METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE DATA MINING

Rodríguez Montequín, Mª Teresa; Álvarez Cabal, J. Valeriano; Mesa Fernández, José Manuel; González Valdés, Adolfo

Resumen

La gran cantidad de datos almacenados actualmente en las organizaciones, unido al gran desarrollo tecnológico de las computadoras, ha supuesto la aparición de nuevas posibilidades, agrupadas bajo el término generalmente conocido como "data mining". El aprovechamiento de estos datos requiere el desarrollo de proyectos con características específicas.

Los proyectos de *Data Mining* tienen por objetivo extraer información útil a partir de grandes cantidades de datos y se aplican a todos los sectores y en todos los campos. Así existen proyectos de este tipo en sectores tan dispares como el comercio electrónico, la banca, las empresas industriales o la exploración petrolífera. La extracción de esta información útil es un proceso complejo, que requiere la aplicación de una metodología estructurada para la utilización ordenada y eficiente de las técnicas y herramientas disponibles.

En este artículo se presentan las principales metodologías utilizadas por los analistas para la realización de proyectos de *Data Mining*: CRISP-DM y SEMMA. Estas metodologías comparten la misma esencia estructurando el proyecto de *Data Mining* en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de *Data Mining* en un proceso iterativo e interactivo.

La presentación de las diferentes fases y tareas de cada metodología proporciona una idea más amplia respecto a la realización de proyectos de *Data Mining*, que facilitará la adaptación de las metodologías, al desarrollo de los proyectos de *Data Mining* específicos de cada organización. Así mismo, la presentación de las fortalezas y debilidades de cada una de las metodologías hace posible la selección informada de una técnica de desarrollo apropiada para cada caso.

Palabras clave: Modelado de datos, Data Mining, Descubrimiento del conocimiento

Abstract

The big amount of data stored by organizations and the technological development of the computer has motivated the apparition of new possibility in the data modelling known as "data mining".

The aim of data mining projects is to extract useful information among a big amount of data. The impact of data mining projects has reached nearly all sectors and fields, existing data mining projects in sectors as unlike as e-bussiness, finance sector, industrial sector or oil-bearing prospecting. The extraction of useful information of data is a complex task requiring the project development with specific characteristics and the use of an organised and structured methodology.

In this paper, they are revised the two main methodologies used by analysts for the development of data mining projects: CRISP-DM and SEMMA. Both methodologies share the same essence structuring the data mining project in phases, which are interconnected themselves, turning into the data mining process in an iterative and interactive process.

The exposition of the different phases and tasks of each methodology provides a more general vision of the development of data mining projects, helping in the adaptation of the methodologies to specific data mining projects. Beside the exposition of the strength and weakness of each methodology allows a reported selecting of a development technique proper in each case.

Key words: Data modelling, Data Mining, KDD

Correspondencia
Universidad de Oviedo
Área de Matemática Aplicada
Independencia, 13
33004 Oviedo
Telef: 985 10 4272
Fax: 985 10 4256
mayte@api.uniovi.es

METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE DATA MINING

1. INTRODUCCIÓN.

El gran desarrollo tecnológico de las computadoras en las últimas décadas ha potenciado el almacenamiento de grandes cantidades de datos y ha permitido el desarrollo de herramientas para su tratamiento, dando lugar a una nueva disciplina conocida como "data mining".

Se puede definir *Data Mining* como el conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos [Piatetski-Shapiro 1991]

Los orígenes del *Data Mining* se pueden establecer a principios de la década de 1980, cuando la administración de hacienda estadounidense desarrolló un programa de investigación para detectar fraudes en la declaración y evasión de impuestos, mediante lógica difusa, redes neuronales y técnicas de reconocimiento de patrones. Sin embargo, la gran expansión del *Data Mining* no se produce hasta la década de 1990 originada principalmente por tres factores:

- Incremento de la potencia de los ordenadores
- Incremento del ritmo de adquisición de datos. El crecimiento de la cantidad de datos almacenados se ve favorecido no sólo por el abaratamiento de los discos y sistemas de almacenamiento masivo, sino también por la automatización de muchos trabajos y técnicas de recogida de datos.
- Aparición de nuevos métodos de técnicas de aprendizaje y almacenamiento de datos.

