En español

CMIN - herramienta case basada en CRISP-DM para el soporte de proyectos de minería de datos

Carlos Cobos[1], Jhon Zuñiga[2], Juan Guarin[3], Elizabeth León[4] y Martha Mendoza[5]

RESUMEN

En este artículo se presenta la CMIN, una herramienta CASE (*Computer Aided Software Engineering*) integrada (que soporta todas las fases de un proceso) basada en CRISP-DM 1.0 (*Cross – Industry Standard Process for Data Mining*) para soportar el desarrollo de proyectos de minería de datos. Primero se expone la funcionalidad general de CMIN, lo que incluye la gestión de procesos, plantillas y proyectos, y se destaca la capacidad de CMIN para realizar el seguimiento de los proyectos de una forma fácil e intuitiva y la manera como CMIN posibilita que el usuario incremente su conocimiento en el uso de CRISP-DM o de cualquier otro proceso que se defina en la herramienta a través de las ayudas e información que se ofrece en cada paso del proceso. Después, se detalla cómo CMIN permite enlazar en tiempo de ejecución (sin necesidad de volver a compilar la herramienta) nuevos algoritmos de minería de datos que apoyen la labor de modelado (basada en un flujo de trabajo o *workflow*) en un proyecto de minería de datos. Finalmente, se ofrecen los resultados de dos evaluaciones de la herramienta, las conclusiones y el trabajo futuro.

Palabras clave: minería de datos, CRISP-DM, herramientas CASE, *workflow*, reflexión.

Recibido: julio 21 de 2009
Aceptado: noviembre 15 de 2010

In English

CMIN – a CRISP-DM-based case tool for supporting data mining projects

Carlos Cobos[6], Jhon Zuñiga[7], Juan Guarin[8], Elizabeth León[9], Martha Mendoza[10]

ABSTRACT

This paper introduces CMIN, an integrated computer aided software engineering (CASE) tool based on cross-industry standard process for data mining (CRISP-DM) 1.0 designed to support carrying out data mining projects. It is "integrated" in the sense that it supports all phases of a process. A general overview of how CMIN works is presented first, including a treatment of processes, templates and project management. CMIN's capacity for easily and intuitively monitoring projects is highlighted, as is the manner in which CMIN allows a user to increase knowledge regarding using CRISP-DM or any other process defined in the CASE tool through the help and information presented in each step. Next, it is shown how CMIN can bind new data mining algorithms in runtime (without the need to recompile the tool) to support modelling tasks (based on a Workflow) and evaluate data mining projects. Finally, the results of two evaluations of the tool, some conclusions and suggestions for future work are presented.

Keywords: Data mining, CRISP-DM, CASE tools, workflow, reflection.

Received: july 21th 2009
Accepted: november 15th 2010

[1] Ingeniero de Sistemas. M.Sc., en Informática, Universidad Industrial de Santander, Colombia. Candidato a Ph.D., en Ingeniería de Sistemas y Computación, Universidad Nacional de Colombia, Bogotá, Colombia. Docente de Planta Tiempo Completo Categoría Titular, Universidad del Cauca, Colombia. Investigador del Grupo de I+D en Tecnologías de la Información (GTI), Universidad del Cauca, Colombia. ccobos@unicauca.edu.co.

[2] Ingeniero de Sistemas, Universidad del Cauca, Colombia. Programador, Informática y Gestión S.A., Colombia. Auxiliar de investigación del Grupo de I+D en Tecnologías de la Información, Universidad del Cauca, Colombia. jzunigaparedes@unicauca.edu.co.

[3] Ingeniero de Sistemas, Universidad del Cauca, Colombia. Programador, Solsoft S.A., Colombia. Auxiliar de investigación del Grupo de I+D en Tecnologías de la Información, Universidad del Cauca, Colombia. jguarin@unicauca.edu.co.

[4] Ingeniera de Sistemas. M.Sc., en Ingeniería de Sistemas, Universidad Nacional de Colombia, Colombia. M.Sc., in Electrical and Computer Engineering, University of Memphis, EEUU. Ph.D., in Computer Science and Computer Engineering, University of Louisville, EEUU. Docente de Planta Tiempo Completo Categoría Asistente, Universidad Nacional de Colombia sede Bogotá, Colombia. Investigadora del Laboratorio de Investigación en Sistemas Inteligentes (LISI), Universidad Nacional de Colombia sede Bogotá, Colombia. eleonguz@unal.edu.co.

[5] Ingeniera de Sistemas. M.Sc., en Informática, Universidad Industrial de Santander, Colombia. Estudiante de Ph.D., En Ingeniería de Sistemas y Computación, Universidad Nacional de Colombia sede Bogotá, Colombia. Docente de Planta Tiempo Completo Categoría Titular, Universidad del Cauca, Colombia. Investigadora del GTI, Universidad del Cauca, Colombia. mmendoza@unicauca.edu.co.

[6] Systems Engineer. M.Sc in Computer Science, Universidad Industrial de Santander, Colombia. Ph.D., candidate in Computer and Systems Engineering, Universidad Nacional de Colombia, Bogotá, Colombia. Plant Teachers Full Time Category Holder, Universidad del Cauca, Colombia. Researcher ID Group on Information Technology (GIT), Universidad del Cauca, Colombia. ccobos@unicauca.edu.co.

[7] Systems Engineer, Universidad del Cauca, Colombia Programmer, Informática y Gestión S.A., Colombia. Research Assistant Group ID in Information Technology, Universidad del Cauca, Colombia. jzunigaparedes@unicauca.edu.co.

[8] Systems Engineer, Universidad del Cauca, Colombia. Programmer, Solsoft S.A., Colombia. Research Assistant Group ID in Information Technology, Universidad del Cauca, Colombia. jguarin@unicauca.edu.co.

[9] Systems Engineer. M.Sc., in Systems Engineering, Universidad Nacional de Colombia, Colombia. M.Sc., in Electrical and Computer Engineering, University of Memphis, EEUU. Ph.D., in Computer Science and Computer Engineering, University of Louisville, EEUU. Plant Teachers Full Time Category Assistant, Universidad Nacional de Colombia, Bogotá, Colombia. Laboratory researcher in Intelligent Systems Research (LISI), Universidad Nacional de Colombia, Bogotá, Colombia. eleonguz@unal.edu.co.

[10] Systems Engineer. M.Sc., in Computer Science, Universidad Industrial de Santander, Colombia. Ph.D., student in Engineering Systems and Computing, Universidad Nacional de Colombia sede Bogotá, Colombia. Plant Teachers Full Time Category Holder, Universidad del Cauca, Colombia GTI Researcher, Universidad del Cauca, Colombia. mmendoza@unicauca.edu.co.

En español

Introducción

En ingeniería de *software* se han establecido diversos procesos, metodologías y herramientas para estandarizar y facilitar el desarrollo de sus productos. Entre las herramientas se cuentan las CASE, las cuales soportan en forma automática varios o todos los pasos de dichas metodologías y se enmarcan en la ingeniería del *software* asistida por computador o *Computer Aided Software Engineering* (INEI, 1999). Las herramientas CASE ayudan a reducir el tiempo empleado en el desarrollo de un sistema, lo que mantiene el costo estable y contribuye a mejorar su calidad (Miren Begoña, 2000). Además, permiten al analista documentar y modelar un sistema, desde la definición de requerimientos hasta el diseño, implementación y prueba (Miren Begoña, 2000).

Hoy se encuentran diversas herramientas *software* para apoyar el desarrollo de proyectos de minería de datos (Britos *et al*., 2005; Kdnuggets, 2005; MetaGroup, 2004). Basado en el listado de herramientas que aparecen en MetaGroup (2004) y Kdnuggest (2005), se realizó una valoración de las más representativas, entre ellas: Clementine (Khabaza & Shearer, 1995; SPSS-Inc., 2009), Insightful Miner (Insightful-Corporation), WEKA (Holmes, Donkin & Witten, 1994; University-of-Waikato, 2009), CART (Salford-System, 2009), PolyAnalyst (Mai, Krishna & Reddy, 2005; Megaputer, 2009; Rippa & Lendyuk, 2007) y SAS Enterprise Miner (SAS, 2009a). Los criterios generales para dicha valoración fueron: el *acceso* (costo de las herramientas), la *interfaz de usuario* (facilidad o dificultad que puede llegar a tener el uso de la herramienta por parte de los usuarios), el *proceso* (o metodología) en la que se basan, la *extensibilidad* (capacidad de ampliar fácil y dinámicamente el conjunto de algoritmos que ofrece la herramienta) y el soporte al desarrollo del proyecto por parte de *equipos* de trabajo. Como resultado se encontró que ninguna de las herramientas cumple completamente con CRISP-DM (*Cross – Industry Standard Process for Data Minning*) (CRISP-DM, 2006; Chapman *et al*., 2000), un proceso para el desarrollo de proyectos de minería de datos iterativo, abierto, personalizable y de gran reconocimiento por la industria y la academia; que ninguna de estas herramientas permite la ampliación dinámica y en tiempo de ejecución (sin volver a compilar el código) del conjunto de algoritmos de minería que se entregan inicialmente con la herramienta; y que a pesar de que algunas herramientas cuentan con una interfaz fácil de usar ninguna de ellas guía apropiadamente el desarrollo de un proyecto y mucho menos ayudan a sus usuarios a conocer y profundizar en el manejo del proceso y en general del desarrollo de proyectos de minería. Por lo anterior, el grupo de investigación GTI decidió desarrollar una herramienta CASE integrada (que soporta todas las fases de un proceso), basada en CRISP-DM (CRISP-DM, 2006; Chapman *et al*., 2000), fácilmente extensible en tiempo de ejecución, fácil de usar y que ayude al usuario a mejorar sus conocimientos y habilidades en el desarrollo de proyectos de minería.

