Introducción a la Minería de Datos y el Data Warehousing

Sergio R. Coria
E-mail: sergio@mineriadedatos.com.mx

Resumen. Para hallar patrones significativos en grandes volúmenes de datos se ha usado inicialmente la estadística y, más recientemente, el aprendizaje automático, un área de la inteligencia artificial. La conjunción de estas disciplinas con la teoría y práctica de las bases de datos ha dado origen a la minería de datos, conocida también como descubrimiento de conocimiento en bases de datos. Los patrones hallados constituyen modelos descriptivos, predictivos o clasificadores que posteriormente pueden servir para implementar software de aplicación especializada o para guiar la revisión de políticas o procedimientos de las organizaciones. La necesidad de preservar y organizar datos para facilitar su consulta y análisis ha dado origen a los *data warehouses* y *data marts*. Un data warehouse es una base de datos sumarizados, organizados en tablas de hechos y dimensiones que facilita el procesamiento analítico en línea (OLAP), favoreciendo el desempeño en consultas masivas.

Palabras clave: minería de datos, descubrimiento de conocimiento en bases de datos, *data warehouse*, *data mart*.

Introducción

La motivación principal de la *minería de datos* (MD) y el *data warehousing* (DW) es la necesidad de organizar grandes volúmenes de datos y descubrir patrones significativos no triviales que sirvan a investigadores y administradores para lograr un mayor entendimiento de los fenómenos y procesos de su interés.

En este documento se presentan algunos de los conceptos fundamentales de la MD y el DW, así como las metodologías más comúnmente usadas en estas disciplinas.

1. Minería de datos

La minería de datos (*data mining*), conocida también como descubrimiento de conocimiento en bases de datos (*knowledge discovery in databases*), es una disciplina de las ciencias e ingenierías de la computación que intenta hallar patrones significativos en conjuntos de datos para producir modelos descriptivos, predictivos y clasificadores apoyándose en técnicas de manejo y programación de bases de datos, en estadística y aprendizaje automático (ML, por *machine learning*).

El ML es de especial utilidad para la MD. Es una disciplina de la inteligencia artificial en la que se crean algoritmos y modelos que intentan imitar la capacidad que tienen los sistemas nerviosos de los seres vivos para abstraer patrones. La noción de *patrón* es de gran importancia en la MD; se entiende como la combinación de características o de eventos que presentan alguna regularidad para la percepción por tener algún tipo de orden o de estructura.

En el ML las capacidades de abstracción se emulan al buscar y modelar las interacciones que existan entre los campos (atributos, variables, *features*). Es frecuente que los diversos algoritmos de ML se basen en estadística y en teorías de las probabilidades y de la información.

Existen dos grandes grupos de algoritmos de ML: aprendizaje *supervisado* y *no supervisado*. En ambos casos, el algoritmo recibe como entrada un conjunto de datos (*data set*) y produce como salida un modelo descriptivo, clasificador o predictivo. El data set es una tabla bidimensional, organizada en renglones y columnas. Cada renglón

constituye una instancia, ejemplo, registro o tupla que describe un caso real del proceso o fenómeno analizado. Cada columna constituye un atributo.

En el aprendizaje supervisado, el data set contiene un atributo denominado *atributo de clase*, que especifica a cuál clase o categoría clasificatoria pertenece cada instancia del data set. El atributo de clase se usa como *target* (objetivo) para crear modelos.

En el aprendizaje no supervisado, el *data set* no incluye atributo de clase debido a que en el proceso o fenómeno analizado no se dispone de este dato. El principal propósito de esta modalidad de aprendizaje es hallar las clases o categorías que pudieran existir en el data set. Los modelos construidos intentan representar las similitudes que existen entre las instancias de las clases halladas. Una vez descubiertas las clases, se puede agregar al data set un atributo de clase cuyos valores serán asignados con base en los patrones hallados. Después, el atributo de clase puede usarse como target en la aplicación de algoritmos de aprendizaje supervisado.

1.1 Qué se produce en un proyecto de MD

El principal producto en un proyecto de MD es uno o más modelos descriptivos, clasificadores o predictivos, basados en estadística y/o en aprendizaje automático.

1.2 Método general de la MD

Todo proyecto de MD requiere la colaboración entre el analista de MD y el usuario experto del dominio de conocimiento al cual se refieran los datos a analizar. La generación de un modelo basado en MD consiste en los siguientes pasos: 1) definición del objetivo del modelo, 2) selección de datos para análisis y modelación y de sus fuentes, 3) recolección, limpieza y pre-procesamiento de datos, 4) análisis estadísticos básicos, 5) selección y aplicación de algoritmos de aprendizaje automático, 6) reporte y evaluación de hallazgos con el experto de dominio, 7) explotación de los hallazgos. A continuación se describe cada uno de los pasos.