Desafortunadamente esta expansión implica el desarrollo de proyectos cada vez más grandes en un sector en el que difícilmente se pueden extraer conclusiones a priori y en el que la selección de la mejor técnica no se puede hacer en las primeras fases sino que se precisa un modelo evolutivo, similar al modelo espiral del ciclo de vida de desarrollo software.

Por otra parte el hecho de que más del 75% del esfuerzo se produzca en las primeras fases (en este caso en el pretratamiento de datos) provoca que este tipo de proyectos sea en general subestimado en cuanto a coste y tiempo y que las desviaciones producidas excedan con mucho el 90%.

Ante la necesidad existente en el mercado de una aproximación sistemática para la realización de los proyectos de *Data Mining*, diversas empresas y consultorías han especificado un proceso de modelado diseñado para guiar al usuario a través de una sucesión de pasos que le dirijan a obtener buenos resultados.

Así SAS propone la utilización de la metodología SEMMA (Sample, Explore, Modify, Model, Assess). En 1999 un importante consorcio de empresas europeas, NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra) y OHRA (Holanda), unieron sus recursos para el desarrollo de la metodología de libre distribución CRISP-DM (Cross-Industry Standard Process for *Data Mining*). Esta metodología, junto con la metodología SEMMA, son las dos principales metodologías utilizadas por los analistas en los proyectos de *Data Mining* (Figura 1)

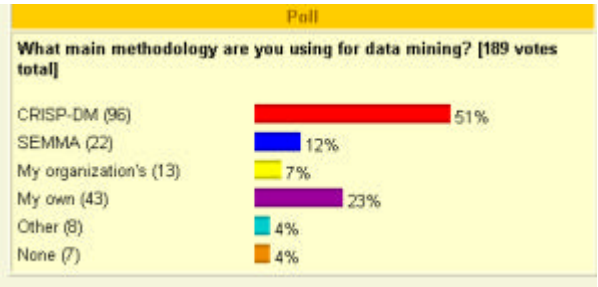

Figura 1: Resultados de la encuesta realizada en http://www.kdnuggets.com

2. METODOLOGÍA SEMMA.

SAS Institute desarrollador de esta metodología, la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos.
El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso (Figura 2)

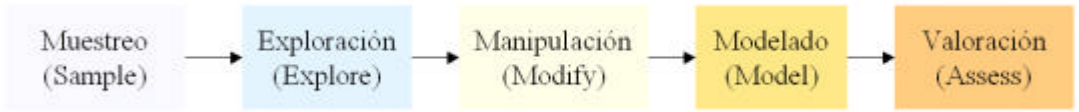

Figura 2: Fases de la metodología SEMMA

El proceso se inicia con la extracción de la población muestral sobre la que se va a aplicar el análisis. El objetivo de esta fase consiste en seleccionar una muestra representativa del problema en estudio. La representatividad de la muestra es indispensable ya que de no cumplirse invalida todo el modelo y los resultados dejan de ser admisibles. La forma más común de obtener una muestra es la selección al azar, es decir, cada uno de los individuos de una población tiene la misma posibilidad de ser elegido. Este método de muestreo se denomina muestreo aleatorio simple.
La metodología *SEMMA* establece que para cada muestra considerada para el análisis del proceso se debe asociar el nivel de confianza de la muestra.
Una vez determinada una muestra o conjunto de muestras representativas de la población en estudio, la metodología *SEMMA* indica que se debe proceder a una exploración de la información disponible con el fin de simplificar en lo posible el problema con el fin de optimizar la eficiencia del modelo. Para lograr este objetivo se propone la utilización de herramientas de visualización o de técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables. De esta forma se pretende determinar cuáles son las variables explicativas que van a servir como entradas al modelo.
La tercera fase de la metodología consiste en la manipulación de los datos, en base a la exploración realizada, de forma que se definan y tengan el formato adecuado los datos que serán introducidos en el modelo.
Una vez que se han definido las entradas del modelo, con el formato adecuado para la aplicación de la técnica de modelado, se procede al análisis y modelado de los datos. El objetivo de esta fase consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un

nivel de confianza determinado. Las técnicas utilizadas para el modelado de los datos incluyen métodos estadísticos tradicionales (tales como análisis discriminante, métodos de agrupamiento, y análisis de regresión), así como técnicas basadas en datos tales como redes neuronales, técnicas adaptativas, lógica fuzzy, árboles de decisión, reglas de asociación y computación evolutiva.