CRISP-DM: Cross-Industry Standard Process for Data Mining

Existen varias metodologías para orientar el proceso de minería de datos; ellas pretenden facilitar la realización de nuevos proyectos con características similares, optimizar la planificación y dirección de éstos, reducir su complejidad y permitir hacerle un mejor seguimiento a ellos (Gondar Nores, 2004). Entre esas metodologías se destacan CRISP-DM (2006) y SEMMA —*Sample, Explore, Modify, Model, Assess*— (SAS, 2009b). SEMMA se centra en las características técnicas del desarrollo del proceso, mientras que CRISP-DM mantiene como foco central los objetivos empresariales del proyecto. Debido a ello, CRISP-DM comienza realizando un análisis del problema em-

In English

Introduction

A variety of processes, methodologies and tools have been established in software engineering to standardise software product development and make it simpler. CASE tools are among the available tools; they automatically support a number or all of the aforementioned methodologies' steps and together are known as computer aided software engineering (CASE) (INEI, 1999). CASE tools help reduce the time required for developing a system, in turn helping to stabilise costs and contribute to quality enhancement (Miren Begoña, 2000). CASE tools further allow an analyst to document and model a system, from initially defining the requirements, through to design, implementation and testing (Miren Begoña, 2000).

A range of software tools are available today that help in carrying out data mining software projects (Britos *et al*., 2005; Kdnuggets, 2005; MetaGroup, 2004). Based on the list of such tools that appear in MetaGroup (MetaGroup, 2004) and Kdnuggets (Kdnuggets, 2005), an evaluation was made of the most representative, including: Clementine (Khabaza & Shearer, 1995; SPSS-Inc., 2009), Insightful Miner (Insightful-Corporation), WEKA (Holmes, Donkin, & Witten, 1994; University-of-Waikato, 2009), CART (Salford-System, 2009), PolyAnalyst (Mai, Krishna, & Reddy, 2005; Megaputer, 2009; Rippa & Lendyuk, 2007) and SAS Enterprise Miner (SAS, 2009a). The general criteria for such evaluation were: its access (cost of the tools), its user interface (how easy or complex the tool was to use according to the user), the process (or methodology) on which it was based, its extensibility (the capacity to easily and dynamically expand the set of algorithms it offers) and support in project development for individuals to work together in groups. It thus came to light that not one of the tools fully complied with the cross–industry standard process for data mining (CRISP-DM) (CRISP-DM, 2006; Chapman *et al*., 2000), a process for carrying out data mining projects that is at once iterative, open, customisable and widely recognised by industry and academia. It also emerged that none of the tools allowed dynamic real time expansion (without recompiling the tool) of the set of algorithms the tool initially produced and that, despite the fact that some of the tools boasted an easy user interface, not one of them properly guided the carrying out of a project, much less aided the user to learn and deepen their knowledge of process management in conducting a data mining project. As such, the research group (GTI) decided to develop an integrated CASE tool based on CRISP-DM (CRISP-DM, 2006; Chapman *et al*., 2000), easily extensible in run-time, easy to use and which helps a user to increase his/her knowledge and abilities in carrying out data mining projects.

Cross-industry standard process for data mining (CRISP-DM)

A variety of methodologies exists for directing data mining. These aim at facilitating new projects having similar characteristics, optimase their planning and management, reduce their complexity and allow smoother execution (Gondar Nores, 2004). Two of these methodologies stood out: CRISP-DM (CRISP-DM, 2006) and sample, explore, modify, model, assess (SEMMA) (SAS, 2009b). The latter concerns itself with the technical characteristics or process development, while CRISP-DM mainly focuses on a project's business objectives. CRISP-DM begins by carrying out an analysis of a business pro-

En español

presarial para su transformación en un problema técnico de minería de datos. CRISP-DM puede ser integrada con una metodología de gestión de proyectos específica que complemente las tareas administrativas y técnicas, además es de libre distribución, sin costo alguno, a diferencia de SEMMA (SAS, 2009b). CRISP-DM define una estructura para proyectos de minería de datos y suministra la orientación para su ejecución. Consta de un modelo de referencia y una guía de usuario (Chapman *et al*., 2000). El *modelo de referencia* da una visión general del ciclo de vida de un proyecto de minería de datos, contiene las fases con sus objetivos, las tareas y las relaciones entre éstas, y las instrucciones paso a paso que se deben llevar a cabo. Las fases definidas por el modelo de referencia son: comprensión del negocio, análisis de datos, preparación de los datos, modelamiento, evaluación y despliegue. Cada una de estas fases (nivel 1) está compuesta de tareas genéricas (nivel 2), que se dividen en tareas específicas (nivel 3), y finalmente, en el nivel 4 se encuentra la instancia del proceso, que describe las actividades específicas a efectuar en un proyecto de minería de datos. La *guía del usuario* brinda consejos detallados, pistas por cada fase, y cada operación dentro de una fase, y ejemplifica cómo hacer un proyecto de minería de datos. Esta guía de usuario es una excelente opción para desarrolladores que tienen poca experiencia en el desarrollo de este tipo de proyectos.

Modelo conceptual de CMIN

Para comprender mejor el funcionamiento de la CMIN primero se presenta el modelo conceptual del sistema, con los principales conceptos y las relaciones existentes entre éstos (Figura 1):

- *Usuarios:* comprende a las personas que pueden utilizar el sistema, los cuales pueden ser novatos o expertos en proyectos de minería de datos.
- *Módulo de procesos:* es el que permite la gestión de procesos, entre ellos CRISP-DM. La *definición de procesos* representa la acción de registrar un proceso mediante la agregación y definición de sus pasos, campos o actividades que se proponen para el desarrollo de un proyecto de minería de datos. Los *reportes* son los documentos o entregables que se deben proveer durante un proyecto, y que son soporte de la ejecución de él.
- *Procesos:* son los pasos que se han agregado a la CMIN y que sirven como base para gestionar los proyectos de minería con la herramienta.
- *Módulo de proyectos:* representa el módulo de gestión de proyectos de minería de datos basado en uno de los pasos previamente adicio-nado en el módulo de procesos. Los *proyectos* comprenden el con-junto de procesos que se han creado en la CMIN y que están en curso o han sido terminados. Los *campos o actividades* de un paso son las tareas específicas que se deben realizar para cumplir con el objetivo del paso al que pertenecen. Los *resultados* representan los productos de la realización de una actividad, que pueden ser: una sugerencia, un texto explicativo o una plantilla de información que se debe diligenciar.
- *Workflow (WF):* entorno gráfico que permite a los usuarios gestionar modelos de minería de datos basados en las tareas de minería definidas en la CMIN.
- *Agregación dinámica de DLL (librerías de enlace dinámico, o por las siglas en inglés de Dynamic Link Library):* es el módulo que permite la gestión de objetos (nuevos algoritmos) que sirven para la ejecución del WF, por medio de DLL. Los *tipos de objetos del flujo de trabajo* representan el conjunto de *tipos de objetos* reconocidos por la CMIN para ser agregados y posteriormente utili-

In English

blem for transforming it into a technical data mining problem. CRISP-DM can also be integrated with a specific project management methodology complementing administrative and technical tasks. It is also widely distributed at no cost, unlike SEMMA (SAS, 2009b). CRISP-DM defines a structure for data mining projects and provides orientation for their execution. It serves both as a reference model and a user guide (Chapman *et al*., 2000). The **reference model** gives a general view of a data mining project's life-cycle, containning each phase with its objective, the tasks, the relationships between them and the step-by-step instructions that must be carried out. The phases defined for the reference model are: understanding the business, data analysis, data preparation, modelling, evaluation and display. Each phase (level 1) is composed of generic tasks (level 2) divided into specific tasks (level 3) and an instance of the process is found in level 4, describing the specific activities to be done in a data mining project. The **user guide** offers detailed advice, tracks for each phase and each operation within a phase, and provides an example of how to do a data mining project. The user guide is an excellent option for researchers having little experience of data mining.

CMIN conceptual model

The conceptual model is presented first to understand better how CMIN works, with its main concepts and the relationships amongst them (see Figure):

- *Users:* people who use the system. They may be experts or novices in data mining;
- *Process module*: this is the module that allows process management, among which is found CRISP-DM. Process definition represents the action of registering a process during aggregation and defining its steps, fields or activities required for carrying out a data mining project. Reports are the documents or deliverables that need to be provided in the course of a project and which aid executing such project;
- *Processes:* processes that have been added to CMIN and that serve as a basis for managing data mining projects in the tool;
- *Project module:* the module for managing data mining projects, based on one of the processes previously added to the process module. Projects represent the set of projects already created in CMIN and can be found in two stages (in progress or completed). The fields or activities of a step are the specific activities that have to be carried out to meet the objective of the step to which they belong. The results represent the products of carrying out an activity, which may comprise a suggestion, an explanatory text or an information template that needs to be observed;
- *Workflow (WF)*: a graphical environment that allows users to manage data mining models based on mining tasks defined in CMIN;
- *Adding dynamic link libraries (DLL):* this module allows the management of objects (new algorithms) that serve to implement the workflow, using DLLs. Types of workflow objects (or types of objects) represents the set of object types recognised by CMIN to be added and in turn used by the WF. Interfaces represents the set of software contracts (e.g. for classification, clustering or

En español

- zados por el WF. Las *interfaces* abarcan el conjunto de contratos de *software* (por ejemplo, en clasificación, agrupación o reglas de asociación) que deben cumplir las DLL para agregarlas al conjunto de objetos que serán utilizados por el WF. Las *DLL* son el conjunto de logaritmos que posee actualmente la CMIN en su batería (objetos del WF).
- *Objetos de WF:* comprenden el conjunto de objetos que se agregan a la CMIN y pueden utilizarse en el WF, el cual puede crecer a medida que los usuarios hagan nuevas implementaciones de cualquiera de los tipos de objetos del WF especificados en la CMIN.
- *Servidor CMIN:* es el que aloja nuevas definiciones de procesos, así como nuevas implementaciones de objetos (algoritmos) del WF por medio de DLL para que los usuarios actualicen la CMIN si así lo requieren, ya que ella se ejecuta independientemente de este servidor.

