Convocatorias 2014
Proyectos de I+D "EXCELENCIA" y Proyectos de I+D+I "RETOS INVESTIGACIÓN"
Dirección General de Investigación Científica y Técnica
Subdirección General de Proyectos de Investigación

AVISO IMPORTANTE

En virtud del artículo 11 de la convocatoria **NO SE ACEPTARÁN NI SERÁN SUBSANABLES MEMORIAS CIENTÍFICO-TÉCNICAS** que no se presenten en este formato.

Lea detenidamente las instrucciones que figuran al final de este documento para rellenar correctamente la memoria científico-técnica.

Parte A: RESUMEN DE LA PROPUESTA/SUMMARY OF THE PROPOSAL

INVESTIGADOR PRINCIPAL 1 (Nombre y apellidos)**:**

Sebastián Ventura Soto

INVESTIGADOR PRINCIPAL 2 (Nombre y apellidos)**:**

TÍTULO DEL PROYECTO: Minería de datos con Representaciones más FLexibles

ACRÓNIMO: MARFIL

RESUMEN Máximo 3500 caracteres (incluyendo espacios en blanco):

El proyecto MARFIL (Minería de dAtos con Representaciones más FlexIbLes) tiene como objetivo el desarrollo de nuevos modelos de extracción de conocimiento para problemas que requieren una representación más flexible de la información:

• Modelos de aprendizaje multi-instancia y aprendizaje relacional, para representar el espacio de entrada de una forma más flexible.

• Modelos de aprendizaje con múltiples salidas, especialmente aprendizaje multi- etiqueta, para representar el espacio de salida de forma más flexible.

• Modelos de aprendizaje multi-fuente y multi-vista, que nos permiten combinar varios conjuntos de datos que describen el mismo problema a través de modelos extraídos para cada uno de estos conjuntos.

Para los paradigmas enumerados anteriormente, desarrollaremos modelos en los ámbitos de clasificación, agrupamiento, asociación y descubrimiento de subgrupos. También adaptaremos dichos modelos a problemas con características especiales, tales como gran número de variables, datos ruidosos, o desbalance de datos, proporcionando métodos de preprocesamiento adecuados y/o modelos que tengan en cuenta estas peculiaridades. Algunos de estos problemas pueden considerarse del ámbito de lo que se ha dado en denominar big data, por lo que nuestras propuestas se adaptarán a este tipo de entornos, y se desarrollarán implementaciones escalables que sean capaces de aportar soluciones apropiadas en estos contextos.

Además de su dimensión teórica, comentada anteriormente, este proyecto presenta una dimensión aplicada, dado que pretendemos resolver distintos problemas del mundo real aplicando los modelos desarrollados. En concreto, abordaremos problemas en los ámbitos

de la minería de datos educativa (predicción del rendimiento académico, modelado de la autoevaluación y de la evaluación por pares y desarrollo de modelos de recomendación de recursos y actividades para estudiantes) y de la biomedicina (diagnóstico precoz a partir del análisis de historias clínicas y predicción de riesgo de enfermedades relacionadas con el metabolismo de la insulina). Es notorio el interés social que ambos campos de aplicación despiertan actualmente en nuestra sociedad, así como la repercusión que cualquier pequeño avance pueda tener en las comunidades educativa y sanitaria. De hecho, además de nuestra estrecha colaboración con las instituciones universitarias involucradas en el proyecto, así como con el instituto Maimónides de investigación biomédica, empresas de ambos sectores ya han mostrado su interés en los resultados que pudieran derivarse de estos trabajos. Así pues, en una primera fase, analizaremos si estos nuevos modelos de representación suponen un avance en la resolución del problema con respecto a las propuestas tradicionales. En una segunda fase, compararemos las propuestas ya existentes y descritas en la bibliografía con nuestras propias soluciones, esperando que estas produzcan mejores resultados.

Por último, pero no menos importante, a fin de dar suficiente promoción a la investigación llevada a cabo, desarrollaremos repositorios de datos de prueba para cada uno de los paradigmas analizados, que permitan a la comunidad científica la replicación de nuestra experimentación y sirvan de benchmark para la comparación exhaustiva de resultados. Además, integraremos los modelos desarrollados en las plataformas software de mayor relevancia actual para facilitar la difusión de los mismos.

PALABRAS CLAVE: minería de datos, datos relacionales, multi-instancia, multi-etiqueta, multi-fuente y multi-vista, big data

TITLE OF THE PROJECT: Data Mining with More Flexible Representations

ACRONYM: MARFIL

SUMMARY Maximum 3500 characters (including spaces):

Project MARFIL (“Minería de datos con representaciones más flexibles”, “Mining data with more flexible representations”) has as objective to develop novel approaches for knowledge extraction in those contexts demanding some additional flexibility in data representation:

• Multi-instance and relational learning models that enable a more flexible representation of the input space.

• Learning models with multiple outputs, especially multi-label learning, that allow representing the output space with more flexibility.

• Multi-source and multi-view learning models, which make possible to combine together several data sets describing the same problem using models individually chosen for each of these data sources.

Having all the approaches aforementioned, we will develop new models in the scope of classification, clustering, association and subgroup discovery. We will also enable mechanisms to adapt these models to problems with special characteristics, such as a large number of variables, or very large data sets, as the circumstance dictates. Some of these problems fit into the so-called big data term, and therefore our proposals will be adapted to this new landscape, supplying scalable implementations that are able to provide innovative, appropriate solutions in these contexts.

In addition to its theoretical dimension, previously introduced, this project has got an applied orientation, since we expect to solve several real life problems making use of the developed models. More specifically, we will address some issues related to the context of educational data mining (predicting students’ academic performance, modelling self-assessment and peer assessment plans, and developing resource and activities recommendation models for

students), and biomedicine (early diagnosis by studying electronic health records, and predicting the risk of insulin metabolism diseases and related pathologies). It is remarkable the interest that nowadays arouses both application fields in our society, as well as the significant impact that any small step forward would have on the health and educational communities. In fact, in addition to our close cooperation with the Universities involved in this project and the Maimónides health research institute, several companies in both sectors have already shown their interest in the results derived from this proposal. Therefore, in a first stage, we will analyse whether these representation models really represent an important step forward to serve the problem resolution with respect to traditional approaches. In a second stage, the existing state of the art methods will be compared to our own proposals, where we expect to achieve significantly improved outcomes.

Last but not least, in order to promote the conducted research, we plan to build test data repositories together with each one of the resulting models in order to allow the scientific community to replicate our experimentation and thoroughly compare the results. Furthermore, we will integrate the developed models into the today's most relevant software platforms in order to facilitate their dissemination.

KEY WORDS: data mining, multiple-instance, multi-label, relational, multi-source and multi-view data, big data

Parte B: INFORMACIÓN ESPECÍFICA DEL EQUIPO

B.1. RELACIÓN DE LAS PERSONAS NO DOCTORES QUE COMPONEN EL EQUIPO DE TRABAJO (se recuerda que los doctores del equipo de trabajo y los componentes del equipo de investigación no se solicitan aquí porque deberán incluirse en la aplicación informática de solicitud). Repita la siguiente secuencia tantas veces como precise.

1. Alain Guerrero Enamorado:
   Titulación: Ingeniero en Ingeniería Informática y Máster en Informática
   Tipo de contrato: entidad extranjera (Universidad de Ciencias Informáticas de la Habana, Cuba)
   Duración del contrato: indefinido
2. Carmen Luque Guzmán:
   Titulación: Graduada en Ingeniería Informática y Máster en Informática
   Tipo de contrato: otros (Servicio Andaluz de Salud)
   Duración del contrato: indefinido
3. Oscar Gabriel Reyes Pupo:
   Titulación: Ingeniero en Informática y Máster en Informática
   Tipo de contrato: entidad extranjera (Universidad de Holguín, Cuba)
   Duración del contrato: indefinido
4. Aurora Ramírez Quesada:
   Titulación: Ingeniera en Informática y Máster en Informática
   Tipo de contrato: en formación (beca FPU)
   Duración del contrato: temporal
5. Hermes Robles Berumen:
   Titulación: Ingeniero de Minas y Suficiencia Investigadora
   Tipo de contrato: entidad extranjera (Universidad Autónoma de Zacatecas, Méjico)
   Duración del contrato: indefinido

B.2. FINANCIACIÓN PÚBLICA Y PRIVADA (PROYECTOS Y/O CONTRATOS DE I+D+I) DEL EQUIPO DE INVESTIGACIÓN (repita la secuencia tantas veces como se precise hasta un máximo de 10 proyectos y/o contratos).

1. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): E. Gibaja, C. Romero, J.R. Romero, S. Ventura y A. Zafra
   Referencia del proyecto: TIN2011-22408
   Título: *New challenges in knowledge discovery: a genetic programming approach*
   Investigador principal: Sebastián Ventura Soto
   Entidad financiadora: Ministerio de Ciencia e Innovación
   Duración: 01/01/2012 – 31/12/2014
   Financiación recibida: 73.053 €
   Relación con el proyecto que se presenta: está muy relacionado
   Estado del proyecto o contrato: concedido

2. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): E. Gibaja, M. Luque, C. Romero, J.R. Romero, S. Ventura y A. Zafra
   Referencia del proyecto: P08-TIC-3720
   Título: *Aplicación de Técnicas de Extracción de Conocimiento en los Sistemas Educativos (ATECSE)*
   Investigador principal: Sebastián Ventura Soto
   Entidad financiadora: Consejería de Innovación. Junta de Andalucía
   Duración: 01/01/2009 – 31/12/2013
   Financiación recibida: 172.743,68 €
   Relación con el proyecto que se presenta: está relacionado
   Estado del proyecto o contrato: concedido

3. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): E. Gibaja, C. Romero, J.R. Romero, S. Ventura y A. Zafra
Referencia del proyecto: TIN2008-06681-C06-03
Título: *Tendencias Actuales y Nuevos Retos en KEEL: Aprendizaje Multi-Instancia, Redes Neuronales Evolutivas, Minería de Datos Educativos y Minería de Datos Web*
Investigador principal: Sebastián Ventura Soto
Entidad financiadora: Ministerio de Educación y Ciencia
Duración: 01/01/2009 – 31/12/2011
Financiación recibida: 112.530 €
Relación con el proyecto que se presenta: está muy relacionado
Estado del proyecto o contrato: concedido

4. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): C. Romero, S. Ventura y A. Zafra
Referencia del proyecto: TIN2005-08386-C05-02
Título: *Modelos de Aprendizaje Evolutivo de Redes Neuronales de Unidades Producto. Programación genética y reglas de asociación. Librería Evolutiva JCLEC. Métodos inferenciales robustos asociados.*
Investigador principal: César Hervás Martínez
Entidad financiadora: Ministerio de Educación y Ciencia
Duración: 01/01/2006 – 31/12/2008
Financiación recibida: 66.640 €
Relación con el proyecto que se presenta: está relacionado
Estado del proyecto o contrato: concedido

5. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): C. Romero y S. Ventura
Referencia del proyecto: P05-TIC-00531
Título: *MINDAT-PLUS: Minería de Datos para los Usuarios*
Investigador principal: Francisco Herrera Triguero
Entidad financiadora: Consejería de Innovación. Junta de Andalucía
Duración: 01/01/2006 – 31/12/2008
Financiación recibida: 84.000 €
Relación con el proyecto que se presenta: está relacionado
Estado del proyecto o contrato: concedido

6. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos): M. Luque, C. Romero, A. Zafra y S. Ventura
Referencia del proyecto: P05-TIC-00602
Título: *Desarrollo de Sistemas de Acceso a la Información WEB Basados en Técnicas de Inteligencia Artificial (SAINFOWEB)*
Investigador principal: Enrique Herrera Viedma
Entidad financiadora: Consejería de Innovación. Junta de Andalucía
Duración: 01/01/2006 – 31/12/2008
Financiación recibida: 74.400 €
Relación con el proyecto que se presenta: está relacionado
Estado del proyecto o contrato: concedido

Parte C: DOCUMENTO CIENTÍFICO

C.1 Propuesta científica

1. Introducción. Antecedentes

En los últimos años han aparecido nuevos paradigmas en el ámbito del aprendizaje automático y la minería de datos que representan la información de forma más flexible, para resolver problemas que no habían podido ser resueltos con los paradigmas tradicionales, o bien quedaban resueltos de manera poco satisfactoria con estos enfoques. Por ejemplo, en *aprendizaje multi-instancia* (*multiple instance learning*, MIL), cada patrón está representado por una bolsa que contiene un número variable de instancias, teniendo cada instancia el mismo número de atributos [Die97]. Esta representación permite presentar cada ejemplo mediante varias observaciones, las cuales suelen estar asociadas a diferentes perspectivas o configuraciones de éste. Por otra parte, el *aprendizaje relacional* (*relational learning*, RL) tiene como objetivo trabajar directamente con datos almacenados en un sistema relacional, lo cual contrasta con la mayoría de las técnicas desarrolladas hasta la fecha, que tratan con una única tabla que contiene todos los datos [Dze01].