1.2.1 *Definición del objetivo del modelo*. El modelo puede ser descriptivo, clasificador o predictivo. En una situación inicial, si el conjunto de datos no ha sido sometido previamente a modelación con MD, el modelo debe ser descriptivo y particularmente con un propósito exploratorio, intentándose obtener una visión general del proceso o fenómeno modelado. Si ya se tiene conocimiento previo de los patrones generales presentes en los datos, se puede optar por la generación de modelos predictivos o clasificadores, en cuyo caso se requiere la colaboración de usuarios expertos del dominio correspondiente.

1.2.2 *Selección de datos y sus fuentes*. Una vez definido el objetivo, se tiene que determinar cuáles datos se usarán para construir los modelos. Se identifican los nombres y tipos de datos disponibles, así como su ubicación en los diversos sistemas de la empresa o institución y, eventualmente, en fuentes externas. Las fuentes de datos de la empresa o institución son generalmente sus sistemas de Procesamiento de Transacciones en Línea (*On Line Transaction Processing*, OLTP), sus *data marts* o su *data warehouse*. Los datos se seleccionan con base en el objetivo del análisis, apoyándose en los conocimientos del experto de dominio.

1.2.3 *Recolección, limpieza y pre-procesamiento.* Los datos seleccionados deben ser copiados desde sus fuentes originales. Se verifican sus características de formato, errores tipográficos y valores faltantes, principalmente. Eventualmente se puede requerir la corrección o eliminación de valores erróneos o de campos con valores faltantes. El pre-procesamiento consiste en organizar los datos generalmente en una única tabla para alimentar a los algoritmos de aprendizaje automático. En algunos casos

es necesario generar datos derivados a partir de los datos originales aplicando procesos aritméticos o textuales.

1.2.4 *Análisis estadísticos*. Éstos consisten en la búsqueda de patrones generales en los datos mediante el uso de herramientas estadísticas básicas, tales como el análisis de Pareto, los histogramas y los diagramas de barras. Si el análisis estadístico sugiere la existencia de patrones significativos, se tendrán mayores posibilidades de producir modelos útiles; si no, la aplicación de aprendizaje automático podría ser infructuosa. El análisis también contribuye en la selección de un algoritmo de aprendizaje automático adecuado en función del objetivo del proyecto y de las características del data set. Además, ofrece una referencia para la prueba estadística de los modelos.

1.2.5 *Selección y aplicación de algoritmos de aprendizaje automático*. Con base en el objetivo del modelo, en las características de los datos y en los patrones generales hallados, se eligen los algoritmos de aprendizaje automático. Los criterios de selección incluyen, entre otros aspectos, el hecho de que las instancias se encuentren o no *etiquetadas* y que los datos constituyan o no series de tiempo. Instancias etiquetadas son aquellas en las cuales cada instancia de los datos analizados tiene especificada la clase o categoría clasificatoria a la cual pertenece. *Serie de tiempo* es un conjunto de valores de un atributo numérico que se van produciendo a lo largo de un período determinado. La selección del algoritmo o tipo de modelo más adecuado requiere que el analista de MD tenga los conocimientos mínimos elementales de ML. La mayoría de los algoritmos de ML se encuentran ya implementados en herramientas con interfaces gráficas de fácil utilización. La precisión y confiabilidad de los modelos deben ser evaluadas estadísticamente, calculando, entre otros, los siguientes indicadores: *accuracy*, estadístico *Kappa*, *precision* (es distinto de *accuracy*), *recall*, medida *F*, etc.

1.2.6 *Reporte y evaluación de hallazgos con el experto de dominio.* Los análisis estadísticos y los modelos de aprendizaje automático se presentan a los usuarios expertos de dominio en un reporte detallado y claramente explicado. El usuario determina si los hallazgos son consistentes con su conocimiento experto del proceso o fenómeno estudiado y decide si éstos son útiles para ser explotados.

1.2.7 *Explotación de los hallazgos*. Con los usuarios expertos de dominio se pueden explotar los hallazgos en una o varias de las siguientes modalidades: a) implementando los modelos sobre los sistemas OLTP de la organización para realizar clasificación o pronóstico automáticamente. b) implementando Sistemas de Soporte a las Decisiones (*Decision Support Systems*, DSS), tales como sistemas expertos o software para elaboración de presupuestos, planes, etc. c) revisando las políticas y procedimientos de la empresa o institución.