In English

- association rules) to be met by DLLs before being added to the set of objects to be used by the WF. DLLs represent the set of DLLs that CMIN currently holds in its array, or algorithms set (WF objects);
- Workflow objects: the set of objects added to CMIN and which can be used in the workflow, which can grow in such a way that users make new implementations of any of the types of WF objects specified in CMIN; and
- CMIN server: the server that hosts new process definitions and new implementations of workflow objects (algorithms) by way of DLLs, so that users can upgrade CMIN if that is what is desired because CMIN is able to run independently of the server.

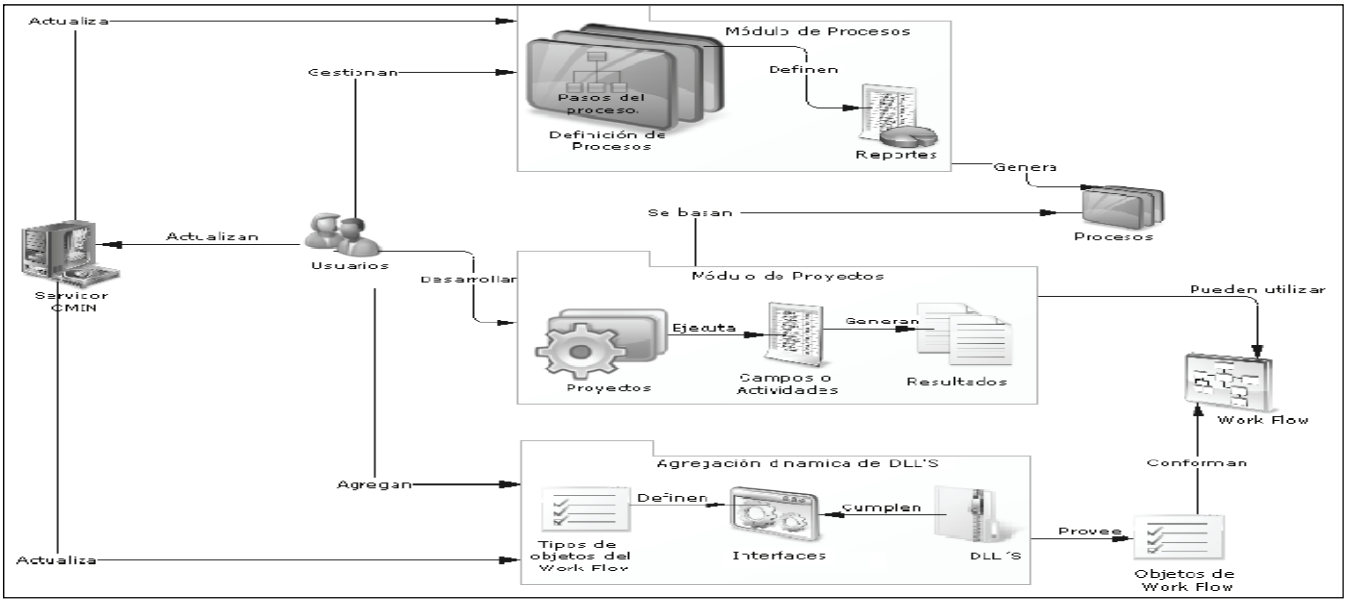

Figura 1. Modelo conceptual de CMIN

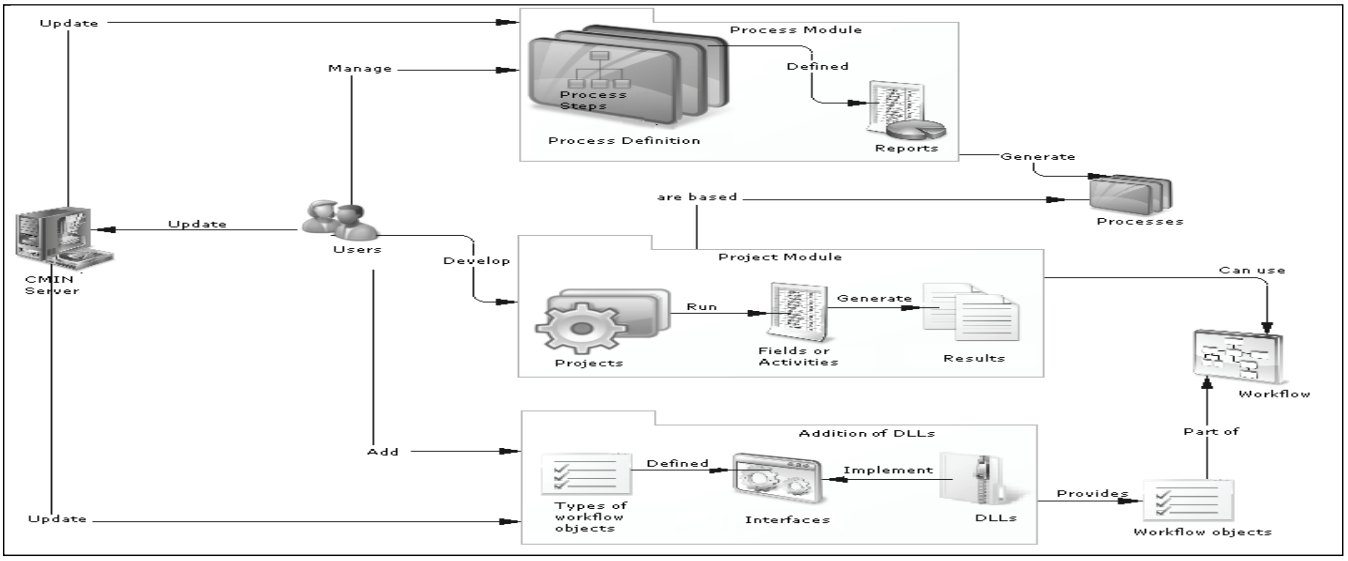

Figure 1. CMIN conceptual model

En español

Casos de uso de la CMIN

En la CMIN se consideran dos tipos de usuario (funciones): *usuarios* finales y *editores expertos* (Figura 2). Los casos de uso del sistema son: entrar al sistema (precondición para usar la herramienta) y gestionar procesos, proyectos, plantillas y DLL. Los usuarios, al ingresar al sistema, deben configurar el servidor de bases de datos de SQL Server con la finalidad de cargar la información necesaria para el funcionamiento del sistema (puede ser una versión *express* que es gratuita). Al gestionar proyectos los usuarios pueden desarrollar los pasos propuestos por el proceso en el que se basa el proyecto, de tal manera que ejecutan los campos que se definen para cada paso, y en algunos campos se puede utilizar el flujo de trabajo (*workflow*) si se necesita utilizar técnicas o algoritmos propios de minería de datos.

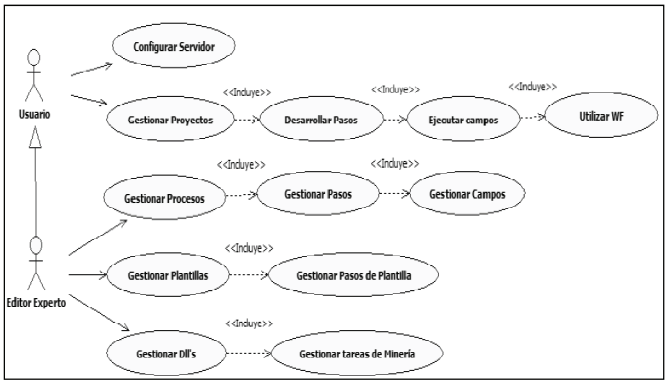

Figura 2. Casos de uso de CMIN.

En la Figura 2 también se presentan los casos de uso de los editores expertos. Estos usuarios, además de usar la funcionalidad de un usuario final, pueden gestionar procesos (crearlos, modificarlos y eliminarlos, y los pasos y campos asociados a ellos), gestionar plantillas (personalizaciones de un proceso en un área específica de aplicación, eliminando pasos que no son apropiados en esa área) y gestionar las librerías de enlace dinámico (DLL) que se utilizan en el sistema. La división de funciones es lógica, ya que la herramienta permite a cualquier usuario desempeñar el papel de editor experto, pero este usuario debe poseer buenos conocimientos de los procesos de minería para definirlos y personalizarlos en plantillas, así como conocer la forma apropiada de crear y cargar nuevos algoritmos de minería de datos en la CMIN. Finalmente, la CMIN cuenta con un conjunto de servicios *web* XML que permiten centralizar nuevos procesos y DLL de algoritmos de minería de datos y estos recursos pueden trasladarse a los clientes con una opción sencilla de sincronización, haciendo que el trabajo del experto sea más sencillo.