Finalmente, la última fase del proceso consiste en la valoración de los resultados mediante el análisis de bondad del modelo o modelos, contrastado con otros métodos estadísticos o con nuevas poblaciones muestrales.

En la Figura 3 se puede ver un esquema de la dinámica general de la metodología.

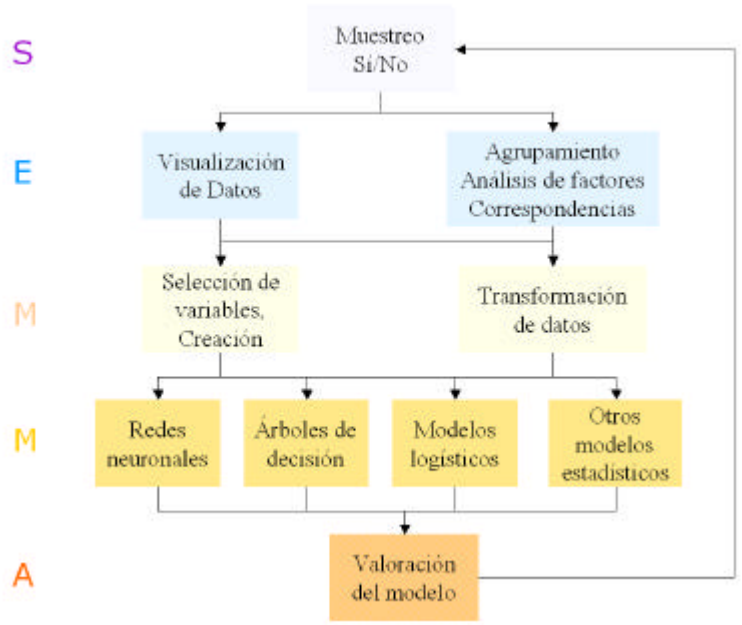

Figura 3: Metodología SEMMA

3. METODOLOGÍA CRISP-DM.

La metodología *CRISP-DM* (Chapman, 1999) consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos (Figura 4)

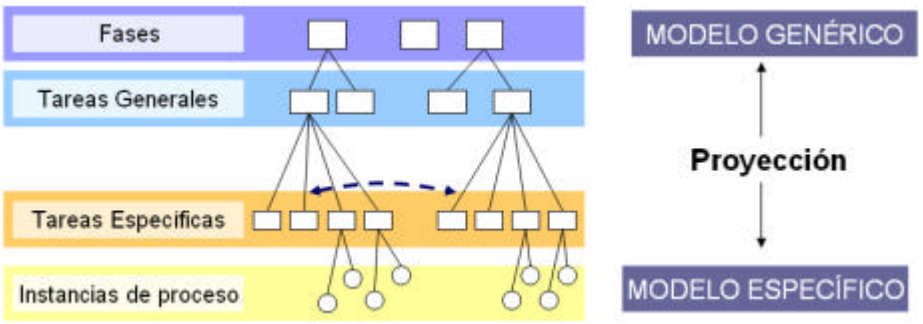

Figura 4: Esquema de los cuatro niveles de abstracción de la metodología CRISP-DM

A nivel más general, el proceso está organizado en seis fases (Figura 5), estando cada fase a su vez estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas. Así, si en el segundo nivel se tiene la tarea general "limpieza de datos", en el tercer nivel se dicen las tareas que tienen que desarrollarse para un caso específico, como por ejemplo, "limpieza de datos numéricos", o "limpieza de datos categóricos".
El cuarto nivel, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de *Data Mining* específico.
La metodología *CRISP-DM* proporciona dos documentos distintos como herramienta de ayuda en el desarrollo del proyecto de *Data Mining*: el modelo de referencia y la guía del usuario.
El documento del modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de *Data Mining* en general. La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de *Data Mining* específicos, proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase.
La metodología *CRISP-DM* estructura el ciclo de vida de un proyecto de *Data Mining* en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto (Figura 5).