2. Data warehousing

El *data warehousing* es el conjunto de técnicas para diseñar, construir y mantener datotecas. Una datoteca es una colección de datos organizados de modo que se optimice el desempeño de las consultas de grandes volúmenes de información. Las datotecas se diferencian de las bases de datos localizadas en los sistemas OLTP porque el propósito principal de las datotecas es facilitar y eficientar las operaciones de consulta de grandes volúmenes de datos para hacer Procesamiento Analítico en Línea (*On Line Analytical Processing*, OLAP). En cambio, las bases de datos de los sistemas OLTP intentan favorecer el desempeño de operaciones de actualización con volúmenes pequeños de datos. Generalmente las datotecas se ubican en servidores separados de los sistemas OLTP para evitar que el procesamiento de consultas voluminosas disminuya el desempeño del OLTP. Otra

diferencia importante es tipo de usuarios típicos de cada uno: los de las datotecas son mayoritariamente de nivel gerencial o directivo, realizando tareas nivel táctico o estratégico; los de OLTP realizan actividades a nivel operativo.

En la mayoría de las datotecas se realiza un proceso denominado Extracción-Transformación-Carga (*Extract-Transform-Load*, ETL). Una datoteca puede construirse en alguna de tres modalidades: 1) repositorio, 2) *data mart* o 3) *data warehouse*. A continuación se describe cada modalidad y se explican los métodos comúnmente aplicados para su construcción.

2.1 Repositorio

Un repositorio es una copia (*réplica*) de una base de datos proveniente de un sistema OLTP. Es la forma más sencilla de datoteca porque los datos generalmente se mantienen organizados en estructuras de tablas que son iguales a las de la base original. Los datos preservan sus valores originales y en caso de que algunos campos contengan valores erróneos o faltantes se les aplican procesos de corrección o eliminación (eliminando campos y/o tuplas), que se establecen a conveniencia de los usuarios. En el repositorio se van añadiendo datos periódicamente conforme se acumulan en el sistema OLTP. En general, en un repositorio no se hacen transformaciones de los datos ni se generan datos derivados, aunque esto no está impedido.

El repositorio se coloca generalmente en un servidor separado del correspondiente al sistema OLTP. El motivo es evitar que los accesos al repositorio para consultar grandes volúmenes de datos reduzcan el desempeño del sistema OLTP. También se intenta que las altas cargas de trabajo originadas por las transacciones realizadas por los usuarios del sistema OLTP no reduzcan el desempeño en las consultas realizadas por los usuarios del repositorio.

2.1.1 Construcción de un repositorio

Un repositorio se construye usando las funcionalidades de copia simple o de replicación de tablas y de bases de datos disponibles en los sistemas manejadores de bases de datos. Para la corrección o eliminación de campos con valores erróneos o faltantes se usan scripts en lenguaje SQL y también se puede recurrir a herramientas especializadas de *higienización de bases de datos*. La higienización consiste principalmente en corrección semi-automática de: ortografía, errores tipográficos, inconsistencias en datos de domicilios.

La implementación del repositorio tiene dos etapas: la *carga inicial* y el *refrescamiento*. La carga inicial consiste en introducir datos al repositorio cuando éste se encuentra completamente vacío. El refrescamiento se aplica después de la carga inicial y consiste en incorporar para propósitos de actualización los datos más recientemente generados por el OLTP.

2.2 Data mart

Un *data mart* (mercado de datos) es una base de datos multidimensional (*Multidimensional Database*, MDD) que contiene información de un área, departamento o proceso determinado de la empresa o institución. Por ejemplo, información de ventas, de compras, de producción, etc. Una MDD es aquella que se organiza en tablas de hechos (*facts*), llamados también métricas (*measures*), y tablas de dimensiones (*dimensions*).

Un hecho es un valor numérico sumarizado; *p. ej*. un monto de ventas expresado en una unidad monetaria. La sumarización puede consistir en una suma simple, o bien, en un conteo de frecuencias, un promedio, un porcentaje, un valor máximo o mínimo, etc. Una dimensión es un dato que determina el contexto a partir del cual se sumariza un hecho; *p. ej*. un determinado período de

tiempo, un alcance territorial, un tipo de producto, etc. Cada dimensión puede organizarse en *jerarquías*. Una jerarquía es un nivel de agregación para contextualizar hechos; *p. ej*. una dimensión territorial puede jerarquizarse en: sucursal, ciudad, Estado y país. Un hecho puede estar contextualizado por una o varias dimensiones; *p. ej*. las ventas generadas por un conjunto específico de sucursales en un mes determinado.