Los paradigmas de aprendizaje multi-instancia y relacional tienen como objetivo encontrar representaciones más flexibles para el espacio de datos de entrada. Los paradigmas de *aprendizaje multi-salida* (*multi-target learning*, MTL) [Zha14] tienen, por el contrario, la intención de presentar de un modo más flexible el espacio de salida. En estos paradigmas, a diferencia de en los tradicionales, el espacio de salida no está formado por una única variable, sino por varias que representan las características del espacio de salida que deben ser aprendidas. De entre estos enfoques, el más popular es el denominado *aprendizaje multi-etiqueta* (*multi-label learning*, MLL), donde los ejemplos del conjunto de aprendizaje pueden pertenecer a más de una clase simultáneamente. Para representar este hecho, se utiliza una variable de tipo binario que representa la pertenencia a cada una de las clases de interés de forma análoga a como se hace en aprendizaje multiclase, pero con la diferencia de que en este caso un ejemplo puede tener más de un valor binario activo [Tso07]. Otros paradigmas de aprendizaje multi-salida son el *aprendizaje multi-dimensional* (*multi-dimensional learning*, MDL), que presenta múltiples salidas de tipo categórico no binario [Bie11], y la *regresión multi-salida* (*multi-output regression,* MOR) que presenta múltiples salidas de tipo continuo [Liu09].

Todos los paradigmas presentados anteriormente están pensados para obtener un modelo o conjunto de patrones a partir de un único conjunto de datos, pero ¿qué pasa cuando disponemos de datos procedentes de diferentes fuentes? Clásicamente, se han agregado los distintos conjuntos en uno solo, pero esta metodología no funciona bien en todos los casos, pudiendo aumentar la ambigüedad del conjunto de datos y dificultar el proceso de aprendizaje. Para dar soluciones más flexibles a este tipo de problemas aparecen el *aprendizaje multi-fuente* (*multi-source learning*, MSL) [Cra08] y el *aprendizaje multi-vista* (*multi-view learning*, MVL) [Xu13]. El propósito de estos paradigmas es el de combinar información heterogénea y complementaria para aprender modelos que obtengan mejores resultados que modelos independientes generados a partir de informaciones independientes.

El propósito principal de este proyecto es el desarrollo de nuevas propuestas que proporcionen soluciones más eficaces y útiles a los problemas de extracción de conocimiento, basadas en los enfoques mencionados previamente. Además, dichas propuestas se aplicarán a problemas reales en dos dominios de aplicación: minería de datos educativos y minería de datos biomédicos. Las soluciones a desarrollar deberán incluir métodos apropiados para el preprocesado de datos (por ejemplo, rebalanceo, selección de características o de instancias) y ser escalables (utilizando técnicas de paralelización apropiadas), por la alta dimensionalidad de los problemas a tratar. Algunos de estos problemas pertenecerán al ámbito de lo que se ha dado en denominar *big data.* En estos casos, se desarrollarán implementaciones altamente escalables, capaces de aportar soluciones ante una cantidad masiva de información.

En lo que sigue, se detallarán brevemente los problemas y dominios de aplicación que serán el objetivo de esta propuesta.

Aprendizaje multi-instancia

El aprendizaje multi-instancia es un paradigma de aprendizaje propuesto por Dietterich en 1997 para dar solución a un problema de modelado entre la relación estructura-actividad de fármacos [Die97]. En este paradigma, cada patrón está representado por una bolsa que contiene un número variable de instancias, cada una con el mismo número de atributos. Esta representación permite presentar un patrón mediante varias observaciones, las cuales suelen estar asociadas a diferentes perspectivas o configuraciones del mismo. La gran flexibilidad de este tipo de representación ha propiciado su uso en aplicaciones tan diversas como la recuperación de imágenes [Xu14b], reconocimiento de escenas [Xu14a], clasificación de documentos [Zha11] y de páginas web índice [Zaf09].

Existen multitud de algoritmos propuestos para abordar el problema del MIL. Tenemos por una parte, algoritmos diseñados específicamente para MIL tales como *APR* [Die97] y *Diverse Density* [Pao08] y por otra algoritmos que adaptan las hipótesis del aprendizaje tradicional al entorno multi-instancia [Zho06] (puede encontrarse una completa bibliografía sobre el tema en http://www.uco.es/grupos/kdis/wiki/mil_bib). Nuestro grupo ha realizado aportaciones en el aprendizaje de reglas y su paralelización [Zaf10; Can14], en selección de características [Zaf12] y en la resolución de algunos problemas reales [Zaf11] (en la página web anterior pueden también consultarse nuestras contribuciones al tema). Sin embargo, existen temas abiertos que se pretenden abordar a lo largo de este proyecto:

- *Clustering multi-instancia*. Hasta la fecha, se han realizado algunas propuestas de clustering con multi-instancias, incluyendo clustering jerárquico y particional [Zha09]. Nuestro objetivo es proponer mejores alternativas a estas propuestas basadas en el uso de nuevas métricas de distancia entre objetos.
- *Hipótesis de trabajo generalizadas en MIL*. La mayoría de los estudios realizados en MIL están basados en la hipótesis de Dietterich [Die97]. Sin embargo, existen hipótesis generalizadas [Fou10] que han mostrado ser más apropiadas en ciertos problemas. En este sentido, se pretende explorar la adaptación de nuestras propuestas a estas nuevas hipótesis y analizar su efectividad en la resolución de distintos problemas.
- *Representación de textos como multi-instancias*. El uso de multi-instancias se ha revelado *recientemente* como una alternativa muy interesante para la representación de documentos [Kum11]. Nuestro objetivo será evaluar la efectividad de este enfoque y aplicar nuestras propuestas a la clasificación de documentos biomédicos.

Aprendizaje relacional

El *aprendizaje relacional* [Dze10] tiene como objetivo trabajar directamente con datos almacenados en un esquema relacional, lo cual contrasta con la mayoría de las técnicas desarrolladas hasta la fecha, que tratan con una única tabla que contiene todos los datos. El interés de este tipo de aprendizaje es claro, ya que en multitud de ocasiones no es posible pasar de un modelo relacional completo a un modelo en forma de una única tabla sin replicar información o sin dejar una gran cantidad de datos perdidos, lo que distorsiona el proceso de extracción de nuevo conocimiento [Dze01].

El uso de diversas relaciones de datos, unido a la dificultad que conlleva visualizar y comprender la información almacenada en dichas relaciones, ha originado el creciente interés en propuestas relativas al aprendizaje relacional. Este interés se ha visto reflejado en dominios de aplicación tales como el diseño de medicamentos, donde se trata de relacionar propiedades de diversos componentes con efectos específicos [Kin95]. En la actualidad, existen propuestas que tratan los datos relacionales mediante modelos de predicción (clasificación y regresión), de agrupamiento (*clustering,* si usamos el término inglés) y de descripción de datos mediante minería de asociaciones [Dze01].

El grupo de investigación ha desarrollado una primera propuesta para la extracción de conocimiento en datos relacionales [Lun13]. Sin embargo, consideramos que aún queda mucho trabajo en esta línea, entre el que destacamos:

- Desarrollo de nuevos modelos predictivos (clasificación) para datos relacionales.
- Desarrollo de nuevos modelos de *clustering* en bases de datos relacionales.
- Diseño de algoritmos de asociación que permitan describir el comportamiento de los datos en múltiples relaciones de datos.

Aprendizaje multi-etiqueta

El aprendizaje multi-etiqueta se caracteriza, a diferencia del aprendizaje supervisado clásico, en que un patrón puede tener asociadas varias clases o etiquetas de forma simultánea. Este modelo de aprendizaje supervisado se ha utilizado con éxito en tareas como la clasificación de textos o multimedia y problemas como el marketing directo (ej. a un cliente potencial es útil recomendarle varios productos) entre otros.

Los métodos desarrollados para MLL pueden categorizarse en *métodos de transformación* y *métodos de adaptación de algoritmos*. Los primeros transforman un *dataset* multi-etiqueta en uno o varios *datasets* de una sola etiqueta y, posteriormente, se aplica cualquier algoritmo de clasificación clásico. Algunos consideran que las etiquetas son independientes [Tso10a]. Otros consideran las relaciones entre etiquetas [Tso10a], lo que conlleva un elevado coste computacional. Las propuestas más recientes se han centrado en considerar las relaciones entre etiquetas con un coste razonable [Rea11]. Los métodos de adaptación de algoritmos adaptan algún algoritmo clásico para trabajar directamente con datos multi-etiqueta [Zha06]. Algunos autores diferencian una tercera categoría de métodos de gran repercusión debido a sus buenos resultados predictivos: *multi-clasificadores de clasificadores multi-etiqueta*, en la que los clasificadores base son multi-etiqueta [Rea11; Tso10b].

El equipo cuenta con experiencia en el área, habiendo publicado varios modelos de aprendizaje evolutivo para MLL [Avi10; Avi11; Can13] (puede encontrarse una completa bibliografía sobre el tema, incluyendo las contribuciones del grupo en el tema, en http://www.uco.es/grupos/kdis/wiki/mlc_bib).

Entre los nuevos retos del MLL cabe destacar la necesidad de explotar las relaciones entre etiquetas, la presencia de datos no balanceados y la gran dimensionalidad de los datos. El proyecto que se presenta pretende abordar algunos de estos retos:

- *Optimización de multi-clasificadores multi-etiqueta*, mediante algoritmos o estrategias para seleccionar y combinar los clasificadores base de entre un conjunto de candidatos.
- *Desarrollo de técnicas de reducción de datos para MLL* incluyendo selección de características, selección y generación de prototipos y discretización.
- *Hibridación del MLL con otros paradigmas de aprendizaje*, fundamentalmente MIL (para trabajar con representaciones más flexibles del espacio de entrada) y MVL (para complementar los modelos obtenidos desde diferentes fuentes de información).

