MINERÍA DE DATOS: HERRAMIENTA DE APOYO EN LA SELECCIÓN DE EQUIPOS DE PROYECTOS INFORMÁTICOS

Resumen / *Abstracts*

El proceso de adquisición o selección de equipos, dentro del área de la gestión de los recursos humanos de proyectos informáticos se considera un elemento crítico para garantizar el éxito. Lograr una selección idónea del capital humano de un equipo de desarrollo implica un análisis profundo de las características y resultados de proyectos concluidos, así como de las medidas y evaluaciones del desempeño individual y colectivo. En el presente trabajo se expone el estudio realizado con el propósito de valorar la aplicabilidad de la minería de datos para el reconocimiento y extracción de información desconocida, que apoye la toma de decisiones vinculadas a dicho proceso.

The process of acquisition or selection of teams, inside the human resources management area of informatics projects is considered a critical issue to guarantee the success. Accomplishing a suitable selection of the human resources of a development team implies a profound analysis of the features and results of finished projects, and also of the measures and evaluations of the individual and collective performance. The aim of this research is to assess the data mining application in the recognition and extraction of unknown information, in order to support the decisions taking involved in this process.

Palabras clave / *Key words*

Selección de equipos, recursos humanos, proyectos informáticos, minería de datos

Selection of teams, human resources, informatics projects, data mining

INTRODUCCIÓN

Como consecuencia del crecimiento en el volumen y complejidad de los productos de software, los proyectos se desarrollan fundamentalmente en equipos de personas, de ahí la importancia de formar equipos de proyectos sólidos y exitosos.[1]

Utilizando metodologías como PSP (Personal Software Process, Proceso de Software Personal) [2] y TSP (Team Software Process, Proceso de Software en Equipo),[3] es posible obtener numerosos datos que pueden resultar de gran utilidad para apoyar la toma de decisiones en el proceso de adquisición o selección de equipos de proyectos.[1] Sin embargo, esta valiosa información histórica puede ser ignorada; pues se dificulta su análisis, debido al volumen de información y el poco tiempo del que disponen los gerentes de proyectos informáticos. Por esta razón, la selección estratégica de equipos de proyectos se ve obstaculizada, lo cual atenta contra la estabilidad y el desarrollo exitoso de las empresas cubanas de software.

El presente trabajo tiene como objetivo general: Valorar la aplicabilidad de la minería de datos para el reconocimiento y extracción de información desconocida, que apoye la toma de decisiones relativas al proceso de selección de equipos de proyectos de software.

Ingrid Wilford Rivera, *Ingeniera Informática, Instructora, Centro de Estudios de Ingeniería de Sistemas (CEIS), Instituto Superior Politécnico José Antonio Echeverría, Cujae, Ciudad de La Habana, Cuba*

e-mail:ingrid@ceis.cujae.edu.cu.

Recibido:mayo del 2006
Aprobado:julio del 2006

La minería de datos es un campo multidisciplinario que se ha desarrollado como extensión de otras tecnologías: bases de datos, recuperación de información, estadística, aprendizaje automático, visualización, sistemas de tomas de decisiones, computación paralela y otras. Es por ello que, la investigación y los avances en la minería de datos se nutren de los que se producen en estas áreas relacionadas. De estas disciplinas que contribuyen a la minería de datos, el aprendizaje automático, es el área de la inteligencia artificial que se ocupa de desarrollar algoritmos capaces de aprender, y constituye, junto con la estadística, el corazón del análisis inteligente de los datos.

MINERÍA DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS

Existen numerosas definiciones de minería de datos (MD): Usama Fayyad, la define como un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos.[4] En esencia, la minería de datos, es un mecanismo de explotación y análisis, consistente en la búsqueda y extracción de información valiosa en grandes volúmenes de datos.[5,6] Los *data warehouses* (DW) o *data marts* (DM) proporcionan la información histórica necesaria, con la cual operan los algoritmos de minería de datos.[6,7]

La minería de datos revela patrones o asociaciones que generalmente eran desconocidos, por lo que algunos autores le llaman también **descubrimiento de conocimiento en las bases de datos** o *knowledge discovery in databases* (KDD); sin embargo, muchos autores se refieren al proceso de minería de datos como una etapa, como la aplicación de algoritmos para extraer patrones de datos, y nombran KDD al proceso completo (identificación del problema, preprocesamiento, minería de datos, posprocesamiento).[5,6]

