Una Aproximación a la Asistencia Inteligente en Minería de Datos

Matías A. Nicoletti[1], Héctor O. Nigro[2] y Sandra E. González Císaro[2]

[1] ISISTAN *Research Institute* - Fac. de Ciencias Exactas - UNCPBA - Tandil, Argentina
[2] INTIA - Fac. de Ciencias Exactas - UNCPBA - Tandil, Argentina
{mnicolet, onigro, sagonci}@exa.unicen.edu.ar

Resumen Extendido

La Minería de Datos, la cual forma parte del proceso de KDD (*Knowledge Discovery in Databases*), es uno de los campos de crecimiento más veloz en la industria de la computación. Una de sus principales fortalezas es su amplitud en cuanto a los posibles dominios de aplicación. La industria bancaria, telecomunicaciones, ventas, medicina, deportes y variedad de ingenierías son algunos ejemplos de su aplicación efectiva [4]. En la actualidad existen gran variedad de herramientas para facilitar su práctica, entre las cuales pueden encontrarse Weka, Orange y KNIME, de libre distribución.

A pesar de que dichas herramientas suponen facilitar la práctica de esta disciplina, definir y ejecutar procesos de Minería de Datos no resulta ser una tarea trivial. Esto es debido a que numerosas variables y decisiones a ser tenidas en cuenta se ven involucradas durante el proceso. Adicionalmente, las actuales aplicaciones para KDD proveen extensos conjuntos de operadores (por ej. en KNIME existen más de 500 operadores actualmente). Pero la complejidad en la Minería de Datos no solo se debe a la extensa cantidad de operadores, sino que también debe considerarse la correcta combinación de dichos operadores respectando las restricciones asociadas. De esta forma, es posible establecer una analogía entre la Minería de Datos y la resolución de rompecabezas [4]. Con el tiempo, los usuarios aprenderán y se familiarizarán con el adecuado uso de los operadores. No obstante, investigar formas de acelerar la curva de aprendizaje motiva el desarrollo de este proyecto.

En ciertas ocasiones, la construcción de procesos se convierte en una tarea repetitiva, sobre todo cuando el objetivo de los mismos es similar. El registro y análisis de las combinaciones entre técnicas y de los procesos que se construyen en el tiempo podrían permitir la implementación de un esquema de asistencia basada en experiencias pasadas. Además, definir modelos de los criterios que condicionan la combinación de los diferentes operadores podría complementar dicho esquema. No obstante, las actuales herramientas carecen de estas características que podrían facilitar el trabajo de los analistas.

Bajo este contexto, se propone la implementación de un esquema de asistencia basado en agentes inteligentes de interfaz y técnicas de Inteligencia Artificial que proporcione ayuda al usuario de las herramientas de Minería de Datos actualmente utilizadas. El agente podrá organizar la creación de procesos y asistir al usuario, inexperto o no, en diversos contextos de la construcción. Una de las características más interesantes de este enfoque es que a pesar de tratarse de asistencia orientada a las aplicaciones, no depende de la herramienta específica donde se aplique. Esto es posible dado que se diseñó un *framework* orientado a objetos que abstraiga las características particulares de cada herramienta. De esta forma, se implementó una solución particular para KNIME (denominada *KAgent*) que permite la evaluación del esquema y sirve como guía para la correcta utilización del *framework*.

Los agentes inteligentes de interfaz son entidades de software que tienen el objetivo de asistir a usuarios en sus tareas basadas en computadoras. A través del aprendizaje de sus hábitos de trabajo y de sus preferencias, estas entidades pueden mejorar la productividad de los usuarios e incluso reducir su carga de trabajo [8]. La inteligencia en las aplicaciones puede identificarse sencillamente, cuando las mismas exhiben comportamiento similar al razonamiento humano. Lógicamente, un agente no puede alcanzar la inteligencia humana, ya que la misma es extremadamente compleja como para ser modelada en un software. Sin embargo, existen aproximaciones donde se consiguen emular algunas características, como por ejemplo el razonamiento, la inferencia o conocimientos específicos para ciertos dominios.

En el contexto de este trabajo, el conocimiento que poseen los usuarios experimentados claramente le resulta de utilidad al novato, y comúnmente se ve desperdiciado. Dentro de dicho conocimiento se pueden identificar a) restricciones conceptuales entre los operadores (de carácter estático), b) restricciones contextuales entre los operadores (referidas al tipo de datos), c) combinaciones entre operadores comúnmente utilizados, y d) procesos completos que han demostrado ser útiles y podrían reutilizarse en un futuro.

En los casos a) y b), se diseñó un esquema flexible en cuanto a la definición de nuevas restricciones, soportando actualmente los dos tipos mencionados. Para a), el usuario experto puede definir la restricción entre dos operadores que conceptualmente no pueden combinarse, y que deben restringirse bajo cualquier contexto. En b), los usuarios expertos deben definir los tipos de datos (nominal, discreto, etc.) que un operador acepta, de modo que el asistente interprete del contexto si dicho operador puede ser utilizado.

En c), el asistente observa como los expertos componen los operadores y genera un modelo predictivo en base al historial de combinaciones. En este caso, se consideró el uso de Cadenas de Markov, realizando una implementación

similar a la del estudio de [2]. Posteriormente, el agente sugiere la utilización del conjunto de operadores más comúnmente utilizados bajo el actual contexto de construcción, generando un ranking por relevancia.