Aprendizajes multi-vista y multi-fuente

El aprendizaje a partir de múltiples vistas y fuentes de datos ha atraído el interés de la comunidad científica en los últimos años [Cra08; Xu13]. Este campo de investigación plantea el reto de unir información heterogénea y complementaria para aprender modelos que obtengan mejores resultados que modelos independientes a partir de informaciones independientes. Esta metodología se basa en los principios de consenso y complementariedad de los datos.

Los datos del mundo real se encuentran almacenados de forma distribuida, particionada y con una organización heterogénea en múltiples fuentes de información. Cada fuente se considera una vista parcial de los datos. Tradicionalmente, los algoritmos de aprendizaje han inferido modelos a partir de un único conjunto de datos homogéneo. Por otro lado, el aprendizaje multi-vista introduce un nuevo paradigma que modela la unión de vistas para aprovechar la redundancia y complementariedad de los datos para mejorar el aprendizaje.

Este nuevo campo de estudio demanda el análisis y desarrollo de nuevas métricas y modelos para su resolución [Wan11; Xie11; Zha12].

Los algoritmos de aprendizaje multi-vista se pueden categorizar en tres grupos: *algoritmos de co-aprendizaje,* que entrenan alternativamente vistas de datos para maximizar la coincidencia en la salida [Ami10], *algoritmos basados en combinaciones de funciones kernel* [Gon11] y *algoritmos de aprendizaje en subespacios latentes* de características compartidas por las diferentes vistas [Sha12].

En este proyecto nos planteamos la resolución de una serie de problemas asociados al aprendizaje multi-vista multi-fuente:

- *Estudio de las interrelaciones entre las vistas y fuentes de datos* para identificar problemas que violen los principios del aprendizaje multi-vista, así como el filtrado de ruido y el estudio de la compatibilidad de datos.
- *Desarrollo de modelos de aprendizaje para la selección o ponderación de subconjuntos de vistas* que mejoren los resultados del subsecuente algoritmo de minería de datos.
- *Diseño e implementación de algoritmos de aprendizaje multi-instancia* que infieran directamente modelos a partir de múltiples vistas y fuentes de datos heterogéneas.

Minería de asociaciones

La minería de asociaciones (*association rule mining*, ARM) tiene como finalidad la extracción y descripción de las relaciones entre los ítems existentes en una base de datos [Agr93]. Estas relaciones se definen mediante implicaciones de la forma X-->Y, siendo X e Y conjuntos disjuntos de ítems. Así, una regla de asociación determina que Y generalmente ocurre siempre que ocurre X.

A pesar de que ARM se centró, originalmente, en dominios puramente discretos, muchos investigadores han propuesto modelos que permiten trabajar en dominios continuos [Lun12; Mar14], así como en contextos muy variados [Rom13c; San08]. En muchos dominios de aplicación, como es el caso de análisis de datos bancarios, la búsqueda de excepciones [Lem08] juega un papel fundamental. En ésta búsqueda de excepciones o irregularidades, la correlación de los patrones es de gran relevancia, por lo que no basta con buscar patrones que se encuentren altamente relacionados, sino que la correlación existente entre ellos debe ser analizada.

En otro sentido, el modo en el que se agrupan y guardan los datos está jugando un papel cada vez más importante en ARM, siendo cada vez más necesaria la búsqueda de relaciones entre datos agrupados en múltiples fuentes de datos con diferentes características [Lun13]. Además, dichos datos pueden estructurarse en bloques indivisibles, por ejemplo transacciones llevadas a cabo por usuarios donde el número de transacciones varía de usuario a usuario. En esta estructura de datos, la extracción de relaciones debe llevarse a cabo entre usuarios, independientemente del número de operaciones llevadas a cabo por cada uno. En http://www.uco.es/grupos/kdis/wiki/arm_bib puede encontrarse una amplia bibliografía sobre el tema.

El grupo de investigación ha desarrollado varias propuestas en el ámbito de ARM [Lun12; Lun13; Rom13c]. Se plantea pues, el estudio de esta tarea desde el punto de vista de nuevas representaciones de datos más flexibles, tratando los siguientes apartados:

- Diseño e implementación de modelos que permitan la extracción de reglas de asociación desde múltiples vistas y fuentes heterogéneas de datos de una manera eficiente y haciendo frente al problema de alta dimensionalidad existente.
- Desarrollo de un modelo que permita la extracción de reglas de asociación excepcionales, permitiendo descubrir comportamientos anómalos entre patrones.
- Diseño e implementación de modelos de asociación a partir de datos multi-instancia. En estos modelos, las relaciones entre patrones deben llevarse a cabo sobre el conjunto total de usuarios o bolsas de instancias. De este modo, todas las bolsas tienen el mismo peso dentro de la base de datos, independientemente del número de instancias que posea cada una.

- Estudio y desarrollo de modelos que permitan describir el comportamiento de los datos incluyendo análisis de correlación en el proceso de ARM, otorgando una mayor descripción del conocimiento extraído.

Big Data

El uso masivo de las TIC genera grandes cantidades de datos. El análisis y el procesamiento de dicha información es un verdadero reto y demanda soluciones computacionales para su procesamiento. Este análisis requiere una metodología revolucionaria donde los métodos tradicionales no son una opción [Sag13]. Big Data se caracteriza por tres propiedades principales: variedad, velocidad y volumen. Estas tres propiedades deben considerarse como un todo para lograr el procesamiento de grandes datos del orden de terabytes, o incluso exabytes, a partir de una gran variedad de fuentes y de una manera rápida y eficiente.

En los últimos años ha crecido el número de investigaciones sobre nuevas metodologías que soportan este gran desafío [Max13]. Los enfoques se centran en la combinación adecuada del análisis, experiencia y arquitectura de hardware [Chu06] en torno a un problema específico.

El software más conocido es Apache Hadoop [Kon13], un marco que permite la computación distribuida de grandes conjuntos de datos a través de clusters de computadores. La clave de Hadoop es su eficiencia y su modelo de programación simplificado, que permite al usuario escribir y evaluar sistemas distribuidos rápidamente. Hadoop se basa en MapReduce [Jia11; Lee11], un marco de programación creado por Google [Lam08], que se ha establecido como la piedra angular en el manejo de grandes datos. MapReduce utiliza una estrategia de divide y vencerás para dividir los datos complejos en unidades más fáciles de manejar.

Big Data es un campo reciente de estudio donde no sólo se considera el problema del procesamiento de datos, sino también el problema de descubrimiento, análisis y mejora en la toma de decisiones. El uso de conjuntos de datos masivos y la metodología de MapReduce en el aprendizaje automático [Con13] es un tema interesante que constituye un nuevo paradigma en el proceso KDD. En este sentido, Mahout [Owe11] y Spark [Zah10] son bibliotecas para el Aprendizaje Automático y Minería de Datos.

Considerado lo anteriormente comentado, en este proyecto se plantean como objetivos:

- *Desarrollar modelos de aprendizaje Big Data en Minería de Datos*, centrándonos en las representaciones flexibles de multi-instancia, multi-etiqueta y multi-vista.
- *Integrar los algoritmos desarrollados en Mahout o Spark*, facilitando la accesibilidad de los modelos desarrollados a la comunidad científica.

Minería de datos educativos

El interés por la aplicación de técnicas de minería de datos para resolver problemas en el ámbito educativo ha crecido enormemente en los últimos años [Rom13a] (la dirección http://www.uco.es/grupos/kdis/wiki/edm_bib presenta una amplia bibliografía sobre el tema). De entre los muchos problemas abordados en este campo de investigación hasta la fecha, en este proyecto nos centraremos en los enumerados a continuación, por el uso que pueden hacer de los modelos obtenidos en investigaciones previas:

- *Predicción del rendimiento académico*. La tarea de estimar el valor desconocido del rendimiento del estudiante es un problema de gran interés que ha sido abordado previamente [Rom13b; Mar13]. En un trabajo previo mostramos que el uso de la representación multi-instancia puede mejorar los resultados producidos con la tradicional [Zaf11]. Nuestro propósito es ampliar esta investigación y hacer uso de modelos de aprendizaje multi-vista para combinar las predicciones realizadas por distintos modelos que hacen uso de diferentes fuentes de información.
- *Modelando la autoevaluación y la evaluación por pares*. La necesidad de disponer de sistemas de evaluación automática ha crecido en los últimos años, sobre todo desde la aparición de los MOOCs (Massive Online Open Courses), donde el número de alumnos por curso hace inmanejable una evaluación directa por parte del profesor [Dar13]. Los sistemas de autoevaluación y de evaluación por pares son una alternativa con un gran

potencial, aunque presentan el problema de producir resultados que discrepan bastante de las evaluaciones realizadas por el profesor. En este sentido, se hace necesario el desarrollo de modelos que sean capaces de (i) correlar las calificaciones del alumno con las calificaciones del profesor y/o (ii) asignar un grado de fiabilidad a la calificación realizada por un alumno. En este proyecto pretendemos utilizar las técnicas de extracción de conocimiento explicadas anteriormente para establecer este tipo de modelos.

- *Recomendación de recursos didácticos a los estudiantes*. Los sistemas recomendadores en educación (*Educational Recommender Systems*, ERSs) realizan una estimación de los mejores recursos/actividades que un alumno debe utilizar para mejorar su aprendizaje en función de sus intereses y/o perfil [San12]. Como en el caso de la predicción del rendimiento, la mayoría de los sistemas recomendadores hacen uso de algoritmos clásicos. Nuestro propósito en este caso es evaluar si los modelos obtenidos con estas nuevas representaciones mejoran las recomendaciones obtenidas.

Minería de datos biomédicos

En las últimas décadas se ha producido un importante crecimiento relacionado con los recursos informacionales disponibles en el dominio de la medicina. La existencia de información diseminada por bases de datos, revistas, portales, buscadores especializados, y un sin fin de recursos repartidos por la red, ha facilitado enormemente la labor diaria de los profesionales de la medicina. Sin embargo, también existe otra información muy importante, tal como la derivada de las historias clínicas de los pacientes, de la que resulta mucho más difícil extraer información útil automáticamente [Jen12]. En estos casos la minería de datos se ha revelado como una herramienta fundamental, que permite realizar desde tareas tan simples como puede ser el análisis de conjuntos de datos clínicos, hasta tareas de mayor complejidad como es el apoyo en la toma de decisiones del diagnóstico y pronóstico médico [Cio02].

Nuestro equipo ha iniciado la investigación en este área con el análisis de una base de datos clínica para establecer qué factores influyen en la readmisión hospitalaria de pacientes de diabetes [Str14]. En este proyecto, se trabajará sobre los siguientes problemas:

- *Modelos de predicción de enfermedades del metabolismo de los hidratos de carbono.* El objetivo de este trabajo es desarrollar modelos para identificar el riesgo personalizado de presentar deterioro de la sensibilidad a la insulina, fallo de célula beta y desarrollo de diabetes de tipo 2 en pacientes, con predicción a 5 y 10 años. Para ello, se dispondrá de una base de datos de 700 pacientes proporcionada por la Unidad de Lípidos y Arteriosclerosis del Hospital Reina Sofía de Córdoba. La base de datos tiene un esquema relacional complejo, por lo que se podrán construir modelos de aprendizaje relacional y comparar los resultados obtenidos mediante estas técnicas y las técnicas tradicionales.
- *Clasificación de documentos para el diagnóstico precoz de enfermedades.* El objetivo de este trabajo es desarrollar modelos basados en los paradigmas enumerados anteriormente (*multi-instance* y *multi-label learning*) para extraer información útil de las historias clínicas de pacientes que conduzca hacia la detección temprana de enfermedades. En este caso construiremos nuestros propios datasets a partir de la información públicamente disponible en la red EURORAD (http://www.eurorad.org/) y el taller SemEval-2014 (http://alt.qcri.org/semeval2014/task7/).