Registro de CRISP-DM en CMIN

El módulo de gestión de procesos permite definir nuevos procesos de minería de datos. A continuación se explica de modo general la forma como se registró CRISP-DM V1.0 en la CMIN. Primero el editor experto registra la información básica del proceso (nombre, estado y descripción) y luego define los pasos y campos del proceso. La Figura 3, en el lado izquierdo, despliega un menú contextual que permite crear dichos pasos (fases, tareas genéricas, tareas específicas, etcétera). En cada paso se define su nombre, el tipo de paso en la jerarquía del proceso, una descripción (que sirve de ayuda al usuario de la CMIN) y el conjunto de campos (información que el desarrollador del proyecto de minería de datos deberá registrar en ese paso). En el lado derecho de la figura se ofrece el resultado de la edición de los pasos del proceso CRISP-DM 1.0 seguidos en la CMIN.

In English

CMIN use cases

Two types of users (roles) are considered in CMIN: end users and expert editors (see Figure 2). The system's use cases are as follows: logging into the system (a pre-condition for using the tool), managing processes, managing projects, managing templates and managing DLLs. On logging into the system, the users must configure the database server to SQL server to load the information necessary for the system's operation (possibly an Express version which comes free of charge). When managing projects, users can carry out the steps suggested by the process that the project is using, in such a way that they implement fields that are defined for each step. In some fields, the workflow can be used if the user needs to use particular data mining techniques or algorithms.

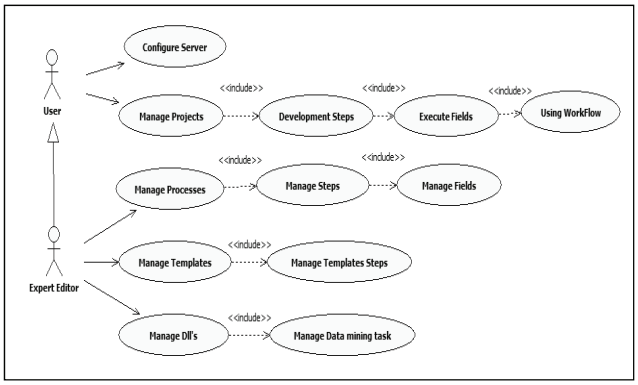

Figure 2. Diagram of CMIN use cases

Figura 2 also shows expert editors' use cases. These users, as well as making use of the functionality available to an end user are also able to manage processes (create, modify and delete processes, and their associated steps and fields), manage templates (customisations of a process in a specific area of application, eliminating steps that are not appropriate in that area) and manage the DLLs used in the system. The division of roles is a logical abstraction, since the tool allows any user to take on the role of expert editor, but such user must have a good knowledge of mining processes to define templates and customise them, as well as learn the proper way to create and load new data mining algorithms in CMIN. CMIN has a set of XML web services that enable the centralisation of data mining algorithms' new processes and DLLs. These resources (processes and algorithms) can be synchronised to customers through a simple synchronisation option, making the job of the expert that much easier.

CRISP-DM register in CMIN

The process management module allows new data mining processes to be defined. The following presents how to register CRISP-DM V1.0 in CMIN. First, the expert editor registers the basic information regarding a process (name, status and description), then defines the steps and process fields. Figure 3 shows, on the left-hand side, how to create a shortcut menu with these steps (phases, generic tasks, specific tasks, etc.). Four things are defined in each step: the name, the type of step in the process hierarchy, a description (which helps the CMIN user) and the set of fields (information that the person carrying out the data mining project must register in that step). The result of editing the steps of CRISP-DM 1.0 registered in CMIN are shown on the right-hand side of the Figure.

En español

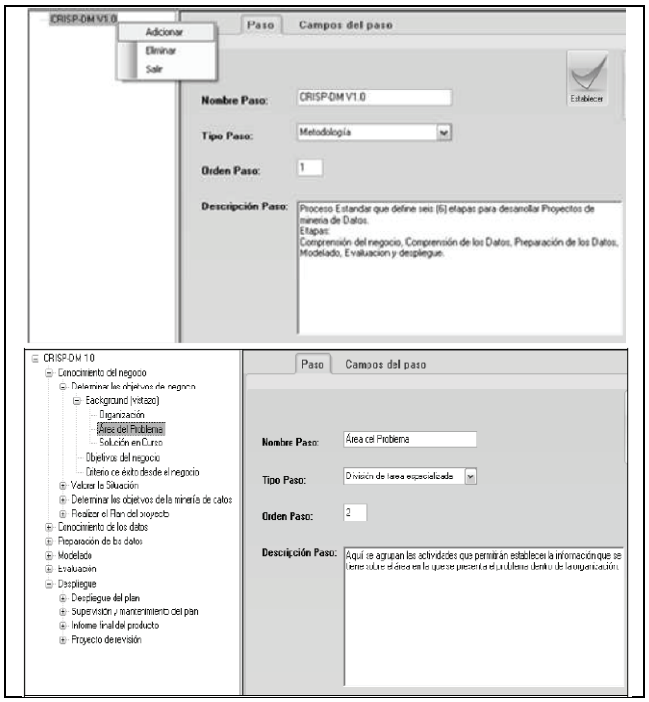

Figura 3. Edición de pasos de un proceso (izquierda) y CRISP-DM en CMIN (derecha)

Después se lleva a cabo la edición de los campos del paso. La Figura 4 contiene un formulario en el que se le solicita al editor o experto en minería el registro de los campos (pueden ser varios) para cada paso. En cada campo se debe incluir una descripción (si es una actividad explica qué se debe hacer, y si es una sugerencia la descripción de ésta); el tipo de campo, que define si es una actividad o sugerencia, y si utiliza workflow (indicando si para realizar la actividad o campo es necesario utilizar el WF).

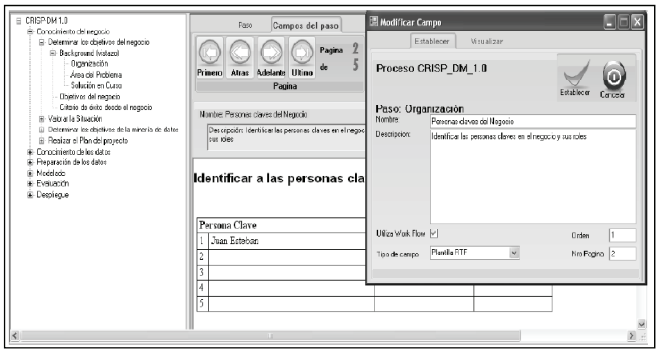

Figura 4. Edición de campos de un paso

Gestión de un proyecto en CMIN

La CMIN permite desarrollar un proyecto de minería de datos basado en un proceso. Para hacer esto, los proyectos heredan la estructura del proceso que el usuario selecciona previamente. En la Figura 5, parte izquierda, se muestra la adición de un nuevo proyecto a la CMIN, lo que implica seleccionar un *proceso base* o una *plantilla* (si se ha definido previamente una), y a su derecha, se presenta el desarrollo de un proyecto. En el numeral (1) se puede observar la estructura del proceso base que es recorrida por el usuario en la medida en que desarrolla el proyecto de minería en la CMIN;

In English

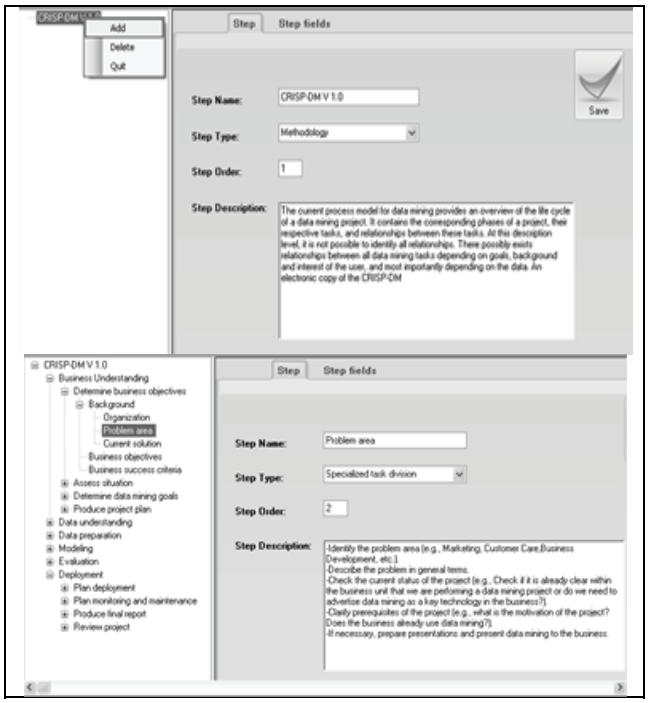

Figure 3. Steps in a process (left) and CRISP-DM in CMIN (right)

Later, the editing of the fields of the step is done. Figure 4 depicts a form that asks the editor or expert in mining to register the various fields (which can be many) for each step. For each field, a *description* must be registered – for example if it is an activity it explains what needs to be done and if it is a suggestion then this is described. The *field type* that defines whether the field is an activity or suggestion is also registered, as is *uses workflow* - indicating whether or not in order to perform the activity or field it is necessary to use the WF.

