿Se definen las entradas de cada actividad? • ¿Se definen las salidas o entregables de cada actividad? • ¿Se provee una guía de buenas prácticas para cada una de las actividades específicas?

Aspecto 2: Escenarios de aplicación.

Los proyectos de explotación de información pueden ser llevados a cabo en distintos escenarios. En algunos casos el usuario desea obtener nuevo conocimiento para abordar algún problema/situación, y en otros se encuentra interesado en explorar sus datos transaccionales en busca de relaciones o patrones que puedan serle útiles. Sin

embargo, en este último caso también existe una situación de trasfondo que motiva el proyecto, ya que difícilmente una organización lo financie si no se establece los beneficios que producirá.

Este aspecto se evalúa en función de 4 características [9] enumerados en la Tabla 2.

Tabla 2. Características del segundo aspecto del marco comparativo

Características a evaluar
• ¿Se especifican actividades para la definición y el análisis del problema u oportunidad con el cual colaborará la minería de datos? • ¿Se consideran puntos de partida alternativos donde el usuario no refiere un problema sino que sólo desea explorar sus datos? • ¿La metodología es independiente del dominio de aplicación? • ¿La metodología es aplicable a proyectos de diferente tamaño?

Aspecto 3: Actividades específicas que componen cada fase.

En este aspecto se pretende analizar la incorporación de ciertas actividades relevantes que deberían estar presentes a lo largo del proceso de minería de datos. Para ello, se propone la evaluación de una serie de características en función de las distintas fases generales que componen al proceso: análisis del problema, selección y preparación de los datos, modelado, implementación y evaluación.

Este aspecto se evalúa en función de 26 características [9], las cuales se agrupan según las diferentes fases del proceso de extracción de conocimiento. Algunas de ellas se mencionan en la Tabla 3.

Tabla 3. Características del tercer aspecto del marco comparativo

Fase	Características a evaluar
Fase de análisis del problema	• ¿Se propone una evaluación general de la organización? • ¿Se identifica al personal involucrado en el proyecto (stackeholders)? • ¿Se especifica de qué forma el usuario utilizará el nuevo conocimiento?
Fase Selección y Preparación de los datos	• ¿Se propone un análisis exploratorio inicial de los datos? • ¿Se contemplan actividades para la transformación de variables y la creación de atributos derivados? • ¿Se verifica con el usuario la completitud del conjunto de datos final?
Fase de Modelado	• ¿Se planifica de qué forma se evaluarán los resultados? • ¿Se proveen directivas para el caso donde se dificulta el descubrimiento de los patrones?
Fase de Evaluación	• ¿Se interpretan los modelos en función de los objetivos organizacionales? • ¿Se comparan y ponderan los modelos obtenidos? • ¿Se propone una revisión general del proceso?
Fase de Implementación	• ¿Se planifica la implementación del nuevo conocimiento? • ¿Se propone la creación de un programa de mantenimiento? • ¿Se documenta la experiencia adquirida por el equipo de trabajo?

Aspecto 4: Actividades destinadas a la dirección del proyecto.

Las actividades de dirección del proyecto consisten en la planificación, ejecución y control de ciertos aspectos importantes para que el mismo se desarrolle exitosamente. Entre estos aspectos se encuentran, por ejemplo, la administración del costo (presupuesto) y la del tiempo del proyecto (cronograma).

Este tipo de actividades se clasifica en dos grupos: actividades de planificación y actividades de control. Las actividades de planificación incluyen la identificación de las tareas a realizar en el proyecto, estimación de la duración de las mismas, estimación de los recursos afectados y la definición del curso de acción. Las actividades de control tienen por objetivo el monitoreo del estado actual del proyecto para su comparación con lo planificado.

Existen ciertos estándares reconocidos internacionalmente, como el PMBOK [10], que establecen cuáles son las áreas que deben administrarse para lograr una adecuada gestión del proyecto, independientemente de su tipo.