Grupos de investigación interesados en la temática

Para terminar esta sección de antecedentes, se presentará brevemente alguna información sobre los grupos de investigación que trabajan en esta temática. En cuanto a los grupos internacionales, indicar que en su gran mayoría publican en las conferencias sobre aprendizaje automático y minería de datos tales como la *ACM International Conference on Knowledge Discovery from Databases* (ACM SIG-KDD), la *IEEE International Conference on Data Mining* (ICDM), la *Pacific-Asian Conference on Knowledge Discovery from Databases* (PAKDD), la *IEEE International Conference on Machine Learning* (ICML) o la *European Conference on Machine Learning* (ECML), que se celebra conjuntamente con *Practice of Knowledge Discovery in Databases* (PKDD), así como en revistas tales como *IEEE*

Transactions on Knowledge and Data Engineering, Data Mining and Knowledge Discovery, Knowledge and Information Systems, Data and Knowledge Engineering, IEEE Transactions on Neural Networks and Learning y *Machine Learning*.

Con respecto a los grupos nacionales que trabajan en esta temática, indicar que, además de publicar en las conferencias y revistas anteriores, se reúnen periódicamente en el Taller de Minería de Datos y Aprendizaje (TAMIDA) que se celebra anualmente, así como en la Conferencia de la Asociación Española para la Inteligencia Artificial (AEPIA). En la dirección http://www.lsi.us.es/redmidas/ puede encontrarse un listado de grupos de investigación que trabajan en minería de datos. Algunos de estos grupos han publicado trabajos en las líneas directamente relacionadas con el desarrollo de este proyecto:

- El grupo SCI2S de la Universidad de Granada, coordinado por Francisco Herrera, ha trabajado en aprendizaje multi-instancia, en minería de reglas de asociación y, recientemente, ha publicado trabajos en el área de *big data*.
- El grupo SIMIDAT de la Universidad de Jaén, coordinado por María José del Jesus, ha trabajado en minería de reglas de asociación, descubrimiento de subgrupos y, más recientemente, en aprendizaje multi-etiqueta.
- El grupo CIG del a Universidad Politécnica de Madrid, coordinado por Concepción Bielza y Pedro Larrañaga ha trabajado en aprendizaje multi-etiqueta entre otros temas.
- El grupo ML de la Universidad de Oviedo en Gijón, coordinado por Antonio Bahamonde, también ha trabajado en aprendizaje multi-etiqueta.
- El grupo MINERVA de la Universidad de Sevilla, coordinado por José C. Riquelme, y el grupo IDBIS de la Universidad de Granada, coordinado por Juan Carlos Cubero, han trabajado en minería de reglas de asociación.
- El grupo MIDAS de la Universidad Politécnica de Madrid, coordinado por Ernestina Menasalvas, ha trabajado en minería de reglas de asociación y, recientemente, trabaja en la línea de *big data analytics*.

Nuestro equipo de trabajo ya ha colaborado con algunos de estos grupos en varias ocasiones, y esperamos poder establecer nuevas colaboraciones con ellos y otros grupos similares (sobre todo internacionales) a lo largo del desarrollo de este proyecto, de cara a poder formar consorcios competitivos para solicitar proyectos en convocatorias de investigación internacionales (tal como la del H2020).

Referencias

[Agr93] R. Agrawal, T. Imielinski, A.N. Swami. Mining association rules between sets of items in large databases. *1993 ACM SIGMOD Int. Conference on Management of Data*, 207-216 (1993).

[Ami10] M.R. Amini, C. Goutte. A co-classification approach to learning from multilingual corpora. *Machine Learning*, 79(1), 105–121 (2010).

[Avi10] J.L. Ávila, E. Gibaja, S. Ventura. Evolving multi-label classification rules with gene expression programming: A preliminary study. *HAIS 2010*, 2, 9-16 (2010).

[Avi11] J.L. Ávila, E. Gibaja, A. Zafra, S. Ventura. A gene expression programming algorithm for multi-label classification. *Multiple-Valued Logic and Soft Computing,* 17(2-3), 183-206 (2011).

[Bie11] C. Bielza, G. Li, P. Larrañaga. Multi-dimensional classification with Bayesian networks. *Int. J. Approx. Reasoning* 52(6), 705-727 (2011).

[Can13] A. Cano, A. Zafra, E. Gibaja, S. Ventura. A grammar-guided genetic programming algorithm for multi-label classification. *EuroGP 2013*, 217-228 (2013).

[Can14] A. Cano, A. Zafra, S. Ventura. Speeding up multiple instance learning classification rules on GPUs. *Knowledge and Information Systems* (2014, *article in press*).

[Cio02] K.J. Cios, G.W. Moore. Uniqueness of medical data mining. *Artificial Intelligence in Medicine,* 26, 1–24 (2002).

[Con13] T. Condie, P. Mineiro, N. Polyzotis, M. Weimer. Machine learning for big data. *Proceedings of the ACM SIGMOD Int. Conf. on Manag. of Data*, 939-942 (2013).

[Chu06] C. Chu, S. Kim, Y. Lin, Y. Yu, G. Bradski, A. Ng, K. Olukotun. Map-reduce for machine learning on multicore. *In NIPS*, 281-288 (2006).

[Cra08] K. Crammer, M. Kearns, J. Wortman. Learning from multiple sources, *J. of Machine Learning Research*, 9, 1757-1774 (2008).

[Dar13] T. Daradoumis, R. Bassi, F. Xhafa, S. Caballe. A review on massive e-learning (MOOC) design, delivery and assessment. *Int. Conf. P2P, parallel, grid, cloud and internet computing*. 208-213 (2013).

[Die97] T.G. Dietterich, R.H. Lathrop, T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence,* 89(1-2), 31-71 (1997).

[Dze01] S. Dzeroski, N. Lavrac. *Relational Data Minin*g. Springer (2001).

[Dze10] S. Dzeroski. Relational Data Mining. *Data Mining and Knowledge Discovery Handbook*. 887-991 (2010).

[Fou10] J. Foulds, E. Frank. A review of multi-instance learning assumptions. *Know. Eng. Review*, 25(1), 1-25 (2010).

[Gon11] M. Gonen, E. Alpaydin. Multiple kernel learning algorithms. *J. of Machine Learning Research*, 12, 2211-2268 (2011).

[Jen12] P.B. Jensen, L.J. Jensen, S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics,* 13, 395-405 (2012).

[Jia11] D. Jiang, A. Tung, G. Chen. MAP-JOIN-REDUCE: Toward scalable and efficient data analysis on large clusters. *IEEE Trans. on Know.and Data Eng.*, 23 (9), 1299-1311 (2011).

[Kin95] R.D. King, A. Srinivasan, M.J.E. Sternberg. Relating chemical activity to structure: An examination of ILP successes. *New Generation Computin*g, 13, 411-433 (1995).

[Kon13] X. Kong. Hadoop MapReduce in cloud environments for scientific data processing. *Journal of Theoretical and Applied Information Technology*, 48(3), 1822–1826 (2013).

[Kum11] J. Kumar, J. Pillai, D. Doermann. Document image classification and labeling using multiple instance learning. In *Proc. of the International Conference on Document Analysis and Recognition*, ICDAR 2011, 1059-1063 (2011).

[Lam08] R. Lam. Google's MapReduce programming model - Revisited. *Science of Computer Programming*, 70(1), 1-30 (2008).

[Lee11] K.H. Lee, Y.J. Lee, H. Choi, Y. Chung, B. Moon. Parallel data processing with MapReduce: A survey. *SIGMOD Record*, 40(4), 11–20 (2011).

[Lem08] D. Leman, A. Feelders, A.J. Knobbe. Exceptional model mining. *In Proc. of the European Conf. in Machine Learning and Knowledge Discovery in Databases, ECML/PKDD* 2008, 1-16 (2008).

[Liu09] G. Liu, Z. Lin, Y. Yu. Multi-output regression on the output manifold. *Pattern Recognition*, 42(11), 2737-2743 (2009).

[Lun12] J.M. Luna, J.R. Romero, S. Ventura. Design and behaviour study of a grammar guided genetic programming algorithm for mining association rules. *Knowl. and Inform. Syst.* 32(1), 53-76 (2012).

[Lun13] J.M. Luna, C. Romero, J.R. Romero, S. Ventura. Extracción de reglas de asociación frecuentes en bases de datos relacionales, *IX Congreso Español sobre Metaheurísticas, Alg. Evol. y Bioinsp.*, *MAEB 2013*, 753-762 (2013).

[Mar13] C. Marquez-Vera, A. Cano, C. Romero, S. Ventura. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, 38(3), 315-330 (2013).

[Max13] V. Marx. The big challenges of big data. *Nature*, 498(7453), 255–260 (2013).

[Mar14] M. Martínez-Ballesteros, I.A. Nepomuceno-Chamorro, J.C. Riquelme. Discovering gene association networks by multi-objective evolutionary quantitative association rules**.** *Journal of Computer and System Sciences*, 80(1), 118-136 (2014).

[Owe11] S. Owen, R. Anil, T. Dunning, E. Friedman. *Mahout in Action*. Manning (2011).

[Pao08] H.T. Pao, S.C. Chuang, Y.Y. Xu, H. Fu. An EM based multiple instance learning method for image classification. *Expert Syst. Appl.*, 35(3), 1468–1472 (2008).

[Rea11] J. Read, B. Pfahringer, G. Holmes, E. Frank. Classifier chains for multi-label classification. *Machine Learning* 85(3), 1-27 (2011).

[Rom13a] C. Romero, S. Ventura. Data mining in education. *WIREs Data Mining and Knowledge Discovery*, 3, 12-27 (2013).

[Rom13b] C. Romero, P.G. Espejo, A. Zafra, J.R. Romero, S. Ventura. Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education,* 21(1), 135–146 (2013).

[Rom13c] C. Romero, A. Zafra, J.M. Luna, S. Ventura. Association rule mining using genetic programming to provide feedback to instructors from multiple-choice quiz data. *Expert Systems*, 30(2), 162-172 (2013).

[Sag13] S. Sagiroglu, D. Sinanc. Big data: a review. *Proceedings of the Int. Conf. on Collaboration Technologies and Systems*, 42-47 (2013).

[San08] D. Sánchez, J.M. Serrano, L. Cerda, M. A. Vila. Association rules applied to credit card fraud detection. *Expert Systems with Applications,* 36, 3630-3640 (2008).

[San12] O.C. Santos, J.G. Boticario. Educational recommender systems and technologies. *Practices and Challenges*. IGI Global (2012).

[Sha12] A. Sharma, A. Kumar, H. Daume III, D.W. Jacobs. Generalized multiview analysis: A discriminative latent space. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2160-2167 (2012).

[Str14] B. Strack, J.P. DeShazo, C. Gennings, J.L. Olmo, S. Ventura, K.J. Cios and J.N. Clore. Impact of HbA1c measurement on hospital readmission rates: An analysis of 70,000 clinical database patient records. *BioMed Research International*, ID#781670 (2014).

