Un marco para democratizar la minería de datos: propuesta inicial y retos

Diego García-Saiz[1], Roberto Espinosa[2], José Jacobo Zubcoff[3], José-Norberto Mazón[4],Marta Zorrilla[2]

[1] Dpto. de Ingeniería Informática y Electrónica, Universidad de Cantabria, España
{diego.garcias,marta.zorrilla}@unican.es
[2] WaKe Research, Universidad de Matanzas, Cuba
roberto.espinosa@umcc.cu
[3] WaKe Research, Dpto. Ciencias del Mar y Biología Aplicada, Universidad de Alicante, España
jose.zubcoff@ua.es
[4] WaKe Research, Dpto. Lenguajes y Sistemas Informáticos, Instituto Universitario de Investigación Informática, Universidad de Alicante, España
jnmazon@dlsi.ua.es

Abstract. Movimientos como el de datos abiertos posibilitan que cada vez haya una mayor disponibilidad de datos accesibles para su reutilización. A pesar de que el número de herramientas analíticas que están a nuestra disposición crece cada día, lamentablemente ninguna permite realizar un proceso de extracción de conocimiento directo a usuarios con poca o nula experiencia en el uso de la estadística y de algoritmos de minería de datos. En este artículo se presenta una aproximación a un marco KaaS (*Knowledge as a Service*) que posibilite a usuarios no expertos la extracción de conocimiento a partir de un conjunto de datos. Se muestra que la propuesta es viable y se plantean los retos aún abiertos.

Keywords: Conocimiento como Servicio, Minería de datos, Analítica, Meta-aprendizaje

1 Introducción

Actualmente estamos inmersos en la era de la "datificación", esto es, la era de la generación de datos constante (redes sociales, sitios Web, sensores desplegados, dispositivos móviles). Además, muchos de estos datos están disponibles para su reutilización debido a la, cada vez más importante, tendencia de exponer los datos como datos abiertos. Este escenario "*big data*" abre un gran abanico de posibilidades analíticas, ya que ahora es posible utilizar técnicas algorítmicas eficientes para procesar y analizar estos datos con el fin de extraer conocimiento novedoso, útil y que directamente pueda ser traducido en acciones concretas. Sin embargo, "*big data*" también conlleva un problema: abre una nueva brecha digital entre aquellas personas que saben cómo analizar datos y aquéllas que no. Es

preciso, por ello, investigar mecanismos para acorten esta brecha, y que democraticen la minería de datos para que cualquier ciudadano pueda aprovecharse del ingente volumen de datos disponible.

El proceso de extracción del conocimiento (de sus siglas *knowledge discovery process*) conlleva varios pasos [4]: 1) traducir el objetivo en un problema de minería de datos, 2) seleccionar y preprocesar las fuentes de datos para el objetivo, 3) seleccionar las técnicas para resolver el tipo de problema, 4) construir y evaluar modelos y 5) mostrar al usuario final el resultado de forma interpretable. Cada una de estas fases, a su vez, conlleva la posibilidad de realizar varias subtareas de forma alternativa, por lo que se puede decir que se trata de un proceso iterativo que se finaliza cuando se alcanza el fin del proyecto o del dinero.

Nuestra propuesta se basa en la creación de un marco KaaS (*Knowledge as a Service*) en el cuál un usuario no experto pueda utilizar servicios de minería de manera sencilla, esto es, un servicio que automatice y oculte las fases mencionadas. Esta propuesta extiende y generaliza el servicio de minería *Elearning Web Miner* (ElWM) que permite a los profesores implicados en enseñanza virtual extraer patrones de comportamiento y de actividad de los alumnos matriculados en sus cursos [9], construir modelos predictivos sobre el rendimiento de sus alumnos y modelos descriptivos que caractericen a sus estudiantes desde un punto de vista social [5]. Esta herramienta está planteada para su uso por profesores totalmente inexpertos en minería de datos y está diseñada en base a preguntas fijas a las que se ofrece respuesta utilizando un fichero de datos y una técnica de minería que experimentalmente ofreció mejor resultado. Por tanto, el objetivo de este nuevo marco es permitir al usuario especificar sus requisitos y que el servicio seleccione el algoritmo que conlleve a la solución más precisa para el conjunto de datos bajo análisis.

2 Descripción del marco

La arquitectura que se propone se recoge gráficamente en la Fig. 1. Como se puede observar, la idea es que el usuario no experto suba al servicio de minería un fichero con los datos a analizar y que, a continuación, este le realice preguntas que le ayuden a reconocer el tipo de problema que quiere resolver. A partir de ahí el servicio se encarga de procesar los datos, elegir la técnica más adecuada haciendo uso de un recomendador, contruir el modelo y ofrecérselo al usuario en forma de reglas o mediante representaciones gráficas fácilmente interpretables. Por tanto, nuestro marco presenta tres módulos esenciales cada uno de los cuales está asociado a uno o varios retos que se deben abordar:

1. Extracción de requisitos: el usuario no experto debe ser capaz de definir sus requisitos de minería sin necesidad de tener conocimientos de este campo.
2. Construcción del modelo de minería: se han de establecer las técnicas de selección de características y la estrategia para construir un recomendador de algoritmos de minería que determine cuál utilizar para construir el modelo.
3. Visualización e interpretación del resultado: la representacin del conocimiento extraído debe ser fácilmente interpretable por el usuario no experto.