En d), los usuarios con experiencia desean almacenar los procesos que han sido exitosos, lo cual es una funcionalidad soportada por las herramientas actuales. Sin embargo, la información generada queda en archivos externos a la aplicación y difícilmente se reutilice. El asistente proveerá un mecanismo de importación / exportación propio que permita catalogar los proyectos generados. En base a esta información, el agente analizará el proceso que se construya en la interfaz y, ocasionalmente, sugerirá al usuario que revise uno de los que se encuentran almacenados, en caso de detectarse un alto grado de similitud entre ambos. Además se tomará en cuenta el feedback o retroalimentación de los usuarios para mejorar sus criterios de recomendación y poder ajustarse a los perfiles de los usuarios. Esta característica es posible gracias al uso del método AHP [6] que le otorga flexibilidad al algoritmo de comparación entre procesos.

Analizando los trabajos relacionados, los agentes inteligentes han sido ampliamente utilizados en diversas áreas, sobre todo cuando se requiere manejar grandes cantidades de información o ejecutar tareas repetitivas, como filtrado de información, búsqueda y navegación en la Web, manejo de e-mail, planificación de reuniones y comercio electrónico. Sin embargo, el uso de agentes para asistir el desarrollo de proyectos de Minería de Datos es un campo realmente poco explorado. Por el contrario, la Minería de Datos ha sido utilizada ampliamente en el diseño de sistemas inteligentes, como por ejemplo en [5]. Considerando esto, la asistencia de agentes inteligentes en el desarrollo de proyectos de Minería de Datos resulta ser un enfoque novedoso y de notable valor práctico.

Actualmente existen líneas de investigación sobre el uso de ontologías para la generación de procesos en Minería de Datos [7]. Uno de los proyectos que podemos encontrar es e-LICO, en donde existe el sub proyecto DMOP (Ontologías en Minería de Datos para la Optimización de Flujos de Trabajo) [3]. Entre las bases de este proyecto, se encuentra la investigación de [1], en donde se introduce el concepto de Asistentes Inteligentes de Descubrimiento (IDA) que permiten la generación semi-automática de procesos válidos y su posterior priorización, a través de la definición de una ontología. En comparación con nuestro enfoque, el propósito general de ambas propuestas difiere, ya que nuestro esquema propone una guía paso a paso para la asistencia y educación de los usuarios. Además, la transferencia de conocimiento se lleva a cabo de formas distintas.

En este trabajo, hemos presentado un esquema de asistencia inteligente para usuarios de herramientas de Minería de Datos, basado en la captura y modelado del conocimiento de usuarios expertos, que comúnmente se desperdicia.

En base a los resultados preliminares obtenidos con *KAgent* para KNIME, se supone que el asistente conseguiría simplificar la curva de aprendizaje de los usuarios novatos y que los modelos se comportarían de la forma esperada. En este momento estamos realizando pruebas sobre estudiantes de Ingeniería en Sistemas que toman el curso optativo *Data Mining*, de forma de poder evaluar las mejoras introducidas por el uso de las herramientas con y sin el asistente sobre usuarios novatos. La evaluación consiste en realizar una encuesta a los alumnos de la materia sobre su experiencia con *KAgent* durante la resolución del trabajo práctico de la cursada, donde debe resolverse un problema de Minería de Datos utilizando KNIME.

Como trabajo futuro, podría a) realizarse mayor experimentación para comprobar la utilidad efectiva del enfoque, b) analizarse qué otras formas de conocimiento resultan útiles para los usuarios novatos y con qué modelos podrían representarse, c) rediseñarse el esquema para que el conocimiento pueda compartirse fácilmente entre distintos usuarios de la herramienta, por ejemplo mediante el uso algún estilo arquitectónico distribuido, y d) analizarse la posibilidad de realizar una combinación entre el asistente basado en ontologías [1] y el desarrollado por nosotros. Particularmente de d) podría resultar un esquema interesante en donde el asistente le presentará al usuario uno o más planes válidos teniendo en cuenta i) el tipo de proceso a realizarse (*clustering*, clasificación, etc.), ii) su perfil y iii) un indicador de usabilidad (tomado de las observaciones realizadas por el agente y plasmadas a través del método AHP).

Referencias

1. A. Bernstein, F. Provost, and S. Hill. Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):503–518, 2005.
2. B. Dasison and H. Hirsh. Predicting sequences of user actions. In *AAAI/ICML Workshop on Predicting the Future: AI Approaches to Time Series Analysis*, 1998.
3. M. Hilario, A. Kalousis, P. Nguyen, and A. Woznica. A data mining ontology for algorithm selection and meta-learning. In *Proc. ECML/PKDD Workshop on Third-Generation Data Mining: Towards Service-Oriented Knowledge Discovery (SoKD-09)*, 2009.
4. M. Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms*. Number 0471228524. Wiley-IEEE Press, 1 edition, October 2002.
5. Y. Kim and W. N. Street. An intelligent system for customer targeting: a data mining approach. *Decision Support Systems*, 37(2):215–228, 2004.
6. M. Mochol, A. Jentzsch, and J. Euzenat. Applying an analytic method for matching approach selection. In *Ontology Matching*, volume 225 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2006.
7. H. O. Nigro, S. E. G. Cisaro, and D. H. Xodo. *Data Mining with Ontologies: Implementations, Findings and Frameworks*. Number 978-1599046181. Idea Group Reference, 2008.
8. S. N. Schiaffino and A. Amandi. User - interface agent interaction: personalization issues. *Int. J. Hum.-Comput. Stud.*, 60(1):129–148, 2004.