[Tso07] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3), 1-13 (2007).

[Tso10a] G. Tsoumakas, I. Katakis, I. Vlahavas. Mining Multi-label Data. *Data Mining and Knowledge Discovery Handbook*, Part 6, 667-685 (2010).

[Tso10b] G. Tsoumakas, I. Katakis, I. Vlahavas. Random k-labelsets for multi-label classification. *IEEE Trans. on Knowledge and Data Eng.* 23(7), 1079-1089 (2010).

[Wan11] Z. Wang, S. Chen, D. Gao. A novel multi-view learning developed from single-view patterns. *Pattern Recognition*, 44(10), 2395-2413 (2011).

[Xie11] B. Xie, Y. Mu, D. Tao, K. Huang. M-SNE: Multiview stochastic neighbor embedding. *IEEE Trans. on Systems, Man, and Cybernetics. Part B*, 41(4), 1088-1096 (2011).

[Xu13] C. Xu, D. Tao, C. Xu. A survey on multi-view learning, CoRR abs/1304.5634 (2013).

[Xu14a] J. Xu, S. Denman, V. Reddy, C. Fookes, S. Sridharan. Real-time video event detection in crowded scenes using MPEG derived features: A multiple instance learning approach. *Pattern Recognition Letters*, 44, 113-125 (2014).

[Xu14b] Y. Xu, J.-Y. Zhu, E.I.C. Chang, M. Lai, Z. Tu. Weakly supervised histopathology cancer image segmentation and classification. *Med. Image Analysis*, 18 (3), 591-604 (2014).

[Zaf09] A. Zafra, C. Romero, S. Ventura, E. Herrera-Viedma. Multi-instance genetic programming for web index recommendation. *Expert Syst. Appl*., 36, 11470-11479 (2009).

[Zaf10] A. Zafra, S. Ventura. G3P-MI: A genetic programming algorithm for multiple instance learning. *Information Science* 180(23), 4496-4513 (2010).

[Zaf11] A. Zafra, C. Romero, S. Ventura. Multiple instance learning for classifying students in learning management systems. *Expert Syst. Appl.* 38(12), 15020-15031 (2011).

[Zaf12] A. Zafra, M. Pechenizkiy, S. Ventura. ReliefF-MI: An extension of ReliefF to multiple instance learning. *Neurocomputing*, 75(1), 210-218 (2012).

[Zah10] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker, I. Stoica. Spark: cluster computing with working sets. *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, 10-10 (2010).

[Zha06] M.L. Zhang, Z.H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans. on Knowledge and Data Engineering,* 18(10), 1338-1351 (2006).

[Zha09] M.L. Zhang, Z.H. Zhou. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence* 31(1) , 47-68 (2009).

[Zha11] D. Zhang, F. Wang, L. Si, T. Li. Maximum margin multiple instance clustering with applications to image and text clustering. *IEEE Trans. N. Networks*, 22 (5), 739-751 (2011).

[Zha12] D. Zhai, H. Chang, S. Shan, X. Chen, W. Gao. Multiview metric learning with global consistency and local smoothness. *ACM Trans. on Intell. Sys. and Technology (TIST)*, 3(3), 53 (2012).

[Zha14] M.L. Zhang, Z.H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8), 1819-1837 (2014).

[Zho06] Z.H. Zhou. Multi-instance learning from supervised view. *J. of Comp. Sci. and Technology*, 21(5), 800-809 (2006).

2. Hipótesis de partida y objetivos generales

Como ya se ha comentado, el propósito principal de este proyecto es el *desarrollo de nuevas propuestas para extracción de conocimiento basadas en representaciones más flexibles para los espacios de entrada y salida, así como su adaptación a datos con características especiales y su aplicación a la resolución de problemas reales*. Nuestra hipótesis de trabajo es que estos métodos pueden resolver multitud de problemas de forma más efectiva que los denominados tradicionales. De este modo, los modelos de aprendizaje multi-instancia nos permitirán presentar cada ejemplo mediante varias observaciones asociadas a diferentes perspectivas o configuraciones de éste y los modelos relacionales nos permitirán trabajar directamente con los datos contenidos en un esquema relacional. Por otra parte, los modelos multi-etiqueta nos permitirán tratar el problema de la multiplicidad en clasificación, es decir, poder clasificar un mismo objeto en varias categorías de forma simultánea. Por último, los modelos multi-vista permitirán trabajar con distintas fuentes de datos, combinando información heterogénea y complementaria para aprender modelos que obtengan mejores resultados que modelos generados a partir de informaciones independientes.

El grupo de investigadores implicado ya ha explorado algunos de estos temas, y ha desarrollado propuestas que han sido publicadas en diferentes revistas y congresos internacionales (la dirección http://www.uco.es/grupos/kdis contiene información sobre las publicaciones del grupo). No obstante, pensamos que aún queda mucho trabajo que hacer en este campo, y que podemos aprovechar esta experiencia para desarrollar nuevas propuestas que mejoren las que constituyen el estado del arte en cada una de las áreas en cuestión.

Por último, indicar que esta propuesta se ajusta perfectamente al Plan Estatal de Investigación Científica, Técnica y de Innovación 2013-2016 por las siguientes razones. En primer lugar, se trata de una propuesta de investigación de calidad, por la novedad de los temas planteados, y con proyección internacional, tanto por los componentes del equipo pertenecientes a grupos de investigación internacionales como por la difusión de los resultados que pretendemos llevar a cabo en congresos y revistas internacionales. Por otra parte, se trata de una propuesta claramente formativa, ya que pretendemos que al menos las personas pertenecientes al equipo de trabajo realicen sus tesis doctorales en cuestiones relacionadas con el contenido del proyecto. También es importante resaltar el interés que ha despertado en varias empresas el contenido de este proyecto, lo que puede redundar en transferencia de los resultados a otros niveles productivos e incrementar la colaboración en materia de I+D+i entre el sector público y el sector empresarial. Por último, pero no menos importante, este proyecto también pretende incrementar la cultura científica, tecnológica e innovadora de la sociedad española mediante la difusión de los resultados de la investigación a todos los estratos de la sociedad, como se ha indicado en el plan de difusión.

Con respecto al programa europeo H2020, las temáticas asociadas a este proyecto, fundamentalmente los problemas que se tratan en la parte de aplicaciones están alineados con algunos de los objetivos del programa. Por ejemplo, la aplicación de técnicas de minería de datos sobre datos educativos para mejorar el proceso enseñanza-aprendizaje se ajusta perfectamente al objetivo ICT 20-2015 (*technologies for better human learning and teaching*). También, las aplicaciones que hemos planteado en el ámbito de la minería de datos médicos caen perfectamente en el domino de los objetivos PHC 25-2015 (*advanced ICT systems and services for integrated care*) y PHC 30-2015 (*digital representations of health data to improve diagnosis and treatment*). Por último, la adaptación de nuestros modelos a big data se ajustan a los objetivos ICT 15-2014 (*big data innovation and take up*) e ICT 16-2015 (*big data research*).

3. Objetivos específicos

El objetivo principal expuesto anteriormente, puede desglosarse en los siguientes objetivos específicos:

1. *Desarrollar nuevas propuestas para extracción de conocimiento basadas en el uso de representaciones de datos flexibles:*
   a. Modelos de aprendizaje multi-instancia.
   b. Modelos de aprendizaje multi-etiqueta.
   c. Modelos de aprendizaje relacional.
   d. Modelos de aprendizaje multi-fuente / multi-vista.
2. *Adaptar las propuestas anteriores a conjuntos de datos con características especiales:*
   a. Modelado de documentos como multi-instancias.
   b. Modelos para datos no balanceados y/o con elevada dimensionalidad.
   c. Modelos para conjuntos de datos de gran tamaño.
1. *Aplicar los modelos desarrollados en los objetivos 1 y 2 a problemas reales en el ámbito de la educación y la biomedicina:*
   a. Problemas educativos:
      - Predicción del rendimiento académico.
      - Modelado de la autoevaluación y la evaluación por pares.
      - Recomendación de recursos didácticos y actividades para la personalización de la enseñanza.
   b. Problemas biomédicos:
      - Diagnóstico precoz a partir del análisis de historias clínicas.
      - Predicción de riesgo de enfermedades relacionadas con el metabolismo de la insulina.
2. *Desarrollar repositorios de datos, para facilitar la comparación de nuestras propuestas con otras desarrolladas posteriormente, y herramientas para la generación de datos sintéticos y la caracterización de conjuntos de datos:*
   a. Repositorios de datos multi-instancia.
   b. Repositorios de datos multi-etiqueta.
   c. Repositorios de datos relacionales.
   d. Repositorios de datos multi-vista.
   e. Herramientas de simulación de datos.
   f. Herramientas para la caracterización de conjuntos de datos.
3. *Integrar las propuestas desarrolladas en las plataformas de minería de datos más populares:*
   a. Integración en WEKA y/o en KEEL para los modelos de tamaño convencional.
   b. Integración en MAHOUT y/o SPARK para los modelos de Big Data.

4. Metodología y Plan de Trabajo

La metodología propuesta tiene una parte teórica y otra práctica. Con respecto a la parte teórica, se desarrollarán nuevos algoritmos de extracción de conocimiento y se analizará su comportamiento. En cuanto a la parte práctica, consistirá en la integración de los algoritmos en diferentes entornos software y en el desarrollo de repositorios de datos de prueba mencionados con anterioridad. El método de estudio para la parte teórica consistirá en la aplicación del método científico para la resolución de problemas, es decir, establecimiento de hipótesis, recolección de datos, comprobación de las hipótesis mediante experimentación y readaptación de las hipótesis iniciales en virtud de los resultados obtenidos.

Más específicamente, el desarrollo de este proyecto se guiará por la propuesta metodológica de Acción-Investigación (*Action-Research*), especialmente adecuada para

situaciones y estudios focalizados en problemas del mundo real. Según sus principios, esta propuesta está orientada a resolver los problemas planteados desde el momento de su identificación siguiendo una filosofía colaborativa de "aprender haciendo" ("*learning by doing*"), permitiendo comprender adecuadamente sus causas y combinando teoría y práctica. Inicialmente diseñada para problemas sociales y médicos, se ha venido utilizando recurrentemente en la construcción de sistemas de información experimentales. No obstante, ha demostrado ser particularmente eficaz en propuestas y desarrollos que pretenden ser utilizados y probados por expertos o usuarios finales, como es el caso del proyecto propuesto, donde además se utilizarán fuentes de datos obtenidas a partir de interacciones en contextos reales. De hecho, esta estrategia de Acción-Investigación es ampliamente utilizada por los grupos de investigación españoles, fundamentalmente por la idoneidad y simplicidad de su propuesta de mejora iterativa.

Siguiendo la esencia de "aprender haciendo" propuesta por esta metodología, se pretenden explorar diferentes técnicas que permitan al equipo del proyecto observar, comentar y reflexionar sobre los hallazgos. Así, sucesivos ciclos de investigación y experimentación guiarán y evolucionarán el trabajo siguiendo los pasos descritos en el proceso metodológico. Para cada ciclo, los pasos a seguir serán los descritos a continuación:

1. Diagnosis: identificación del problema.
2. Plan de acción: análisis de las diferentes alternativas y elección de la más apropiada.
3. Toma de acción: diseño, implementación y prototipado de la solución.
4. Evaluación: estudio de las consecuencias de la acción realizada a partir de la exhaustiva generación de pruebas sobre el prototipo.
5. Aprendizaje: análisis y evaluación de los resultados obtenidos a fin de identificar los cambios que permitirían la mejora de la solución.