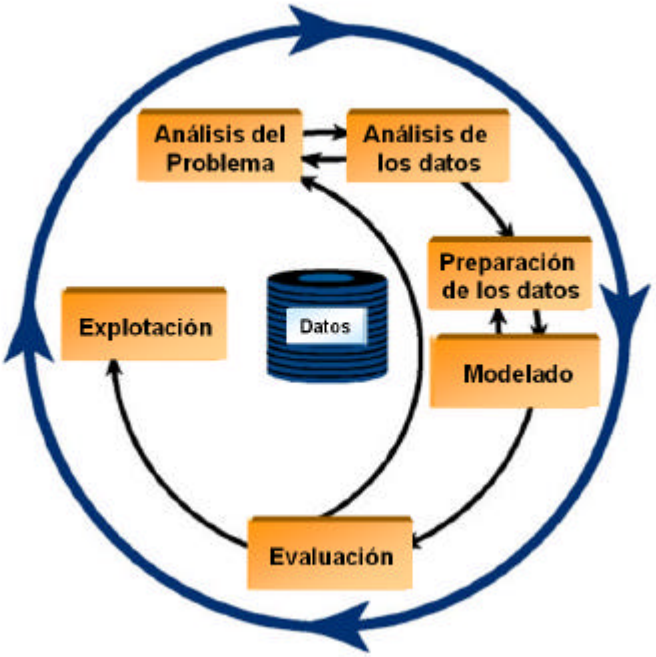

Figura 5: Fases del proceso de modelado metodología CRISP-DM. Las flechas indican relaciones más habituales entre las fases, aunque se pueden establecer relaciones entre cualquier fase. El círculo exterior simboliza la naturaleza cíclica del proceso de modelado.

La primera fase análisis del problema, incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación.
La segunda fase de análisis de datos comprende la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis.

Una vez realizado el análisis de datos, la metodología establece que se proceda a la preparación de los datos, de tal forma que puedan ser tratados por las técnicas de modelado. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.
La fase de preparación de los datos, se encuentra muy relacionada con la fase de modelado, puesto que en función de la técnica de modelado que vaya a ser utilizada los datos necesitan ser procesados en diferentes formas. Por lo tanto las fases de preparación y modelado interactúan de forma sistemática.
En la fase de modelado se seleccionan las técnicas de modelado más apropiadas para el proyecto de *Data Mining* específico. La técnicas a utilizar en esta fase se seleccionan en función de los siguientes criterios:

- Ser apropiada al problema
- Disponer de datos adecuados
- Cumplir los requerimientos del problema
- Tiempo necesario para obtener un modelo
- Conocimiento de la técnica

Antes de proceder al modelado de los datos se debe de establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad de los modelos. Una vez realizadas estas tareas genéricas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos.
En la fase de evaluación, se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo.
Normalmente los proyectos de *Data Mining* no terminan en la implantación del modelo, sino que se deben documentar y presentar los resultados de manera comprensible en orden a lograr un incremento del conocimiento. Además en la fase de explotación se debe de asegurar el mantenimiento de la aplicación y la posible difusión de los resultados [Fayyad, 1996]
En la Tabla 1 se puede ver un esquema de las diferentes fases de la metodología y las tareas generales que se deben de desarrollar en cada fase.

4. COMPARACIÓN DE METODOLOGÍAS.

Las metodologías *SEMMA* y *CRISP-DM* comparten la misma esencia, estructurando el proyecto de *Data Mining* en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de *Data Mining* en un proceso iterativo e interactivo.
La metodología *SEMMA* se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología *CRISP-DM*, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto de *Data Mining* donde la metodología *SEMMA* comienza realizando un muestreo de datos, mientras que la metodología *CRISP-DM* comienza realizando un análisis del problema empresarial para su transformación en un problema técnico (Figura 6). Desde ese punto de vista más global se puede considerar que la metodología *CRISP-DM* está más cercana al concepto real de proyecto, pudiendo

ser integrada con una Metodología de Gestión de Proyectos específica que completaría las tareas administrativas y técnicas.