Las MDD se implementan en esquema de estrella (*star*) o de copo de nieve (*snow flake*). Ambos presentan cierta semejanza con el esquema relacional, el más utilizado en sistemas OLTP; pero existen diferencias significativas. Tanto en el de estrella como en el de copo de nieve existen tablas de hechos y de dimensiones. Sin embargo, en el de copo de nieve cada dimensión puede tener relacionada una serie de tablas de *subdimensiones*; mientras que en el de estrella cada dimensión está restringida y no puede tener subdimensiones. Estas diferencias son significativas para la velocidad de procesamiento y para la facilidad de modificación de la base de datos, ya que cada uno de los dos esquemas favorece una u otra.

2.2.1 Construcción de un data mart

Un *data mart* se construye siguiendo los siguientes pasos generales: a) identificar los datos disponibles para los usuarios en sus sistemas OLTP, b) identificar los datos numéricos sumarizables, c) determinar los hechos que pueden calcularse a partir de éstos, d) determinar las dimensiones que pueden producirse para dar contexto a los hechos, e) diseñar las tablas de hechos y dimensiones para la MDD, f) implementar la MDD sobre una plataforma de base de datos, g) realizar la carga inicial y h) refrescar el *data mart* con la periodicidad adecuada.

2.3 Data warehouse

Un *data warehouse* (bodega de datos) es una MDD similar al *data mart* y se caracteriza por contener datos sumarizados de todas las áreas, departamentos y procesos de una empresa o institución. Su diferencia principal con el *data mart* es el tamaño y el alcance.

2.3.1 Construcción de un data warehouse

Un *data warehouse* se construye siguiendo los mismos pasos generales que al construir un *data mart*. El *warehouse* puede construirse sin tener otra fuente que los sistemas OLTP de los usuarios, o bien, a partir de uno o más *data marts* que ya existan. Una forma conveniente es crear primero uno o varios *marts* y después el *warehouse*.

3. Comentarios finales

La necesidad del *data warehousing* y la minería de datos en los ámbitos científico, gubernamental y de negocios ha venido creciendo durante los últimos años y es evidente que esta tendencia prevalecerá. Por ello, es pertinente profundizar en la investigación básica y aplicada de estas disciplinas y ampliar sus aplicaciones prácticas.

REFERENCIAS

Sobre minería de datos

BERRY, M. and LINOFF, G.S. Data Mining Techniques for Marketing, Sales, and Customer Support. Edit. John Wiley & Sons, Inc., 2004.

DILLY, R. (Based on S.S. Anand). Data Mining: an Introduction, Version 2.0, Feb 1996: www.pcc.qub.ac.uk/tec/courses/datamining/ohp/dm-OHP-final_1.html

HAN, J., KAMBER, M., PEI, J.: Data Mining: Concepts and Techniques (3rd ed.). The Morgan-Kaufmann Series in Data Management Systems, Waltham (2011).

HERNÁNDEZ ORALLO, J., RAMÍREZ QUINTANA, M. J., FERRI RAMÍREZ, C. Introducción a la Minería de Datos. Pearson / Prentice-Hall. España.

MIERSWA, I., Wurst, M. and Klinkenberg, R. and Scholz, M. and Euler, T., Yale (now: RapidMiner): Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), 2006.
Software disponible en: http://rapid-i.com

WITTEN, I.H., and FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann Series in Data Management Systems (paperback - Jun 10, 2005). Elsevier, 2005.
Software disponible en: http://www.cs.waikato.ac.nz/~ml/weka

Sobre data warehousing

CHAUDHURI, Surajit, y UMESHWAR, Dayal. An Overview of Data Warehousing and OLAP Technology. VLDB Conference, 1996.

HOBBS, L., HILLSON, S., LAWANDE, S and SMITH, P. Oracle 10g Data Warehousing (paperback). Elsevier, 2005.

KAISER, B.U. Corporate Information with SAP-EIS: Building a Data Warehouse and a MIS-Application with inSight (Efficient Business-computing) (Hardcover). Originally published in German, 1998.

KIMBALL, R. and CASERTA, J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning (paperback - Sep 13, 2004). John Wiley & Sons, Inc. New York, USA

KIMBALL, R. and ROSS, Margy. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (second edition, April 26, 2002). John Wiley & Sons, Inc. New York, USA

KIMBALL, R., ROSS, M., THORNTHWAITE, W., MUNDY, J. and BECKER, B. The Data Warehouse Lifecycle Toolkit (second edition, 2008). John Wiley & Sons, Inc. New York, USA

MUNDY, J., THORNTHWAITE, W. and KIMBALL, R. The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset (Paperback - Feb 13, 2006). John Wiley & Sons, Inc. New York, USA

RAFANELLI, M. Multidimensional Databases: Problems and Solutions. Idea Group Publishing. USA, 2003

STACKOWIAK, R., RAYMAN, J. and GREENWALD, R. Oracle Data Warehousing and Business Intelligence Solutions (paperback - Jan 10, 2007). John Wiley & Sons, Inc. New York, USA