Los pasos descritos en la metodología anterior se han adaptado al proyecto, estableciendo y definiendo una serie de tareas. Dichas tareas se han dividido en tareas de tipo A (desarrollo de algoritmos para extracción de conocimiento), tareas de tipo B (adaptación de los modelos anteriores a datos con características especiales), tareas de tipo C (resolución de problemas reales) y tareas de tipo D (desarrollo de repositorios de datos e integración de las propuestas desarrolladas en las plataformas más populares). Cada tarea tiene asignado un coordinador (en negrita) y varios participantes. A continuación, se detallarán las distintas tareas planteadas, indicando su equipo de trabajo y temporalización.

A. Desarrollo de algoritmos para extracción de conocimiento

Este bloque de tareas consistirá en el desarrollo de nuevos modelos de aprendizaje con datos que requieren una representación flexible. Estos modelos abordarán los problemas planteados en las secciones anteriores de esta memoria.

T.A.1. Nuevos modelos para aprendizaje supervisado con datos multi-instancia

Descripción: Esta tarea consistirá en el desarrollo de nuevos modelos para aprendizaje multi-instancia y su evaluación frente a las propuestas existentes en la literatura.
Participantes: **Amelia Zafra**, María Luque, Sebastián Ventura, Krys Cios, Hermes Robles
Temporalización: M01-M09
Objetivo relacionado: Objetivo 1a
Resultados esperados: Nuevos modelos para clasificación de datos multi-instancia.

T.A.2. Nuevos modelos para *clustering* con datos multi-instancia

Descripción: Desarrollo de algoritmos de *clustering* multi-instancia, definiendo nuevas medidas de distancia entre objetos que mejoren las propuestas actuales.
Participantes: **Amelia Zafra**, María Luque, Hermes Robles, Sebastián Ventura
Temporalización: M10-M21
Objetivo relacionado: Objetivo 1a
Resultados esperados: Nuevos modelos para *clustering* de multi-instancias y obtención de reglas de asociación a partir de datos multi-instancia.

T.A.3. Nuevos modelos para aprendizaje multietiqueta

Descripción: En esta tarea se desarrollarán nuevas propuestas para el aprendizaje multietiqueta que sean competitivas con las descritas hasta la fecha en la literatura.
Participantes: **Eva L. Gibaja**, Cristóbal Romero, Carmen Luque, Oscar G. Reyes, Alain Guerrero
Temporalización: M01-M12
Objetivo relacionado: Objetivo 1b
Resultados esperados: Nuevos modelos para aprendizaje multi-etiqueta.

T.A.4. Nuevos modelos para aprendizaje relacional

Descripción: Diseño y evaluación de modelos para el aprendizaje relacional.
Participantes: **María Luque**, José M. Luna, Mykola Pecheniskyi, José R. Romero
Temporalización: M13-M27
Objetivo relacionado: Objetivo 1c
Resultados esperados: Nuevos modelos para clasificación a partir de datos relacionales.

T.A.5. Nuevos modelos de clasificación multi-vista / multi-fuente

Descripción: Generación de modelos de clasificación multi-vista / multi-fuente competitivos con los de la literatura.
Participantes: **Sebastián Ventura**, Alberto Cano, Krys Cios, Cristóbal Romero
Temporalización: M01-018
Objetivo relacionado: Objetivo 1d
Resultados esperados: Nuevos modelos de clasificación multi-vista/multi-fuente

T.A.6. Nuevos modelos para minería de asociaciones y descubrimiento de subgrupos

Descripción: Desarrollo de algoritmos de obtención de reglas para asociación y/o descubrimiento de subgrupos a partir de datos multi-instancia, datos relacionales y multi-vista.
Participantes: **José R. Romero**, José M. Luna, Mykola Pecheniskyi, María Luque
Temporalización: M01-M09
Objetivos relacionados: Objetivos 1a, 1c y 1d
Resultados esperados: Modelos para la obtención de reglas de asociación y descubrimiento de subgrupos en las condiciones planteadas

B. Adaptación de los modelos anteriores a datos con características especiales

En este bloque de tareas, se adaptarán los modelos desarrollados en las tareas A.1-A.6 a conjuntos de datos con características especiales (datos textuales, con alta dimensionalidad y/o gran tamaño).

T.B.1. Representación de documentos mediante datos multi-instancia

Descripción: Se analizarán distintas representaciones de documentos en formato multi-instancia y se comparará su rendimiento frente al de la representación convencional en las tareas de clasificación y *clustering* de documentos.
Participantes: **María Luque,** Amelia Zafra, Carmen Luque, Sebastián Ventura
Temporalización: M07-M18
Objetivo relacionado: 2.a
Resultados esperados: Representaciones textuales basadas en el uso de multi-instancias e informe de rendimiento de las mismas.

T.B.2. Algoritmos de selección de características para aprendizaje muti-etiqueta

Descripción: Diseño de métodos de selección de características para abordar datos con alta dimensionalidad para aprendizaje multi-etiqueta.
Participantes: **Eva Gibaja**, Amelia Zafra, María Luque, Oscar G. Reyes, Alain Guerrero
Temporalización: M13-M20
Objetivo relacionado: 2.b
Resultados esperados: Métodos para selección de características en aprendizaje multi-etiqueta.

T.B.3. Algoritmos de selección de instancias en aprendizaje multi-instancia y/o multi-etiqueta

Descripción: Se desarrollarán métodos de selección de instancias en aprendizaje multi-instancia y/o multi-etiqueta

Participantes: **Amelia Zafra**, Eva Gibaja, María Luque, Mykola Pecheniskyi, Oscar G. Reyes, Alain Guerrero
Temporalización: M19-M27
Objetivo relacionado: 2.b
Resultados esperados: Procedimientos de selección de instancias para aprendizaje multi-instancia y/o multi-etiqueta.

T.B.4. Desarrollo de modelos de aprendizaje para Big Data

Descripción: El objetivo de esta tarea es desarrollar modelos de aprendizaje con representaciones flexibles para contextos con una cantidad masiva de datos.
Participantes: **Sebastián Ventura**, Alberto Cano, José María Luna
Temporalización: M13-M18 + M25-M30
Objetivo relacionado: 2.c
Resultados esperados: Modelos de aprendizaje para Big Data en problemas que usan representaciones flexibles.

C. Resolución de problemas reales

Las tareas de tipo C consistirán en la resolución de problemas reales, de los ámbitos académico y sanitario, mediante la aplicación de los algoritmos y modelos desarrollados en las tareas A.1-A.6 y B1-B4.

T.C.1. Predicción del rendimiento académico

Descripción: Predecir el rendimiento académico utilizando aprendizaje multi-fuente, multi-vista, bases de datos relacionales, post minería de datos y meta-aprendizaje.
Participantes: **Cristóbal Romero**, Amelia Zafra, Mykola Pecheniskyi, Aurora Ramírez
Temporalización: M04-M12 + M30-M33
Objetivo relacionado: 3.a
Resultados esperados: Nuevos modelos de predicción del rendimiento académico.

T.C.2. Modelado de la autoevaluación y la evaluación por pares

Descripción: Modelar la autoevaluación y la evaluación por pares utilizando técnicas multi-fuente, multi-instancia, asociación, descubrimiento de subgrupos, agrupamiento, detección de *outliers* y regresión.
Participantes: **Cristóbal Romero**, José R. Romero, Aurora Ramírez
Temporalización: M15-M27
Objetivo relacionado: 3.a
Resultados esperados: Nuevos modelos de autoevaluación y evaluación por pares.

T.C.3. Recomendación de recursos didácticos para personalización de cursos

Descripción: Recomendar recursos didácticos, actividades y/o cursos a los estudiantes utilizando técnicas multi-fuente, multi-instancia y multi-etiqueta.
Participantes: **Cristóbal Romero**, Eva Gibaja, José R. Romero, Aurora Ramírez
Temporalización: M19-M30
Objetivo relacionado: 3.a
Resultados esperados: Nuevos modelos de recomendación de recursos didácticos.

T.C.4. Predicción de enfermedades mediante clasificación de historias clínicas

Descripción: En esta tarea se desarrollarán modelos basados en clasificación de documentos para el diagnóstico temprano de enfermedades.
Participantes: **Sebastián Ventura**, Eva Gibaja, María Luque, Krys Cios, Carmen Luque
Temporalización: M13-M33
Objetivo relacionado: 3.b
Resultados esperados: Modelos de predicción de enfermedades.

T.C.5. Obtención de modelos predictivos en biomedicina

Descripción: Esta tarea consiste en el desarrollo de modelos predictivos para detección de enfermedades del metabolismo de la insulina a partir de datos no textuales.
Participantes: **Sebastián Ventura**, Krys Cios, Mykola Pecheniskyi, Carmen Luque
Temporalización: M13-M33
Objetivo relacionado: 3.b

Resultados esperados: Modelos predictivos en distintas enfermedades relacionadas con el metabolismo de la insulina.

D. Desarrollo de repositorios de datos e integración de las propuestas desarrolladas

Este último bloque comprende el desarrollo de repositorios con los datos que se hayan empleado en las tareas de tipo A, B y C. La tarea también comprende el desarrollo de algunas utilidades de manejo de datos y la integración de nuestro software en las plataformas más empleadas en la actualidad.

T.D.1. Repositorios de datos para las distintas tareas de aprendizaje

Descripción: Creación de repositorios con *datasets* para aprendizaje multi-instancia, multi-etiqueta, relacional, multi-vista / multi-fuente y minería de asociaciones.
Participantes: **Cristóbal Romero**, Oscar G. Reyes, Hermes Robles, Carmen Luque
Temporalización: M10-M12 + M22-M24 + M31-M36
Objetivo relacionado: 4.a, 4.b y 4.c
Resultados esperados: Uno o varios repositorios con los datos empleados en cada una de las investigaciones relativas a las tareas de tipo A, B y C.

T.D.2. Herramienta para caracterización de conjuntos de datos multi-etiqueta

Descripción: Se trata de desarrollar una herramienta que permita el cálculo de distintas métricas de evaluación de *datasets* multi-etiqueta, generando un informe de descripción de los mismos, de cara a la publicación de los conjuntos de datos en el repositorio anterior.
Participantes: **Eva Gibaja**, Alain Guerrero
Temporalización: M01-M06
Objetivo relacionado: 4.d
Resultados esperados: Herramienta para caracterización de *datasets* multietiqueta.

T.D.3. Herramienta para simulación de conjuntos de datos multi-vista

Descripción: Se trata de desarrollar una herramienta para la generación de conjuntos de datos multi-vista, que permita la gestión de la información y su particionamiento desde una interfaz gráfica.
Participantes: **Amelia Zafra**, Alberto Cano
Temporalización: M01-M06
Objetivo relacionado: 4.e
Resultados esperados: Herramienta para generación de conjuntos de datos multi-vista.

T.D.4. Integración de algoritmos en WEKA y/o KEEL

Descripción: En esta tarea implementaremos nuestras propuestas para su uso en los sistemas WEKA y/o KEEL.
Participantes: **José R. Romero**, Alberto Cano, José M. Luna; Oscar Reyes, Alain Guerrero
Temporalización: M10-M12 + M22-M24
Objetivo relacionado: 5.a
Resultados esperados: Implementaciones de nuestras propuestas listas para funcionar en los entornos WEKA y/o KEEL.