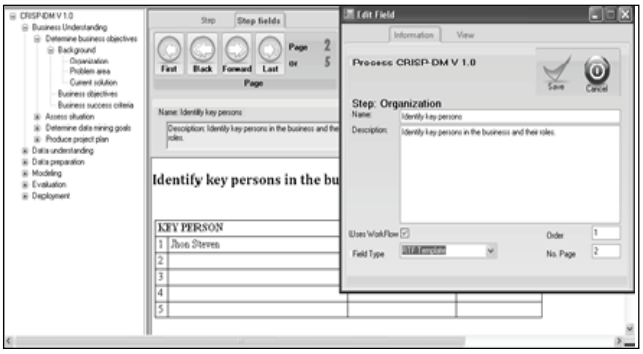

Figure 4. Editing of the fields of the step

Management of a Project in CMIN

CMIN allows a data mining project based on a process to be carried out. In order to do this, the projects inherit the structure of the process that the user selected previously. The left hand part of Figure 5 shows the addition of a new project to CMIN. This process involves selecting a base process or template (if one has been defined previously). The right hand part of Figure 5 shows how a project is conducted. At (1) the structure of the basic process can be seen, which is executed by the user in such a way that the mining project is conducted in CMIN; at (2) the fields or activities to be performed per-

En español

en el (2) se aprecian los campos o actividades a desarrollar pertenecientes al paso en el cual se encuentra; el (3) muestra el botón que guarda la información resultante del campo o actividad; el (4) refiere cómo se puede crear un ciclo de cualquier paso del proceso, siendo esto muy importante, ya que la mayoría de proyectos necesitan reprocesar o repetir ciertos pasos en un momento específico de su evolución; en el (5) se reseña cómo se visualizan los ciclos.

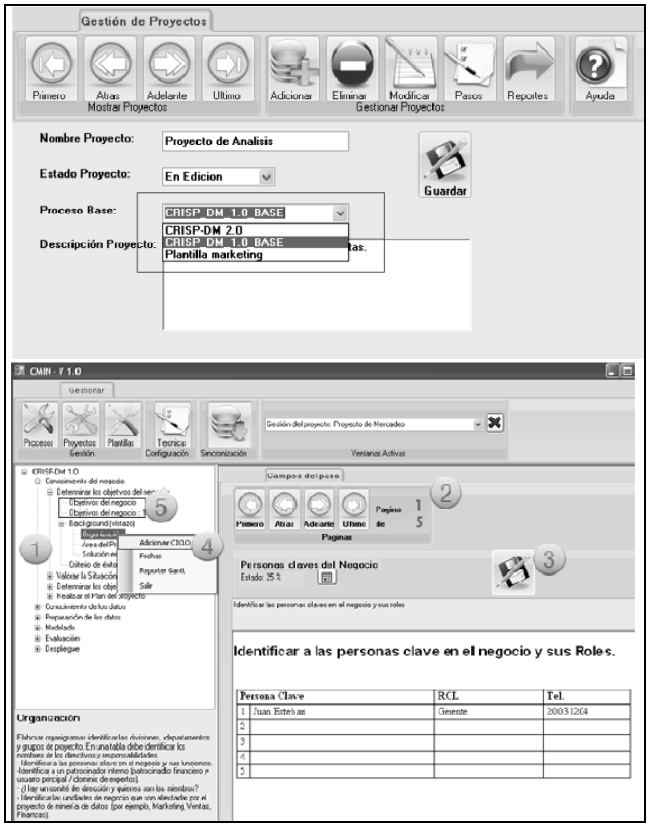

Figura 5. Gestión de proyectos en CMIN

Workflow de minería de datos en CMIN

En la Figura 6 se presenta el *workflow* de la CMIN; el número (1) registra los *tipos de objetos* del *workflow* (fuentes de datos, algoritmos de clasificación, algoritmos de descripción de datos, filtros, visualizadores y algoritmos de agrupamiento o *clustering*); el (2) exhibe un objeto ofrecido del tipo "fuente de datos", y el (3) un objeto en ejecución en el marco del *workflow*.

Figura 6. Workflow de minería en CMIN

Para adicionar algoritmos u objetos a los *tipos de objetos* en tiempo de ejecución, se definió para cada tipo de objeto del *workflow* una interfaz de *software* o contrato (Microsoft-Corporation, 2009a), que agrupa los métodos necesarios para su uso, y otros métodos de interacción con los demás *tipos de objetos del workflow*. Cuando se crea

In English

4) shows how to create a cycle of any step in the process. This last point is very important because most projects need to re-process or repeat certain steps at a specific moment along the way; and (5) indicates how the cycles are displayed.

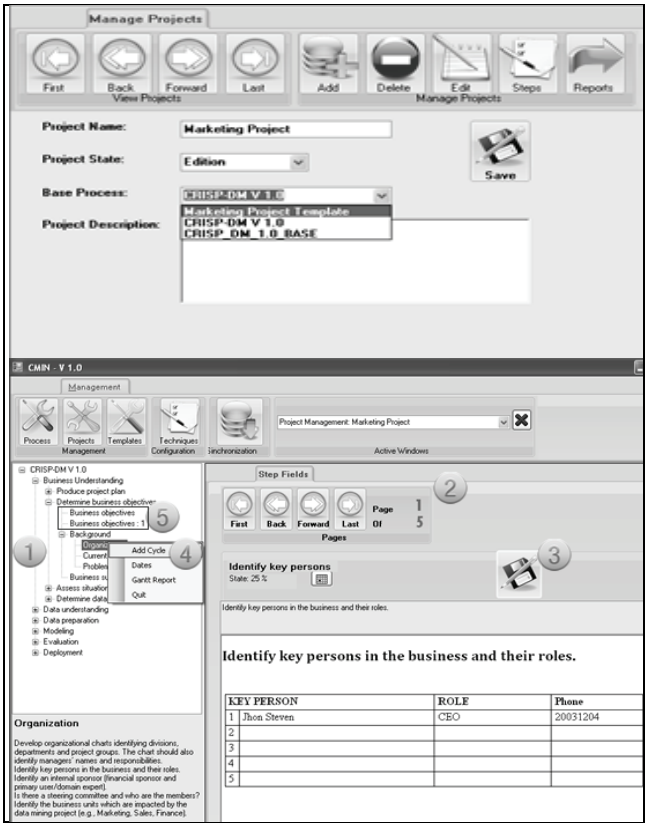

Figure 5. Managing a project in CMIN

Data mining workflow in CMIN

Figure 6 shows the workflow of CMIN. The **types of objects** in the workflow are outlined at (1) (data sources, classification algorithms, data description algorithms, filters, displays, and grouping or clustering algorithms); (2) shows an offered object of the "Data Source" type; and (3) presents an object in execution within the workflow.

Figure 6. Data mining workflow in CMIN

A software interface or contract (Microsoft-Corporation, 2009a) must be defined for each type of object in the workflow to add algorithms, or objects, to types of objects in run time; this groups the methods necessary for its use and other interaction methods with other types of workflow objects. When a new type of object is.

En español

un nuevo tipo de objeto éste se debe reportar a la CMIN con el formulario que se ofrece en el lado izquierdo de la Figura 7.

La interfaz del nuevo tipo se desarrolla previamente con Visual Studio.NET (Chand, 2000), se compila como un ensamblado que se carga en la CMIN. La información del tipo de objeto es almacenado en la base de datos y el archivo ".DLL" es copiado y almacenado en la carpeta local de la CMIN denominada *Assemblies_CMIN*. Después de ingresar el *tipo de objeto* se debe definir con quién se pueden establecer enlaces, es decir, definir qué tipo de objeto puede entregarle información y a qué tipo de objeto se le puede brindar (ver lado derecho de la Figura 7).

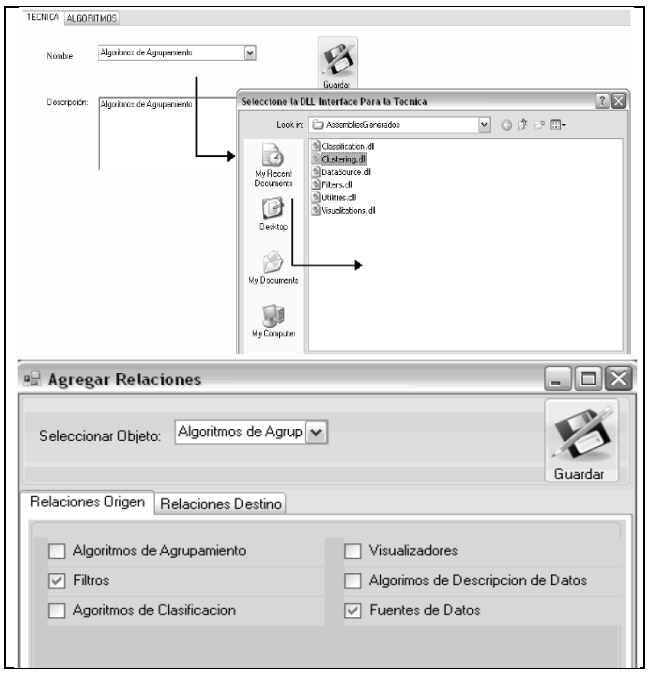

Figura 7. Edición de tipos de objetos (izquierda) y relaciones de los tipos (derecha) en el workflow

Adición de un nuevo algoritmo a la CMIN

El proceso para adicionar un *nuevo objeto* a un *tipo de objeto de CMIN* es el siguiente:

– Un programador crea un proyecto de librería en Visual Studio.NET (Chand, 2000) adicionando como referencia la DLL que define el contrato o interfaz de *software* (Microsoft-Corporation, 2009a) para el tipo de objeto que va a implementar. Es decir, agrega al proyecto la interfaz de *clustering.dll* si va a implementar el algoritmo *k-means* (Figura 8).