En el proceso de KDD, una vez identificado el problema y efectuado el preprocesamiento se van definiendo las posibles tareas a realizar.[8]

ACTIVIDADES O TAREAS QUE INCLUYE LA MINERÍA DE DATOS

Las actividades de minería de datos combinan la tecnología de bases de datos y *data warehousing*, con técnicas de aprendizaje automático y estadística. Son muy diversas las clasificaciones halladas en la bibliografía consultada, referentes a las posibles tareas de MD. Una primera es la siguiente:[4,7]

- Tareas descriptivas
- Tareas predictivas

Las **tareas descriptivas** tienen como objetivo transformar el conjunto modelo (*model set*) en informaciones precisas, que reflejen las propiedades más relevantes y generalidades de los datos.[6] Estas tareas de MD, construyen modelos sobre patrones de hechos ocurridos en el pasado para su presentación de una forma comprensible.

Las **tareas predictivas**, sin embargo, construyen modelos sobre datos ocurridos en el pasado para predecir el comportamiento de nuevos conjuntos de datos (*score sets*), los cuales pueden corresponderse o no con hechos futuros.[6]

Otra clasificación, más desglosada que la anterior, es la siguiente:[5,6,9]

1. Clasificación.
2. Estimación.
3. Predicción.
4. Determinar grupos afines o reglas de asociación.
5. *Clustering* o agrupamiento.
6. Descripción y visualización o tareas de preprocesamiento.

La clasificación, estimación y predicción se agrupan con el calificativo de **minería de datos directa** (MDD) o **métodos supervisados**; mientras que las tres restantes actividades (determinar grupos afines o reglas de asociación, *clustering* y descripción y visualización) conforman el grupo de **minería de datos indirecta** (MDI) o métodos no supervisados.[5,6] En el caso de la MDD el objetivo está bien determinado, a diferencia de la MDI, en la que no se tienen claro los resultados que se desean obtener.[9]

La clasificación consiste en identificar características de un objeto o registro con el propósito de asignarle una clase o categoría predefinida. Su propósito es descubrir si una instancia del conjunto de ocurrencias (*data set*), pertenece a una de varias clases previamente definidas. Para ello se requiere construir un modelo de clasificación. La salida obtenida la componen valores discretos: clases. En la práctica, estas clases son usualmente definidas a partir de valores específicos de determinadas variables o campos.[5,6]

Las actividades de clasificación pueden tener un enfoque tanto descriptivo como predictivo. En el primer caso, se desea conocer las variables y valores más significativos de las instancias de cada tipo de clase. Mientras que, en el segundo caso, se desea explorar las características de una nueva instancia y asignarle una de las clases predefinidas.[5,6]

La estimación es semejante a la clasificación, pero la salida la constituyen valores continuos. En algunos casos es posible hacer estimación y posteriormente clasificación.[6,9] Por ejemplo: estimaciones de altura, gastos, salarios, etcétera.

Los valores estimados por estas técnicas pudieran incluso ser utilizados para ordenar las instancias en un *score set*,[6] y luego seleccionar las **mejores**, según el problema a resolver.

La predicción es similar a las dos anteriores, con la particularidad de que la salida, sea continua o discreta, no ha ocurrido: la variable estimada o la clase asignada se refiere a un evento que ocurrirá en el futuro.[6,9]

Por otra parte, la actividad de MDI: Determinar grupos afines o reglas de asociación, se encarga de descubrir fenómenos que ocurren de conjunto, aunque se desconoce el tipo de relación causal que existe entre estos. A partir de los grupos afines identificados, es posible generar reglas de asociación entre los datos.[6]

Las tareas de *clustering* o agrupamiento, tienen el propósito de formar subgrupos homogéneos (*clusters*), a partir de un grupo diverso, según el grado de semejanza entre ellos. El significado de los grupos identificados, de existir, es una tarea del usuario. Clasifica también como MDI, ya que se desconoce la cantidad y el significado de los grupos que se obtendrán.[6]

Existen dos tipos de técnicas de *clustering*, aquellas que crean una colección de grupos y las que generan una jerarquía de *clusters*. La jerarquía de *clusters* se justifica por el hecho de que al ser el *clustering* una actividad de MDI, sigue una estrategia no supervisada, no existe una respuesta absolutamente correcta.[6]

Por último, la descripción y visualización o las tareas de preprocesamiento, son una actividad de MDI muy importante, pues su aplicación permite enfocar las restantes actividades de minería de datos.