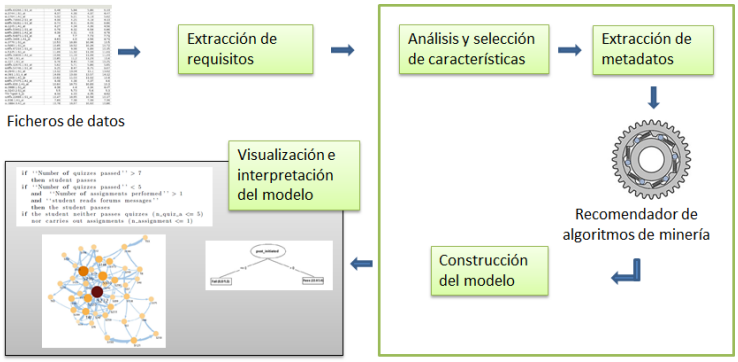

Fig. 1. Arquitectura del servicio de minería propuesto

En relación a la captura de requisitos, nuestra propuesta inicial es hacer uso de la taxonomía descrita en [3]. No obstante ésta deberá ser extendida con controles que verifiquen que el fichero de datos es conforme a los requisitos de la técnica. Como reto de futuro, se debe avanzar para que el usuario exprese su objetivo en lenguaje natural y sea el servicio el que apoyado por ontologías contextuales a cada dominio o bien utilizando lenguajes específicos de dominio [6], determine el tipo de problema a resolver y los atributos que deberían conformar el fichero de datos.

Respecto al módulo de minería, el componente más importante es el recomendador, así que su efectividad será esencial para el éxito de la propuesta. Para su diseño y construcción proponemos el uso de técnicas de meta-aprendizaje sobre las que otros autores y nosotros mismos [8, 7, 10, 3] hemos comprobado experimentalmente, que ofrece resultados bastante buenos y próximos a los que conseguiría un experto. No obstante, aún se debe profundizar más en el estudio de las meta-características que mejor caracterizan a las fuentes de datos (generales, estadísticas, basadas en modelo, medidas de complejidad, etc.), y a los modelos construidos que conforman la base de datos experimental sobre la que se construye el recomendador. Este se construirá a partir de estos experimentos utilizando bien un árbol de decisión (elegir la mejor técnica) o un conjunto de regresores (ranking de técnicas a partir de su predicción predicha). El componente responsable de las tareas de preprocesado y selección de características no entraña ninguna dificultad, ya que son técnicas bien conocidas y, por tanto, se pueden aplicar de forma sistemática siguiendo el modelo de procesos estándar CRISP-DM [2]: evaluar correlaciones entre atributos y con respecto a la clase, desbalanceo de clase, valores nulos, técnicas de selección de atributos, etc. Lo mismo se puede decir del componente responsable de construir el modelo.

El módulo responsable de la visualización del patrón condiciona el tipo de algoritmos de minería que se pueden utilizar pues la salida de muchos de ellos no permiten la extracción de conocimiento accionable, caso de todos los denominados de caja negra (redes neuronales, máquina soporte vector,...). La repre-

sentación en forma de reglas es la más sencilla y fácil de interpretar, pero no es válida para todo tipo de técnicas, por lo que es necesario complementar la salida con elementos gráficos. Aquí se plantean nuevos retos debidos principalmente a la dimensionalidad de las fuentes de datos.

3 Conclusiones y trabajos futuros

Este artículo propone una arquitectura modular que permita desarrollar un servicio de minería de datos para usuarios inexpertos en este campo. Trata así de abordar uno de los retos de investigación recogidos en el informe Beckman [1], en concreto, el etiquetado como *End-to-End Processing and Understanding of Data* que señala la falta de herramientas que permitan procesar datos en crudo y extraer conocimiento sin una intervención significativa del usuario.

Dada la envergadura de la propuesta, los trabajos que se acometerán próximamente son dos: uno, relacionado con el módulo de recomendación que es determinar experimentalmente la metacaracterísticas más significativas para la construcción del recomendador y el método más efectivo para implementar el mismo; y dos, realizar un estudio empírico sobre la utilidad de la taxonomía propuesta para recoger los requisitos de los usuarios.

References

1. Abadi, D., et al.: The beckman report on database research. SIGMOD Rec. 43(3), 61–70 (Dec 2014)
2. Chapman, P.e.a.: Crisp-dm 1.0 step-by-step data mining guide. Tech. rep., The CRISP-DM consortium (August 2000)
3. Espinosa, R., Garca-Saiz, D., Zorrilla, M., Zubcoff, J., Mazn, J.N.: Enabling non-expert users to apply data mining for bridging the big data divide. In: Ceravolo, P., Accorsi, R., Cudre-Mauroux, P. (eds.) Data-Driven Process Discovery and Analysis, LNBIP, vol. 203, pp. 65–86. Springer Berlin Heidelberg (2015)
4. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Commun. ACM 39(11), 27–34 (Nov 1996)
5. Garca-Saiz, D., Palazuelos, C., Zorrilla, M.: Data mining and social network analysis in the educational field: An application for non-expert users. In: Peña Ayala, A. (ed.) Educational Data Mining, SCI, vol. 524, pp. 411–439. Springer (2014)
6. Alfonso de La Vega, Diego Garca-Saiz, P.S., Zorrilla, M.: Domain specific languages for data mining: A case study for educational data mining. In: Symposium on Languages, Applications and Technologies (2015)
7. Reif, M., Shafait, F., Goldstein, M., Breuel, T., Dengel, A.: Automatic classifier selection for non-experts. Pattern Analysis and Applications 17(1), 83–96 (2014)
8. de Souto, M., et al.: Ranking and selecting clustering algorithms using a meta-learning approach. In: Neural Networks, 2008. pp. 3729–3735 (June 2008)
9. Zorrilla, M.E., García-Saiz, D.: A service oriented architecture to provide data mining services for non-expert data miners. DSS 55(1), 399–411 (2013)
10. Zorrilla, M.E., García-Saiz, D.: Meta-learning: Can it be suitable to automatise the KDD process for the educational domain? LNCS, vol. 8537, pp. 285–292. Springer (2014)