Tomando como referencia el PMBOK, en este aspecto del marco comparativo se definen cinco áreas de dirección del proyecto, para cada una de las cuales se evaluarán características relacionadas a la planificación y control de las mismas. Las áreas definidas son: gestión del alcance, gestión del tiempo, gestión del costo, gestión del equipo de trabajo y gestión del riesgo.

Este aspecto se evalúa en función de 17 características [9], las cuales se agrupan según el área de dirección del proyecto. Algunas de ellas se enumeran en la Tabla 4.

Tabla 4. Características del cuarto aspecto del marco comparativo

Área	Características a evaluar
Gestión del alcance	• ¿Se propone la selección de los entregables que se generarán durante el proyecto? • ¿Se especifican actividades de control del alcance?
Gestión del tiempo	• ¿Se construye un cronograma para el proyecto? • ¿Existen actividades de control del cronograma?
Gestión del costo	• ¿Se construye un presupuesto de costos? • ¿Existen actividades de control del presupuesto a medida que avanza el proyecto?
Gestión del equipo de trabajo	• ¿Se efectúa una planificación de los recursos humanos? • ¿Se proponen actividades para motivar la interacción entre los miembros del equipo? • ¿Se efectúa un seguimiento del rendimiento de los recursos humanos?
Gestión del riesgo	• ¿Se efectúa una identificación de los riesgos del proyecto? • ¿Se realiza una cuantificación y priorización de los riesgos? • ¿Existen actividades de supervisión y control de los riesgos?

Utilización del marco comparativo

Como se ejemplifica en la Tabla 5, las características podrán ser evaluadas positiva o negativamente dependiendo si las metodologías en estudio cumplen o no con las mismas, obteniendo como resultado final el porcentaje de valoraciones positivas de cada enfoque.

Tabla 5. Ejemplo de comparación de 3 metodologías.

Característica	Metodología 1	Metodología 2	Metodología 3
Nivel de detalle en actividades de cada fase			
Característica 1.1	NO	NO	SI
Característica …	SI	NO	SI
Escenarios y puntos de partida del proyecto	..	..	..
…	..	..	..
Total de características cumplidas	40 de 52 (77%)	35 de 52 (67%)	51 de 52 (98%)

No se ha realizado una ponderación de cada ítem debido a la subjetividad que podría implicar dicha tarea, quedando esta decisión a criterio del usuario de este marco comparativo. Si se utilizara un sistema de puntajes los resultados probablemente sean diferentes (ya que la presencia de cada característica tendrá su propio peso).

3 Resultados

El marco comparativo se utilizó para evaluar y confrontar a CRISP-DM y a Catalyst [9]. Previamente, para conocer en profundidad ambas metodologías, se aplicaron a un caso de estudio de un centro médico [9] donde se abordó el problema de ausentismo en atenciones médicas programadas (turnos).

La Tabla 6 resume los resultados obtenidos, donde para cada aspecto se han contabilizado las características positivas sobre el total de características evaluadas.

Tabla 6. Evaluación final de todos los aspectos del marco comparativo

Aspecto	CRISP-DM	Catalyst
Nivel de detalle en la descripción de las actividades.	4/5 (80%)	4/5 (80%)
Escenarios de aplicación	3/4 (75%)	4/4 (100%)
Actividades específicas de cada fase	20/26 (77%)	21/26 (81%)
Actividades de dirección del proyecto	8/17 (47%)	8/17 (47%)
Total de características cumplidas	**35/52 (67%)**	**37/52 (71%)**

La evaluación final demuestra que se ha obtenido un resultado muy parejo. La metodología Catalyst logró cumplir con 37 de las 52 características que componen el marco (es decir, el 71%). CRISP-DM ha obtenido un resultado levemente inferior cumpliendo el 67% de los puntos evaluados.