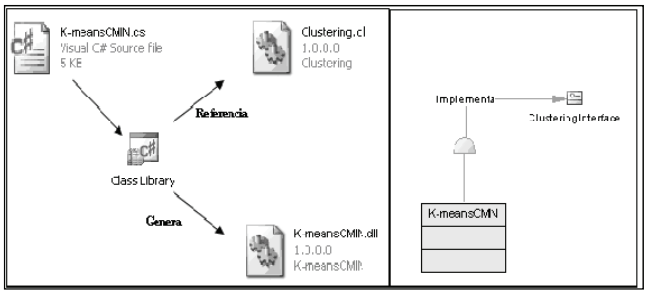

Figura 8. Relación de ensamblados y diagrama de clases dentro del proyecto de librería de VS.NET

In English

created, it should be reported to CMIN using the form seen on the left in Figure 7.

The interface of the new type is developed beforehand using Visual Studio .NET (Chand, 2000); it is compiled as an assembly and this assembly is loaded into CMIN. The information about the object type is stored in the database and the "DLL" file is copied and stored in the local CMIN folder called Assemblies_CMIN. After entering the type of object, the links which can be established must be defined, i.e. define to which type of object you can give information and which type of object can give you information (see the right hand side of Figure 7).

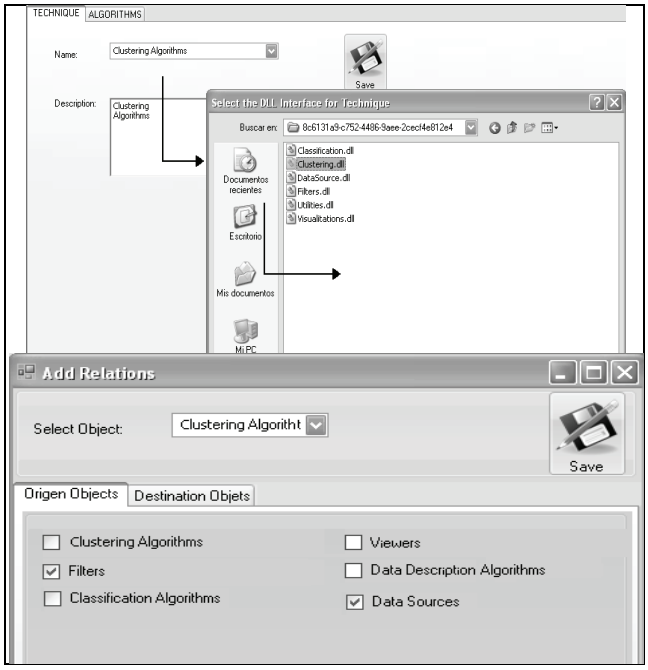

Figure 7. Editing types of objects (left); Relationships among types (right) in workflow

Adding a new algorithm in CMIN

The process for adding a new object (algorithm) to a type of CMIN object is as follows:

A developer creates a library project in Visual Studio .NET(Chand, 2000) adding the DLL that defines the contract or software interface (Microsoft-Corporation, 2009a) as a reference for the type of object that will be implemented. In other words, the developer adds the clustering.dll to the project if the k-means algorithm is going to be implemented (see Figure 8).

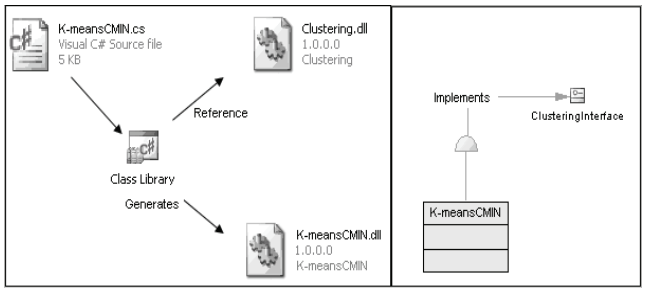

Figure 8. Relationships between assemblies and class diagram in the VS.NET library project

En español

– El programador implementa el algoritmo en el proyecto de librería cumpliendo con el contrato, genera la *nueva DLL* y la comprime en un archivo *.zip* (Figura 9).
– Cuando un usuario necesite usar el nuevo algoritmo en la CMIN primero selecciona el archivo *.zip* con la DLL, luego verifica que cumpla con el contrato —esta comparación se lleva a cabo utilizando reflexión (*System.Reflection*) (Microsoft-Corporation, 2009b) cargando los ensamblados y comparando los métodos—, crea una imagen que represente el nuevo algoritmo y finalmente la carga en la CMIN (Figura 10).

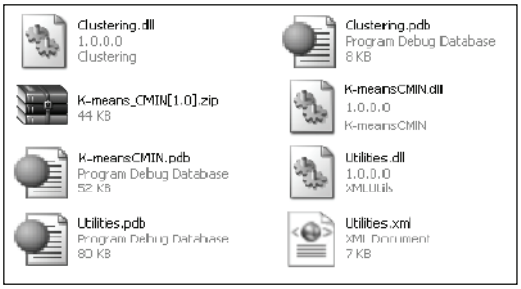

Figura 9. DLL resultado del proyecto de librería

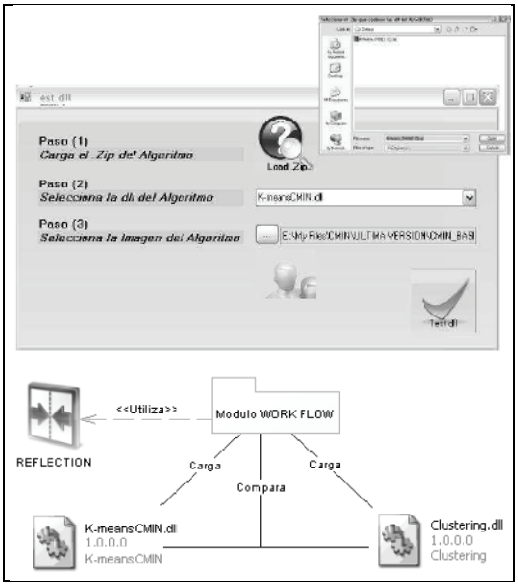

Figura 10. Adición de un nuevo algoritmo (izquierda) y validación de la DLL (derecha) en CMIN

– Si el nuevo algoritmo cumple con la interfaz del *tipo de objeto*, se registra en la base de datos y los archivos del *.zip* son descomprimidos y almacenados en la carpeta local de CMIN denominada *algoritmos*, quedando listo para ser utilizado en el *workflow* (Figura 11).

Invocación de los algoritmos en tiempo de ejecución

Para la invocación de los métodos de los algoritmos que están implementados en las DLL se debe tener en cuenta que la CMIN almacena los ensamblados (*Assemblys*) o DLL de los algoritmos en carpetas locales y que tiene también almacenados los ensamblados de los *tipos de objetos*, es decir, las interfaces. Estos *tipos de objetos del workflow* son estáticos y la parte dinámica la conforman los algoritmos u objetos de cada uno de los tipos, los cuales pueden crecer en tiempo de ejecución. Con este precedente, el grupo definió previa-

In English

– The developer implements the algorithm in the library project (fulfilling the contract), generates the **new DLL** and compresses it in a zip file (see Figure 9);
– When a user needs to use the new algorithm in CMIN, the zip file with the DLL should first be selected, then verified that it complies with the contract - this comparison is done using reflection (System.Reflection) (Microsoft-Corporation, 2009b), loading the assemblies and comparing the methods. An image is then uploaded to represent the new algorithm and finally loaded into CMIN (see Figure 10); and

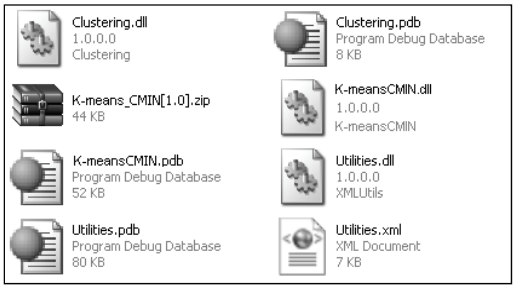

Figure 9. DLL result of the VS.NET library project

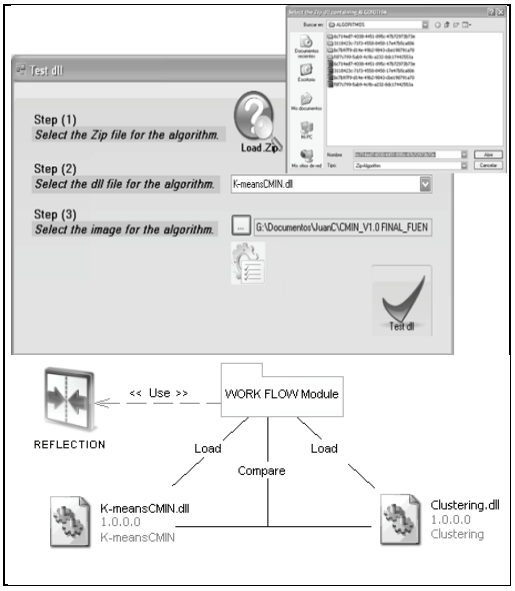

Figure 10. Adding a new algorithm (left) and validating the DLL (right) CMIN

– If the new algorithm meets the requirements of the type of object interface, it is registered in the database and the zip file is decompressed and stored in the local CMIN folder called algorithms, ready to be used in the workflow (Figure 11).