MINERÍA DE DATOS Y SU POSIBLE APLICACIÓN EN LA SELECCIÓN DE EQUIPOS DE PROYECTOS INFORMÁTICOS

El campo de acción de la MD es muy amplio y variado. Son numerosos los ejemplos en los que esta puede ser usada para resolver problemas muy importantes: en la industria farmacéutica como herramienta de investigación, en el deporte, en campañas de marketing, en la bioinformática, en la construcción de sistemas expertos, en la obtención y validación de modelos de estimación de software, en el control y análisis del tráfico, entre muchas otras aplicaciones.[6,10]

En el ámbito de la gestión de proyectos informáticos, una vez definido un nuevo proyecto, se requiere seleccionar el equipo que lo va a desarrollar. Para ello es conveniente considerar características, roles definidos y resultados de los proyectos realizados semejantes al nuevo, plantilla de roles a emplear, experiencia de los desarrolladores en el desempeño de estos roles, **idoneidad** y grado de afinidad entre ellos para el trabajo en equipo. En este sentido, es posible definir los siguientes pasos o tareas esenciales en el proceso de selección de equipos de proyectos informáticos:

1. Identificar grupos de proyectos semejantes ya concluidos.
2. Clasificar el nuevo proyecto de software de acuerdo con los grupos de proyectos semejantes previamente identificados.
3. Seleccionar plantilla de roles a ocupar por los integrantes del equipo de proyecto.
4. Agrupar los ingenieros de la empresa según los roles que requiera el equipo de proyecto. Cada grupo concentrará a los individuos con habilidades demostradas en el desempeño del rol correspondiente.
5. Estimar la **idoneidad** de cada uno de los ingenieros agrupados por roles. Dicha **idoneidad** debe considerar: habilidades, motivación, experiencia en proyectos semejantes y probabilidad de tener un desempeño exitoso en el nuevo proyecto. Constituye un valor numérico.
6. Ordenar los ingenieros de cada uno de los grupos de acuerdo con la **idoneidad** estimada.
7. Determinar grupos de ingenieros afines, que se complementen y armonicen entre sí. Los ingenieros candidatos a agrupar son los ya seleccionados previamente y organizados por roles.

Llevar a cabo estas acciones cada vez que se defina un nuevo proyecto, puede llegar a ser una labor muy compleja e ineficiente si se realiza por los métodos tradicionales, considerando el volumen de datos históricos y el poco tiempo del que disponen los gerentes de proyectos, por lo que la aplicación de técnicas de minería de datos resulta favorable y necesaria.

A continuación se propone, de acuerdo con las tareas definidas previamente, las técnicas de MD que podrían usarse en cada caso.

Identificar grupos de proyectos semejantes ya concluidos y clasificar el nuevo proyecto de software

Utilizar *clustering* jerárquico para agrupar los proyectos ya concluidos según el grado de semejanza entre estos y asignar el nuevo proyecto a uno de los grupos que se determinen, de acuerdo también con la función de semejanza que se defina. Constituye una tarea de MDI, ya que se desconoce la cantidad y el significado de los grupos de proyectos que se obtendrán.

Seleccionar plantilla de roles

En este caso se propone emplear técnicas de resumen o visualización, para mostrar los roles definidos en los proyectos semejantes al nuevo y facilitar la selección de la plantilla de roles a emplear. El gerente de proyectos podrá observar patrones y generalidades a partir de los roles asignados en los proyectos semejantes al nuevo y de acuerdo con ello, seleccionar la plantilla de roles a emplear.

Crear grupos de ingenieros por roles

Emplear la técnica de clasificación para clasificar los ingenieros en uno de los roles de la plantilla seleccionada, según las habilidades y competencias demostradas en el desempeño de estos roles. La salida, en este caso, son clases, valores discretos.