T.D.5. Integración de algoritmos de Big Data en MAHOUT / SPARK

Descripción: Implementación de los resultados de la tarea B.4 en las plataformas MAHOUT / SPARK
Participantes: **José R. Romero**, Alberto Cano, José M. Luna, Aurora Ramírez, Hermes Robles
Temporalización: M19-M24 + M31-M36
Objetivo relacionado: 5.b
Resultados esperados: Implementaciones de nuestras propuestas listas para funcionar en los entornos MAHOUT y/o SPARK.

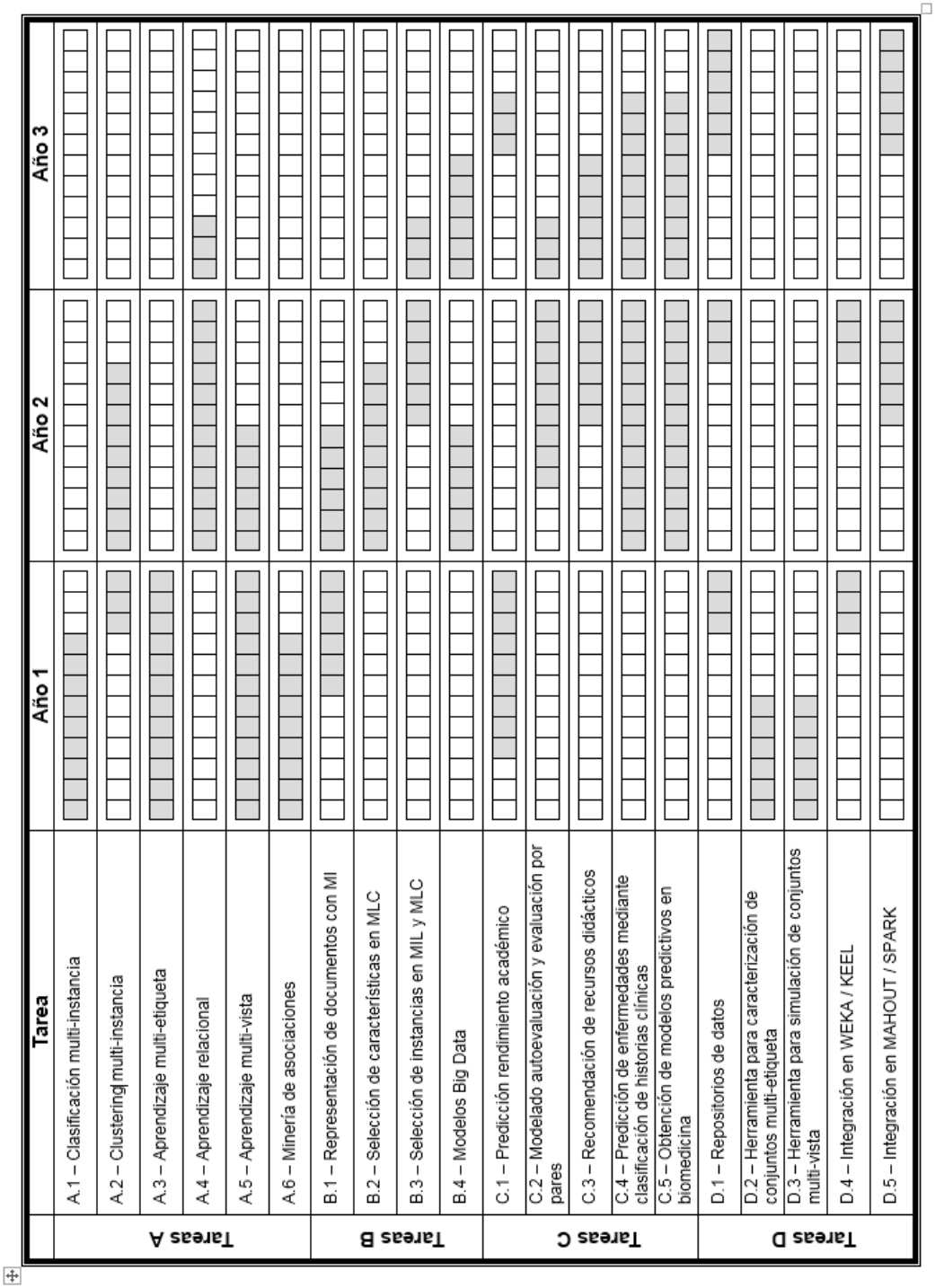
5. Medios materiales, infraestructuras y equipamientos singulares

Con respecto a los medios materiales a emplear en este proyecto indicar que, para poder trabajar en condiciones óptimas, se necesitaría un servidor de cómputo potente sobre el que

implantar software de distribución de colas de procesos, del estilo de HTCcondor (http://research.cs.wisc.edu/htcondor/) o Grid Engine (http://gridscheduler.sourceforge.net/) para automatizar las múltiples baterías de pruebas que se realizarán en las distintas fases de experimentación del proyecto. En este clúster de máquinas se podría instalar también HADOOP y las herramientas Mahout y SPARK, para realizar toda la experimentación relativa a *big data*. El equipo de investigación dispone de un clúster para computación adquirido con los fondos del último proyecto de investigación, concedido en el año 2011. Considerando el ritmo al que avanza la tecnología, y la alta obsolescencia de este tipo de equipos, sería muy deseable disponer de un nuevo clúster como el descrito para disponer de una capacidad de cómputo que nos permita avanzar en la investigación a un ritmo apropiado. Por esta razón, hemos incluido esta máquina en el presupuesto del proyecto. Por lo demás, consideramos que no necesitamos ninguna otra infraestructura que sea digna de mención.

6. Cronograma

La siguiente figura muestra el cronograma de las tareas en las que se organiza este proyecto.



Como puede comprobarse, la mayor carga de trabajo se encuentra durante los primeros 30 meses del proyecto, estando organizadas las tareas en función de la disponibilidad de los miembros de los distintos equipos que se forman para trabajar en cada temática. Hemos preferido organizar el cronograma de este modo para (i) poder reaccionar ante imprevistos que retrasen la ejecución y (ii) dedicar los últimos 6 meses de proyecto a trabajar intensivamente en las tareas de construcción de repositorios de datos y de implementación de algoritmos en Mahout / SPARK, que suponemos tendrá un considerable coste de tiempo.

C.2. Impacto esperado de los resultados

1. Impactos científico-técnico, social y/o económico

El beneficio científico más relevante de este proyecto es el desarrollo de modelos más flexibles para la resolución de nuevos retos en el ámbito del descubrimiento de

conocimiento. De hecho, como se ha comentado previamente, queda mucho trabajo por hacer en estas líneas de investigación, y esperamos que nuestra contribución pueda suponer un avance real dentro de este campo.

Por otra parte, consideramos que la aplicación de nuestros modelos a la resolución de problemas reales puede tener gran interés para nuestro entorno. En este sentido, las cuestiones que se abordarán en el contexto de la minería de datos educativos tienen un enorme impacto social. Por ejemplo, la predicción del rendimiento académico, incluyendo el fracaso escolar (*student drop-out*) es un problema que interesa tanto a los estudiantes como a las instituciones educativas. Por otra parte, el modelado de la autoevaluación y el desarrollo de técnicas de recomendación de cursos pueden contribuir a la mejora de la enseñanza online, la cual está tomando cada vez un mayor auge en todos los ámbitos de la sociedad (por ejemplo, con la aparición de los MOOCs) porque permiten dar soporte a una enseñanza en línea personalizada y de calidad.

Por otra parte, los problemas del ámbito de la biomedicina tienen un enorme impacto en la sociedad, por el interés que ésta tiene en todos los temas relacionados con salud y calidad de vida. Así, los modelos de diagnóstico precoz suponen una gran ventaja tanto para el paciente como para el sistema sanitario. Pero además, los modelos obtenidos pueden arrojar nueva luz en la comprensión de las enfermedades y suponer un avance en el desarrollo de la ciencia médica.

Esperamos que nuestras contribuciones puedan redundar en un avance para la resolución de estos problemas y que nos permitan ponernos en contacto con los distintos agentes socio-económicos para transferir el conocimiento obtenido mediante el desarrollo de aplicaciones software, asesoramiento, etc.

2. Plan de difusión

El plan de difusión de los resultados consistiría en las siguientes acciones:

- Realización y publicación de *tesis doctorales* desarrolladas en el contexto del proyecto. Esperamos que a lo largo de estos tres años de proyecto se defiendan 3-4 tesis doctorales (idealmente, deberían defenderse todas las tesis de personal del equipo no doctor, aunque consideramos que esta cifra es satisfactoria).
- La *asistencia a conferencias nacionales e internacionales* relacionadas con las áreas temáticas de este proyecto, para mostrar los resultados obtenidos a la comunidad científica y contactar con otros grupos de investigación de intereses similares de cara a futuras colaboraciones. A nivel nacional, asistiremos al Taller de Minería de Datos (TAMIDA) y al Congreso de Algoritmos Bioinspirados y Metaheurísticas (MAEB). A nivel internacional, nos centraremos en conferencias de primer nivel (core A* o A) relacionadas con aprendizaje automático, y extracción de conocimiento en bases de datos (incluidos *Big Data* y *Data Science*).

- *Publicación en revistas* relacionadas con la temática del proyecto. En función de las contribuciones científico-técnicas esperadas, el proyecto debería producir entre 9 y 15 trabajos en revistas internacionales de primera línea. Considerando el currículo de los participantes en el proyecto podemos observar su experiencia en la publicación en este tipo de revistas.
- *Publicación de capítulos de libros* en editoriales internacionales de prestigio. En general serán capítulos por invitación o sometidos a un proceso de revisión por pares análogo al de la publicación en revistas.
- Por último, *trasladaremos algunos de los resultados obtenidos a la sociedad* mediante el envío de notas de prensa o haciendo uso de los canales disponibles para la difusión de los resultados que sean directamente transferibles a la sociedad.

3. Transferencia de resultados

Con respecto a la transferencia de resultados al entorno productivo, considerando el interés socio-económico que tienen los problemas que se van a resolver, es posible establecer

colaboraciones con empresas para transferir estos resultados en forma de proyectos con empresas. En este sentido, hemos de indicar que el grupo editorial Santillana (http://www.santillana.com/) ha mostrado un gran interés por los modelos de predicción de rendimiento académico que pueden derivarse de la consecución de este proyecto. Por otra parte, los modelos de biomedicina pueden dar también lugar a desarrollos de interés comercial, como muestra el interés que ha mostrado en nuestro proyecto la empresa CORESOFT S.L. (http://www.coresoft.es/).