Invoking algorithms in run-time

CMIN stores the algorithm assemblies or DLLs in local folders and it also stores the assemblies of the **types of objects**, i.e. the interfaces. These **types of workflow objects** are static and the dynamic part is made up of the algorithms or objects for each type which can be extended in runtime. Taking this into account, the group first defines software interfaces (contracts) that each type of object must fulfil, focusing on methods allowing algorithm interaction with the user and the CMIN core. This means that the CMIN core (the nerve centre of

En español

mente las interfaces de *software* (contratos) que cada tipo de objeto debía cumplir, teniendo en cuenta métodos que permitieran la interacción de los algoritmos con el usuario y el núcleo de la CMIN. Esto quiere decir que el núcleo de la CMIN, el corazón del *workflow*, funciona basado en la información de las interfaces *software*. El núcleo sabe qué métodos debe invocar en los objetos, ya que ellos cumplen con los contratos de cada tipo de objeto. Para la creación de objetos, la carga y la invocación de los métodos, se usó *reflection* (Microsoft-Corporation, 2009b). Además el núcleo valida las relaciones que se pueden dar entre los objetos basado en las reglas que se registran en la parte derecha de la Figura 7. Como resultado, el *workflow* funciona como se muestra en la Figura 12.

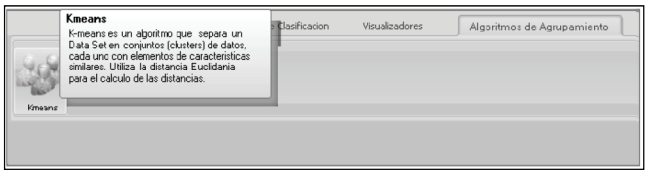

Figura 11. Algoritmo nuevo listo para su uso

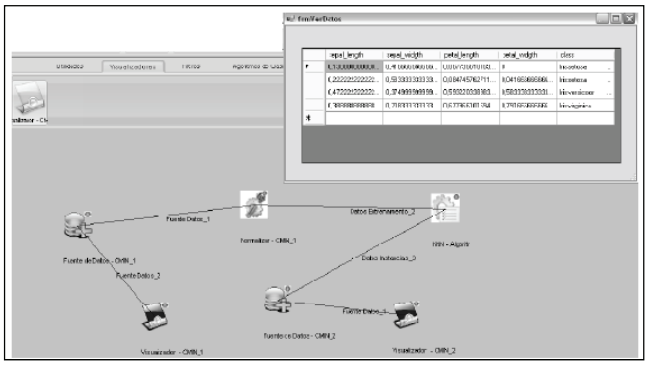

Figura 12. Workflow desarrollado en el taller usando CMIN

Evaluación de la CMIN

La CMIN ha sido sometida a dos evaluaciones:

– *Evaluación preliminar de la gestión de procesos y la gestión de proyectos*. Realizada en febrero de 2008 con 16 estudiantes de la asignatura electiva de minería de datos, en la Universidad del Cauca (UC). En esta evaluación se asignó cada fase de CRISP-DM a dos estudiantes del curso y basados en la versión 1.0 de CMIN realizaron una evaluación general del cumplimiento de las fases de CRISP-DM por parte de la herramienta y además evaluaron la facilidad de uso de ésta. Como conclusión general, la herramienta cumplió en un 100% con CRISP-DM, pero se detectó la necesidad de mejorar algunas plantillas de recolección de información en algunas fases. Teniendo en cuenta los resultados positivos de dicha evaluación, en marzo de 2008 se participó en una convocatoria de proyectos a ser presentados en el Demofest del Microsoft Research Academic Summit. En el proyecto, seleccionado por Microsoft, se presentó un póster científico de CMIN el 16 de mayo de 2008 en Ciudad de Panamá y se ofreció directamente la herramienta a los profesores e investigadores que participaron en el evento. A pesar de que en el Demofest se presentaron proyectos con inversiones muy superiores a la hecha por la CMIN, el proyecto recibió excelentes comentarios y Microsoft lo incluyó en una nota publicitaria que se presentó en el programa *Adelantos*, de *CNN en español* (ver copia del video en http://www.unicauca.edu.co/~ccobos/cnn-adelantos.wmv).

In English

the workflow) functions in a way that is based on the information from the software interfaces. The core knows which methods it must invoke on the objects so that they comply with the contracts for each type of object. For creating and loading objects and invoking methods, the core uses **reflection** (Microsoft-Corporation, 2009b). The core also validates the relationships that can occur between objects, based on the rules presented in the right-hand part of Figure 7. As a result, the workflow functions as shown in Figure 12.

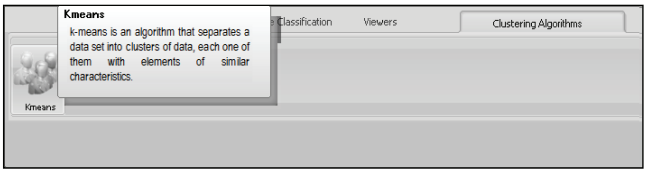

Figure 11. New algorithm ready to be used

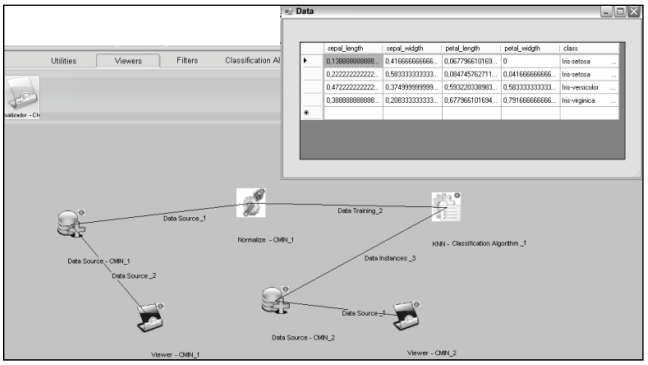

Figure 12. Workflow developed in the workshop using CMIN

CMIN assessment

CMIN has undergone two evaluations:

– A preliminary assessment of process management and project management was held in February 2008 with sixteen students from the University of Cauca's optional Data Mining course. In this evaluation each CRISP-DM phase was assigned to two students on the course. Based on version 1.0 of CMIN, they made an overall assessment of the tool's compliance with CRISP-DM phases and also evaluated the ease of use thereof. As a general conclusion, the tool fulfilled CRISP-DM requirements 100% although some templates for collecting information in some phases needed to be improved. Given the positive results of this evaluation, a description of the tool was sent to a project presentation meeting in March 2008 to be submitted to the Demofest of the Microsoft Research Academic Summit. The project was selected by Microsoft and a scientific poster on CMIN was presented in Panama City on May 16th 2008. A tool was presented in person to the teachers and researchers who attended the event. Despite the fact that many projects presented at the Demofest boasted investments much higher than that of CMIN, the tool received excellent reviews and Microsoft decided to include it in publicity that appeared on CNN television (Spanish language) in their program ADVANCES (see copy of the video http://www.unicauca.edu.co/~ccobos/cnndelantos.wmv).

En español

– *Evaluación de la usabilidad de la herramienta*. Esta evaluación fue hecha en marzo de 2009 con una prueba beta donde participaron ingenieros y estudiantes del programa de Ingeniería de Sistemas de la UC que trabajan en minería de datos. Esta prueba tuvo dos objetivos: la revisión completa de la CMIN en un ambiente diferente al de desarrollo, a través de un test de usabilidad, y verificar con un experimento si mediante el uso de la CMIN se podía mejorar el conocimiento que los usuarios tienen de CRISP-DM. El experimento se efectuó en seis pasos, de la siguiente manera: 1) aplicación de un test previo para valorar los conocimientos del grupo sobre CRISP-DM; 2) presentación básica de la herramienta CMIN; 3) desarrollo de un taller de minería de datos (consistente en resolver un problema típico de clasificación, para el cual se seleccionó el *data set* IRIS disponible en el repositorio de la UCI (Asuncion & Newman, 2007), mientras que los usuarios emplearon el *workflow* y obtuvieron el resultado desplegado en la Figura 12; 4) interacción con el grupo a través de preguntas y sugerencias; 5) aplicación de un test posterior para valorar el nuevo nivel de conocimientos del grupo sobre CRISP-DM (el contenido del test no cambió con respecto al del paso 1); y 6) aplicación de un test de usabilidad basado en un cuestionario de la Universidad Politécnica de Cataluña (Borges de Barros Pereira, 2002).

En términos generales la prueba fue exitosa, ya que la herramienta no tuvo errores y todos los participantes lograron resolver el problema de clasificación presentado. Los resultados del test de usabilidad fueron muy buenos. Se puede afirmar que la CMIN cuenta con una interfaz amigable, entendible y, sobre todo, que el manejo de los proyectos que contemplan aspectos repetitivos y en cierta medida complejos pueden ser manejados con facilidad. La interfaz minimiza lo que el usuario debe aprender y en cada paso lo orienta para llevar a feliz término cada una de las tareas correspondientes a un proyecto de minería de datos. En la Figura 13 se indican los principales resultados del test de usabilidad, donde los usuarios expresan para cada uno de los indicadores de evaluación una valoración mayoritariamente excelente y buena.

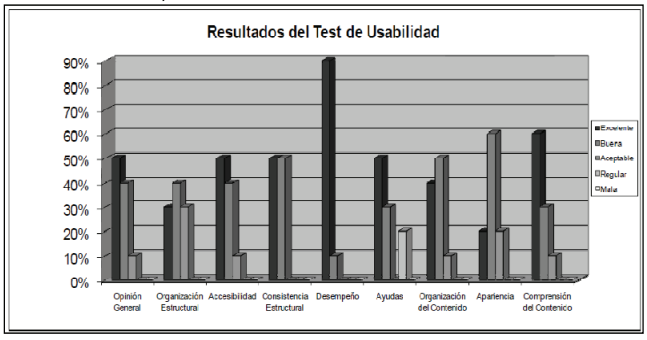

Figura 13. Principales resultados del test de usabilidad

En cuanto al test de conocimiento sobre CRISP-DM se logró un incremento del 5 al 10% en el conocimiento del proceso en el poco tiempo que duró el taller (1 hora), destacando que éste no tenía como objetivo que los usuarios memorizaran las fases, las tareas genéricas y específicas de CRISP-DM y, lo más importante de esto, el cambio en los términos de las respuestas dadas por los usuarios en el test posterior, las cuales fueron más precisas, técnicas y directamente relacionadas con las fases del proceso.