Análisis del problema	Análisis de los datos	Preparación de los datos	Modelado	Evaluación	Explotación
Determinación de los objetivos empresariales *Conocimiento previo* *Objetivos* *Criterios de éxito* **Evaluación de la situación** *Recursos disponibles* *Requerimientos, supuestos y restricciones* *Terminología* **Determinación de los objetivos técnicos** *Objetivos del modelado* *Criterios de éxito* **Elaboración de la estrategia** *Planificación* *Valoración inicial tec./herramientas*	**Adquisición de datos** *Informe fuentes datos* *Estudio datos disponibles* *Instalación base datos* **Descripción de datos** *Tipo* *Unidades* *Significado* *Procedencia* **Exploración de datos** **Verificación de la calidad de los datos**	**Preprocesado de datos** *Conversión a valores numéricos* *Rellenado de datos* *Identificación de valores no usuales* **Reducción de la dimensionalidad** *Variables* *Muestras* **Transformación datos** *Normalización* *Transformaciones matemáticas* *Discretización*	**Selección de la técnica de modelado** *Técnicas de modelado* *Supuestos de la técnica de modelado* **Diseño del método de evaluación** *Medidas de error* **Generación del modelo** *Parámetros del modelo* *Modelos* *Descripción del modelo* **Evaluación del modelo** *Verificación de los resultados* *Obtención de más información*	**Evaluación de los resultados** *Valoración de los resultados* *Modelos válidos* **Revisión del proceso** *Detección de errores en el proceso de modelización* **Determinación de las siguientes acciones** *Lista de las posibilidades* *Decisión*	**Planificación de la explotación** *Plan de utilización* **Planificación de la monitorización y mantenimiento** *Plan de monitorización y mantenimiento* **Revisión del proyecto** *Extracción de conclusiones*

Tabla 1: Esquema de las tareas generales y las salidas (cursiva) de las seis fases de la metodología *CRISP-DM*

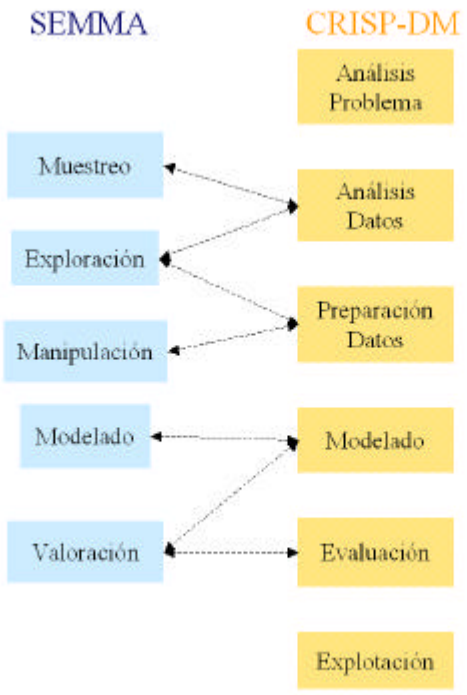

Figura 6: Comparativa de las interrelaciones entre las fases de las metodologías SEMMA y CRISP-DM

Otra diferencia significativa entre la metodología *SEMMA* y la metodología *CRISP-DM* radica en su relación con herramientas comerciales. La metodología *SEMMA* sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se

encuentra implementada. Por su parte la metodología *CRISP-DM* ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo del proyecto de *Data Mining* siendo su distribución libre y gratuita.

5. CONCLUSIONES.

El desarrollo de las bases de datos y los sistemas de computación han generado gran cantidad de información que sólo puede ser justificada si se utiliza como fuente de información para mejorar el proceso en el que es generada. Sin embargo dada la novedad del sector y la característica de I+D del proceso de análisis, éste no se realiza de forma suficientemente estructurada, por lo que se producen grandes errores en las estimaciones de coste y plazo en este tipo de proyectos.
La utilización de una metodología estructurada y organizada presenta las siguientes ventajas para la realización de proyectos de *Data Mining*:

- Facilita la realización de nuevos proyectos de *Data Mining* con características similares
- Facilita la planificación y dirección del proyecto
- Permite realizar un mejor seguimiento del proyecto

En este trabajo se han presentado las dos principales metodologías utilizadas para el desarrollo de proyectos de *Data Mining*, así como las fases establecidas por cada metodología para el desarrollo del proceso. De este análisis se concluye que

REFERENCIAS.

- De Abajo Martínez, N. *Optimización mediante data mining de modelos para el diagnóstico de calidad en hojalata.* Tesis doctoral, Universidad de Oviedo, 2001
- Fayyad, U. *Mining science data,* Comunications of ACM, Vol. 39. Nº 11, 1996
- Chapman, P, Clinton, J. Khabaza, T. Reinartz, T. Rüdiger, W. The CRISP-DM Process Piatetski-Shapiro G., Frawley W.J: Knowledge discovery in databases. Ed. AAAI/MIT Press, 1991
- Rodríguez Montequín M.T. *Técnicas de análisis de datos*. Departamento de Matemáticas. Universidad de Oviedo. 2002.
- http://www.sas.com/technologies/analytics/datamining/miner/semma.html : Descripción de la metodología SEMMA
- http://www.crisp-dm.org/: Metodología CRISP-DM