Estimar idoneidad y ordenar los ingenieros en cada grupo

En este caso se propone utilizar la estimación, ya que el objetivo es asignar un valor numérico (salida continua) correspondiente a la **idoneidad** (considera: habilidades, motivación, experiencia en proyectos semejantes y probabilidad de tener un desempeño exitoso en el nuevo proyecto; constituye un valor numérico) de cada ingeniero. Posteriormente basado este valor se ordenan los objetos o registros y se realiza una clasificación por **idoneidad** (alta, media, baja).

Determinar grupos de ingenieros afines

Para determinar el grado de afinidad entre los ingenieros candidatos a formar el equipo de proyecto, se propone combinar las actividades de *clustering* jerárquico con determinación de grupos afines o reglas de asociación, considerando resultados individuales y colectivos alcanzados en proyectos realizados, resultados de evaluaciones, encuestas o test de afinidad efectuados.

CONCLUSIONES

En esencia, la minería de datos es un mecanismo de explotación y análisis, consistente en la búsqueda y extracción de información valiosa en grandes volúmenes de datos. Las actividades de minería de datos combinan la tecnología de bases de datos y almacenes de datos, con técnicas de aprendizaje automático y estadística.

Se considera adecuada la clasificación de tareas de minería de datos, en seis categorías: tareas de clasificación, estimación, predicción, determinación de grupos afines o reglas de asociación, *clustering* o agrupamiento y descripción y visualización (tareas de preprocesamiento).

En el proceso de selección de equipos de proyectos informáticos, se considera necesaria la aplicación de técnicas de minería de datos para descubrir relaciones ocultas entre los datos almacenados, e información desconocida, que resulte de utilidad para la toma de decisiones.

Para apoyar dicho proceso, y de acuerdo con la clasificación de tareas de minería de datos aceptada, se propone emplear las siguientes actividades de minería de datos: clasificación, estimación, determinar grupos afines o reglas de asociación, *clustering* jerárquico y resumen o visualización, como se describe previamente.

REFERENCIAS

1. "Guía de los Fundamentos de la Dirección de Proyectos (Guía del PMBOK®), 3ra. ed., Project Management Institute, 2004.
2. **LÓPEZ TRUJILLO, YUCELY Y MARGARITA ANDRÉ AMPUERO:** *Disciplina personal en el proceso de desarrollo de software. Primeros pasos a seguir en el contexto cubano*, Memorias de la XI Convención y Feria Internacional de Informática 2005, Ciudad de la Habana, Cuba, 2005.
3. **ANDRÉ AMPUERO, MARGARITA:** *El proceso de software en equipo: de la disciplina personal a la disciplina organizacional*, Memorias de la XI Convención y Feria Internacional de Informática 2005, Ciudad de la Habana, Cuba, 2005.
4. **FAYYAD; PIATESKY-SHAPIRO; SMITH AND UTHURUSAMY:** *Advance in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, Mass, 1996.
5. **BERRY, M. AND G. LINOFF:** *Mastering Data Mining, The Art and Science of Customer Relation-ship Management*, John Wiley & Sons, Inc, 2000.
6. **———.:** *Data Mining Techniques for Marketing Sales, and Customer Relationship Management,* Second Edition, Wiley Pub., Inc., 2004.
7. **HAN, JIAWEI AND MICHELINE KAMBER:** *Data Mining: Concepts and Techniques*, Morgan Kauf-Mann, 2001.
8. **MOLINA FÉLIX, LUIS CARLOS:** "Data Mining: Torturando a los datos hasta que confiesen", 2002, Disponible en: http://www.uoc.edu/molina1102/esp/molina1102/molina1102.html, Fecha de consulta: marzo de 2005.
9. **ROSETE SUÁREZ, ALEJANDRO:** *Minería de datos: El camino de la academia a la realidad cotidiana*, VIII Congreso Internacional de Ciencias de la Computación (CICC 2003), Bolivia, 2003.
10. **WEBER, R.:** "Data Mining en la empresa y en las finanzas utilizando tecnologías inteligentes", *Revista Ingeniería de Sistemas*, Vol. XIV, Universidad de Chile, Junio, 2000.