Conclusiones y trabajo futuro

La CMIN es una herramienta CASE integrada que orienta el desarrollo de los proyectos a través de procesos, facilita la integración del

In English

– An evaluation of the usability of the tool. This evaluation was carried out in March 2009 using a Beta test with the participation of the University of Cauca (UC) Engineers and Systems Engineering students who work in data mining. This test had two objectives: a thorough revision of CMIN in a different environment to that of its development, by way of a usability test, and the verification (through an experiment) of whether or not using CMIN could increase the knowledge users had of CRISP-DM. The experiment was conducted in six steps, as follows: 1) a pre-test evaluated the group's initial knowledge of CRISP-DM; 2) a basic presentation of the CMIN tool was given; 3) a workshop on data mining was held (the aim of the workshop was to set a typical classification problem for the group to solve. The IRIS data set - available from the UCI repository (Asuncion & Newman, 2007) - was selected for the workshop. The participants used the workflow and obtained the result shown in Figure 12.); 4) interaction with the group was done by questions and suggestions; 5) a further test was taken, to evaluate the group's new level of knowledge regarding CRISP-DM (the content of this test did not change regarding the pre-test); and 6) a usability test was set, based on a questionnaire from the Universidad Politécnica de Cataluña (Borges de Barros Pereira, 2002).

Overall, the test was successful in that the tool did not throw up any errors while all participants were able to resolve the classification problem presented. The usability test results were very good. CMIN can be said to have a friendly interface that is understandable and through which – most importantly - the management of projects that may involve repetitive and somewhat complex aspects can be handled easily. The interface minimizes what the user needs to learn in the tool. At each step it provides guidance for successfully carrying out data mining project tasks. Figure 13 shows the main results of usability testing wherein, for each indicator, the users expressed an assessment mainly consisting of excellent and good.

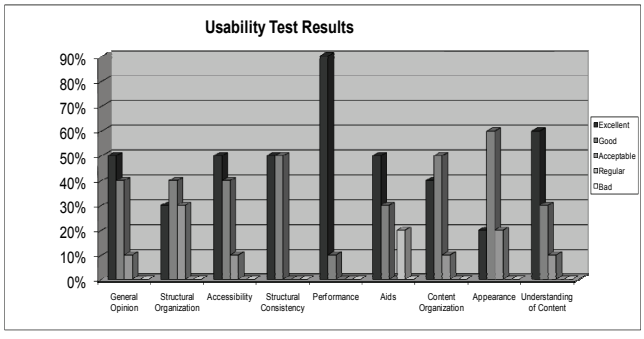

Figure 13. Main results of usability test

As regards the CRISP-DM knowledge test, an increase of between 5% and 10% in knowledge of the process was achieved in the short period of the workshop (1 hour), noting that it was not intended that users memorize CRISP-DM phases and its generic and specific tasks. Most important was the change seen in the terminology in test users' responses. Compared to the pre-test, responses proved to be more accurate, more technical and more directly related to the phases of the process.

Conclusions and future work

CMIN is an integrated CASE tool that guides the carrying out of projects through processes, facilitates the integration of the process with

En español

proceso con el proyecto y asegura el cumplimiento del proceso en la ejecución del proyecto; su funcionalidad extensible (ampliación dinámica y en tiempo de ejecución de la batería de algoritmos) motiva y facilita el desarrollo en comunidad, ya que una nueva funcionalidad puede ser programada por miembros de la comunidad, y después puede ser probada y evaluada por un grupo de expertos y finalmente incluida y distribuida a los demás miembros de la comunidad de usuarios de la herramienta a través de la opción de sincronización. Mediante la información detallada y apropiada en cada paso de un proceso y de un proyecto en la CMIN se posibilita que el usuario conozca progresivamente sobre un proceso de minería de datos (por ejemplo, CRISP-DM).

Como trabajo futuro, el grupo de investigación planea implementar una versión mejorada del componente de seguimiento a proyectos que tenga en cuenta la administración de los recursos para cada actividad, de tal forma que se puedan hacer reportes de costos en cada paso del proyecto y en general, integrar a la CMIN una metodología de gestión de proyectos; además, centrar esfuerzos en el establecimiento de la comunidad que permita un rápido crecimiento de la batería de algoritmos que se puedan usar en la CMIN y potenciar de esta forma el uso del *workflow*.

In English

the project and ensures the process's compliance in the execution of the project. CMIN is a tool with expandable functionality (capable of dynamic extension of the algorithm array in runtime) that encourages and facilitates cooperation within the development community, as new functionality can be programmed by community members, then tested and evaluated by a panel before being finally included and distributed to other members of the tool user community through the synchronisation option. Using detailed and appropriate information in each step of any process or in any project in CMIN, it is likely that the user will progressively come to know more about any data mining process (for example, CRISP-DM).

Regarding future work, the research group plans to implement an improved version of the component for project monitoring that takes into account the management of resources for each activity. Cost reports can thus be produced for each step of the project; the group thus recognises the need for integrating suitable project management methodology within CMIN. Additionally, the intention is to focus efforts on building up the tool development community. This ought to allow rapid growth in the existing battery of algorithms that can be used in CMIN and thus enhance workflow use.

Bibliografía / References

Asuncion, A., Newman, D. J., UCI Machine Learning Repository 2008., 2007. from http://www.ics.uci.edu/~mlearn/ML Repository.html

Borges de Barros Pereira, H. Análisis experimental de los criterios de evaluación de usabilidad de aplicaciones multimedia en entornos de educación y formación a distancia Unpublished Doctoral., Universitat Politecnica de Catalunya, Barcelona, 2002.

Britos, P., Fernández, E., Ochoa, M., Merlino, H., Diez, E., García, R., Metodología de Selección de Herramientas de Explotación de Datos., Paper presented at the II Workshop de Ingeniería del Software y Bases de Datos. XI Congreso Argentino de Ciencias de la Computación, 2005.

CRISP-DM., CRoss Industry Standard Process for Data Mining., 2006. from http://www.crisp-dm.org/

Chand, M., Creating C# Class Library (DLL) Using Visual Studio .NET [Electronic Version]., C# Corner, (2000). from http://www.c-harpcorner.com/UploadFile/mahesh/dll12222005 064058AM /dll.aspx

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., CRISP-DM 1.0: Step-by-step data mining guide: CRISP-DM Consortium., 2000.

Gondar Nores, J.-E., Metodologías para la Realización de Proyectos de Data Mining [Electronic Version]., 2004. from http://www.estadistico.com/arts.html?20040426

Holmes, G., Donkin, A., Witten, I. H., WEKA: a machine learning workbench., Paper presented at the Intelligent Information Systems,1994., Proceedings of the 1994 Second Australian and New Zealand Conference on, 1994.

INEI., Herramientas CASE. Lima, Perú: Instituto Nacional de Estadística e Informática., 1999.

Insightful-Corporation., Insightful Miner., from http://www.insightful.com/products/iminer/default. asp

Kdnuggets., Tools data mining., 2005. from http://www. kdnuggets.com/polls/2005/data_mining_tools.htm

Khabaza, T., Shearer, C., Data mining with Clementine., Paper presented at the Knowledge Discovery in Databases, [IEE Colloquium on], 1995.

Mai, C. K., Krishna, I. V. M., Reddy, A. V. Polyanalyst application for forest data mining., Paper presented at the Geoscience and Remote Sensing Symposium, 2005, IGARSS '05. Proceedings. 2005 IEEE International, 2005.

Megaputer., PolyAnalyst 6.0 - simplify your analytics., 2009. from http://www.megaputer.com/

MetaGroup., METAspectrum Market Summary., 2004. from http://www.oracle.com/technology/products/bi/odm/pdf/odm_m etaspectrum_1004.pdf

Microsoft-Corporation., interface (C# Reference), 2009a. from http://msdn.microsoft.com/en-us/library/87d83y5b.aspx

Microsoft-Corporation., Reflection Overview [Electronic Version]. .NET Framework Developer's Guide., 2009b. from http://msdn.microsoft.com/en-us/library/f7ykdhsy.aspx

Miren Begoña, A.-R., A retrospective view of CASE tools adoption., SIGSOFT Softw. Eng. Notes, 25(2), 2000, pp. 46-50.

Rippa, S., Lendyuk, T. Selection of Alternative Projects Using Data Mining., Paper presented at the 4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS, 2007.

Salford-System., Classification And Regression Trees (CART)., 2009. from http://www.salfordsystems.com/cart.php

SAS., Data mining with SAS® Enterprise Miner., 2009a. from http://www.sas.com/technologies/analytics/datamining/miner/

SAS. SAS Enterprise Miner – SEMMA., 2009b. from http://www.sas.com/offices/europe/uk/technologies/analytics/da tamining/miner/semma.html

SPSS-Inc., Clementine., 2009. from http://www.spss.com/es/ clementine/

University-of-Waikato., Weka 3: Data Mining Software in Java., 2009. from http://www.cs.waikato.ac.nz/ml/weka/