C.3. Capacidad formativa del equipo solicitante

La capacidad formativa de nuestro equipo de investigación está garantizada si consideramos las siguientes cuestiones:

- 10 de los 15 miembros del equipo son doctores, y los 5 restantes son estudiantes de doctorado, 2 de los cuáles tienen su tesis en un estado avanzado (esperamos que su defensa tenga lugar durante el 2016):
    - Hermes Robles Berumen, dirigido por los Dres. Ventura y Zafra
    - Oscar G. Reyes, dirigido por el Dr. Ventura
- En los últimos años se han defendido siete tesis doctorales dentro del grupo de investigación (para una lista detallada de estas tesis doctorales, ver http://www.uco.es/grupos/kdis). La calidad de dichas tesis está avalada por las publicaciones descritas en los currículum de los miembros del equipo de investigación, habiendo sido algunas de ellas desarrolladas en colaboración con científicos de reconocido prestigio internacional como los Dres. Krzysztof Cios y Mykola Pecheniskyi, los cuales forman parte del equipo de trabajo del proyecto.
- Todos los miembros del equipo participan en el programa de doctorado *"Computación Avanzada, Energía y Plasmas"* de la Universidad de Córdoba (línea de trabajo *"Aprendizaje Automático, Minería de Datos y Modelado de Sistemas"*), lo que permite la incorporación de nuevos doctorandos en temas relativos al proyecto de investigación. Por otra parte, los Dres Ventura y Cios están trabajando en la elaboración de un proyecto de doctorado dual UCO-VCU que permitirá a los alumnos obtener el título de doctor en ambas universidades, lo cual debe revertir positivamente en la visibilidad del programa y en la demanda social del mismo.

En resumen, considerando la ratio entre miembros doctores y no doctores del equipo y las cuestiones expuestas anteriormente, pensamos que la incorporación de nuevos estudiantes de doctorado al programa es muy positiva, y que su formación para la obtención del grado de doctor quedará plenamente garantizada.

C.4. Implicaciones éticas y/o de bioseguridad

Con respecto a la esta cuestión, sólo indicar que los datos que vamos a recibir en relación a los problemas de aprendizaje y/o diagnóstico médico no dispondrán de información confidencial/sensible, dado que sus propietarios se encargarán de anonimizarlos y prepararlos apropiadamente antes que lleguen a nuestro poder. Por esta razón, hemos marcado “no” en todas las cuestiones relativas a ímplicaciones de tipo ético de nuestra investigación.

INSTRUCCIONES PARA RELLENAR LA MEMORIA CIENTÍFICO-TÉCNICA

AVISO IMPORTANTE

En virtud del artículo 11 de la convocatoria **NO SE ACEPTARÁN NI SERÁN SUBSANABLES MEMORIAS CIENTÍFICO-TÉCNICAS** que no se presenten en este formato.

Este documento está preparado para que pueda rellenarse en el formato establecido como obligatorio en las convocatorias (artículo 11.7.a): letra Times New Roman o Arial de un tamaño mínimo de 11 puntos; márgenes laterales de 2,5 cm; márgenes superior e inferior de 1,5 cm; y espaciado mínimo sencillo. La parte C ("Documento científico") de la memoria deberá tener una extensión máxima de 20 páginas, incluidos todos sus apartados. No se admitirán memorias con contenidos propios de la parte C incluidos en las partes A o B.

La memoria consta de tres partes: la parte A contiene información general y básica de la propuesta; la parte B contiene la relación de los componentes del equipo de trabajo (excepto doctores) y la información específica sobre la financiación pública y privada del equipo de investigación; y la parte C es el documento científico propiamente dicho.

Con carácter general:

1. Las memorias pueden rellenarse en español o en inglés, a excepción de la parte A: RESUMEN DE LA PROPUESTA/SUMMARY OF THE PROPOSAL, que debe rellenarse en ambos idiomas.
2. Se recomienda rellenar la memoria empleando un ordenador con sistema operativo Windows y usando como procesador de textos MS Word (MS Office).
3. Una vez terminada la memoria en Word, deberá convertir el archivo en formato pdf (de no más de 4Mb) y aportarlo en la aplicación informática de solicitud del proyecto en el apartado Añadir documentos > Memoria científico-técnica.

Parte A: RESUMEN DE LA PROPUESTA/SUMMARY OF THE PROPOSAL

Toda la información de este apartado deberá también rellenarse en la aplicación de solicitud para que los campos puedan explotarse informáticamente, aunque se incluyen también en la memoria para facilitar las tareas de evaluación. Se aconseja que se utilice el *copiar* y *pegar* desde la memoria hasta la aplicación informática de solicitud o viceversa para que no haya inconsistencias en el contenido de los textos.

Todos los campos de este apartado deberán rellenarse obligatoriamente en inglés y en español.

El resumen de la propuesta/summary of the proposal (con un máximo de 3500 caracteres, contando los espacios en blanco) contendrá los aspectos más relevantes de la propuesta, así como los objetivos planteados y los resultados esperados. Su contenido podrá ser publicado a efectos de difusión si el proyecto fuera financiado en esta convocatoria, salvo que haya indicado expresamente en la aplicación de solicitud que existen resultados susceptibles de ser protegidos.

Parte B: INFORMACIÓN ESPECÍFICA DEL EQUIPO

B.1. RELACIÓN DE LAS PERSONAS NO DOCTORES QUE COMPONEN EL EQUIPO DE TRABAJO

No se relacionarán en este apartado los datos del personal perteneciente al equipo de investigación ni los datos de los doctores pertenecientes al equipo de trabajo, puesto que esas personas deberán incluirse en la aplicación informática de solicitud.

Deberán rellenarse los siguientes datos del personal perteneciente al equipo de trabajo, excepto los doctores, repitiendo la secuencia que se indica a continuación tantas veces cuantas se necesite. En los campos de titulación, tipo de contrato y duración del contrato deberá tachar o borrar las claves que no procedan.

1. Nombre y apellidos:
   Titulación: licenciado/ingeniero/graduado/máster/formación profesional/otros (especificar)
   Tipo de contrato: en formación/contratado/técnico/entidad extranjera/otros (especificar)
   Duración del contrato: indefinido/temporal

B.2. FINANCIACIÓN PÚBLICA Y PRIVADA (PROYECTOS Y CONTRATOS DE I+D+I) DEL EQUIPO DE INVESTIGACIÓN

Deberá relacionar los proyectos y/o contratos de I+D+I en los que hayan participado los componentes del equipo de investigación y que hayan recibido financiación o que estén pendientes de resolución, en los últimos 8 años, en convocatorias de ámbito nacional, autonómico o internacional hasta un máximo de 10 proyectos y/o contratos. Si la relación fuera muy extensa, se recomienda seleccionar aquellos que estén más directamente relacionados con la propuesta que se presenta.

Deberán rellenarse los siguientes datos repitiendo la secuencia que se indica a continuación tantas veces como se necesite. En los campos de relación temática con el proyecto que se presenta y estado del proyecto o contrato deberá tachar o borrar las claves que no procedan:

1. Investigador del equipo de investigación que participa en el proyecto/contrato (nombre y apellidos):
   Referencia del proyecto:
   Título:
   Investigador principal (nombre y apellidos):
   Entidad financiadora:
   Duración (fecha inicio - fecha fin, en formato DD/MM/AAAA):
   Financiación recibida (en euros):
   Relación temática con el proyecto que se presenta: mismo tema/está muy relacionado/está algo relacionado/sin relación
   Estado del proyecto o contrato: concedido/pendiente de resolución

Parte C: DOCUMENTO CIENTÍFICO

La parte C de la memoria científico-técnica es la única que está limitada en cuanto a extensión. Los cuatro apartados de la parte C no podrán superar las 20 páginas, debiendo mantenerse además los márgenes, espaciado y tipo de letra establecidos en la convocatoria. Se recuerda que no se admitirán memorias con contenidos propios de la parte C incluidos en otras partes del documento. En su caso, los anexos, imágenes, tablas, fórmulas, etc. estarán incluidos en la parte C.

C.1. PROPUESTA CIENTÍFICA

Se recomienda incluir:

1. Los **antecedentes y estado actual** de los conocimientos científico-técnicos de la materia específica del proyecto, incluyendo, en su caso, los resultados previos del equipo de investigación y la relación, si la hubiera, entre el grupo solicitante y otros grupos de investigación nacionales y extranjeros.

Si el proyecto es continuación de otro previamente financiado, deben indicarse con claridad los objetivos y los resultados ya alcanzados de manera que sea posible evaluar el avance real que se propone en el nuevo proyecto. Si el proyecto aborda un tema nuevo, deben indicarse los antecedentes y contribuciones previas del equipo de investigación que justifiquen su capacidad para llevarlo a cabo.

2. La **hipótesis de partida** y los **objetivos generales** perseguidos, así como la **adecuación** del proyecto a la Estrategia Española de Ciencia y Tecnología y de Innovación y, en su caso, a Horizonte 2020 o a cualquier otra estrategia nacional o internacional de I+D+i.

Si la memoria se presenta a la convocatoria de RETOS, deberá identificarse el reto cuyo estudio se pretende abordar y la relevancia social o económica prevista.

3. Los **objetivos específicos**, enumerándolos brevemente, con claridad, precisión y de manera realista (acorde con la duración prevista del proyecto).

En los proyectos con dos investigadores principales, deberá indicarse expresamente de qué objetivos específicos se hará responsable cada uno de ellos.

4. El detalle de la **metodología** propuesta, incluyendo la viabilidad metodológica de las tareas. Si fuera necesario, también se incluirá una evaluación crítica de las posibles dificultades de un objetivo específico y un plan de contingencia para resolverlas.

5. La descripción de los **medios materiales, infraestructuras y equipamientos singulares** a disposición del proyecto que permitan abordar la metodología propuesta.

6. Un **cronograma** claro y preciso de las fases e hitos previstos en relación con los objetivos planteados en la propuesta.

7. Si se solicita ayuda para la **contratación de personal, justificación** de su necesidad y descripción de las tareas que vaya a desarrollar.

C.2. IMPACTO ESPERADO DE LOS RESULTADOS

El contenido de este apartado se solicitará también en la aplicación informática de solicitud (con un máximo de 3500 caracteres) y su contenido podrá ser publicado a efectos de difusión si el proyecto fuera financiado en esta convocatoria.

Se recomienda incluir:

1. Descripción del **impacto científico-técnico social y/o económico** que se espera de los resultados del proyecto, tanto a nivel nacional como internacional.

2. El **plan de difusión** e internacionalización en su caso de los resultados.

3. Si se considera que puede haber **transferencia de resultados**, se deberán identificar los resultados potencialmente transferibles y detallar el plan previsto para la transferencia de los mismos.

C.3. CAPACIDAD FORMATIVA DEL EQUIPO SOLICITANTE

Este apartado solo se rellenará si se solicita la inclusión del proyecto en la convocatoria de "Contratos predoctorales para la formación de doctores". Dicha inclusión solo será posible en un número limitado de los proyectos aprobados.

Para evaluar la capacidad formativa del equipo solicitante, se recomienda incluir:

1. El **plan de formación previsto** en el contexto del proyecto solicitado.

2. Relación de tesis realizadas o en curso (últimos 10 años) con indicación del nombre del doctorando, el título de tesis y la fecha de obtención del grado de doctor o de la fecha prevista de lectura de tesis.

3. Breve **descripción del desarrollo científico o profesional de los doctores egresados** del equipo de investigación.

C.4. IMPLICACIONES ÉTICAS Y/O DE BIOSEGURIDAD

Este apartado solo se rellenará si en la aplicación electrónica de solicitud se contesta afirmativamente a alguno de los aspectos relacionados con implicaciones éticas y/o de bioseguridad allí recogidos.

Se recomienda incluir:

1. Una **descripción de los aspectos éticos** referidos a la investigación que se propone.

2. Una explicación de las **consideraciones, procedimientos o protocolos** a aplicar en cumplimiento de la normativa vigente, así como una descripción de las instalaciones y las preceptivas autorizaciones de las que se dispone para la ejecución del proyecto.