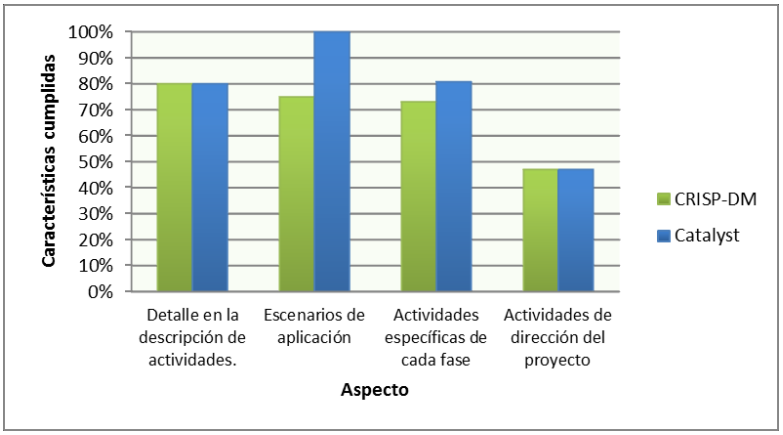

Fig. 2. Porcentaje de características presentes en cada metodología

Como se puede observar en la Figura 2, ambas metodologías resultaron con un puntaje similar en la mayoría de los aspectos, salvo en los escenarios de aplicación donde Catalyst ha logrado cumplir el 100% de las características evaluadas. Se puede apreciar también que ninguna de las dos metodologías tuvo un buen desempeño en las actividades de dirección del proyecto, ya que ambas han obtenido un puntaje inferior al 50%, evidenciando la falta de madurez en este aspecto.

A partir de este análisis, se observa que CRISP-DM y Catalyst reúnen un buen porcentaje de las características propuestas en este marco comparativo, aunque ambos enfoques deberían complementarse con actividades destinadas a la dirección del proyecto.

4 Conclusión

En este trabajo se ha propuesto un marco comparativo como herramienta para la confrontación de metodologías de minería de datos.

El marco presentado incluye cuatro aspectos donde se analiza el nivel de especificación de las tareas, los escenarios de aplicación, las actividades que componen cada fase del proceso y la incorporación de actividades para la dirección del proyecto. Para cada uno de estos aspectos se propone la evaluación de un conjunto de características que deberían estar presentes en una metodología de minería de datos bien definida.

La herramienta se ha utilizado para confrontar las metodologías CRISP-DM y Catalyst. Ambos enfoques han logrado cumplir un gran porcentaje de las

características evaluadas, aunque durante el estudio también se han evidenciado los puntos que se deberían mejorar y seguir desarrollando, como las actividades destinadas a la dirección del proyecto.

Referencias

1. Fayyad, U. y otros: The KDD process for extracting useful knowledge from volumes of data. ACM vol. 39 (1996)
2. SAS Institute: Data Mining and the Case for Sampling. http://nas.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf. (1998) Recuperado el 5 de mayo de 2015
3. Chapman, P., Clinton, J., Kerber, y otros: CRISP-DM 1.0 Step-by-step data mining guide (2000)
4. Pyle, D.: Business Modeling and Data Mining. Morgan Kaufmann Publishers (2003)
5. Moine, J. M., Gordillo, S., Haedo, A.: Análisis comparativo de metodologías para la gestión de proyectos de Minería de Datos. CACIC 2011, VIII Workshop Bases de Datos y Minería de Datos (2011)
6. Mariscal, G. y otros: A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, Vol. 25:2, 137–166 (2010)
7. Kurgan, L. A., Musilek, P.: A survey of Knowledge Discovery and Data Mining process models. Knowledge Engineering Review, 21(1), 1-24 (2006)
8. Azevedo, A.: KDD, SEMMA AND CRISP-DM a parallel overview. AIDIS (2008)
9. Moine, J. M., Gordillo, S., Haedo, A.: Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo. Tesis de Magíster. Universidad Nacional de La Plata, Argentina (2013)
10. Project Management Institute: Guía de los fundamentos de la dirección de proyectos (Guía del PMBOK®). Tercera edición (2005)