RE-IMPLEMENTACIÓN DEL PAQUETE AMORE

Javier Alfonso
Manuel Castejón
Universidad de León
Andrés Sanz
Julio Fernández
Roberto Fernández
Universidad de La Rioja

Abstract

The AMORE package for neural network training and simulation has been developed by the EDMANS research group as a natural consequence of the research projects developed by the group. The virtues and usefulness of the package in its current state of development have been recognized not only by the EDMANS group itself, where it has become a widely used tool, but by the scientific community abroad our borders, as the citing data evidences. Nevertheless, the consolidation of the PMML specification, currently in version 4.0, as a de facto standard, as well as the recent extension of the R language to the object oriented programming field, claims for an upgrade of the package. This paper shows some of the main features that the new version of the AMORE package will display once these improvements are implemented following this new approach.

Keywords*: EDMANS, data mining, neural networks, quality, industrial processes.*

Resumen

La librería AMORE para el entrenamiento y simulación de redes neuronales ha sido desarrollada por el grupo de investigación EDMANS dentro del contexto de los proyectos de investigación desarrollados en el seno del grupo. La utilidad y virtudes de la librería en su estado actual ha sido reconocida no sólo dentro del grupo EDMANS, donde es una herramienta de trabajo habitual, sino también fuera de él, como muestran las citas bibliográficas realizadas por diversos autores en revistas de impacto. Sin embargo, la consolidación de la especificación PMML, actualmente en versión 4.0, como norma de facto para el intercambio de modelos en el contexto de la minería de datos, así como la reciente extensión del lenguaje R en el campo de la programación orientada a objetos, reclama la actualización de la librería. El presente artículo expone algunas de las principales virtudes que caracterizarán a la nueva versión de la librería AMORE una vez implementadas las mejoras correspondientes a esta nueva orientación.

Palabras clave: *EDMANS, minería de datos, redes neuronales, calidad, procesos industriales.*

1. Introducción

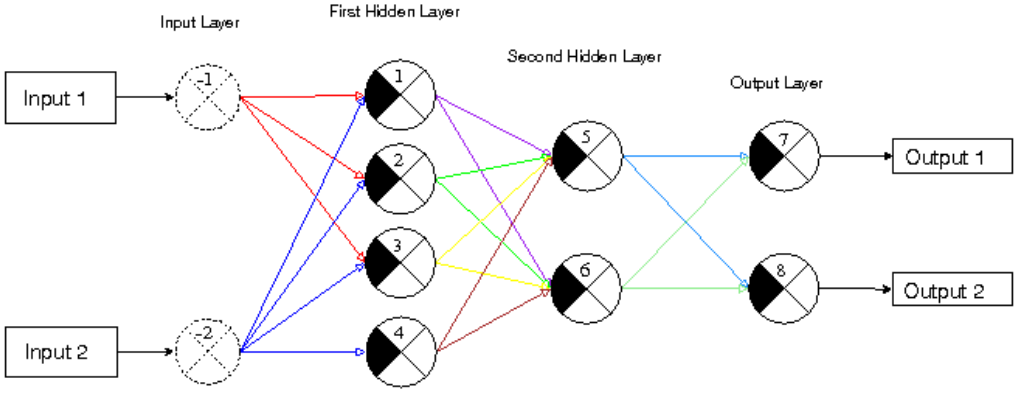
El paquete AMORE fue desarrollado por el grupo EDMANS, con el objetivo de proporcionar al usuario una potente herramienta para el entrenamiento, y la simulación de redes

neuronales, que permitiera obtener un control total de la red, accediendo directamente a cada uno de los parámetros de la misma, permitiendo configurar y personalizar sus diferentes funciones y características, con el fin de adaptarlas a las necesidades específicas de cada usuario.

Las primeras especificaciones del paquete AMORE fueron desarrolladas utilizando el lenguaje S, en su tercera versión S3 para el software R, pero dada la gran aceptación del paquete, y debido a su creciente número de citas en revistas indexadas, surgió rápidamente la necesidad de seguir mejorando su implementación, desarrollando nuevas características y funcionalidades, aumentando así su potencia (Castejón Limas, y otros, 2009)l.

Después de lo dicho anteriormente, y debido a la aparición de la nueva versión de S (S4), se tomó la decisión de re-implementar la librería AMORE, mejorando la misma mediante el potencial que aporta S4, y utilizando la especificación PMML (Predictive Model Markup Language) 4.0 para redes neuronales, que como se analizará más adelante se ha convertido en un estándar de facto para el intercambio de modelos en el contexto de la minería de datos.

Figura 1: Ejemplo de la estructura de una red neuronal simulada con la librería AMORE



En la Figura 1, se puede observar la estructura de una red neuronal, y los elementos principales que intervienen en su simulación utilizando la librería AMORE.

2. El paquete AMORE

El paquete AMORE es una herramienta destinada al entrenamiento y simulación de redes neuronales cuya diferencia con otras alternativas de extendido uso es la flexibilidad con la que cuenta el usuario para adaptar la estrategia de aprendizaje a sus necesidades. Éste es uno de los objetivos fundamentales del paquete y una de las causas de su creación.

Así, a diferencia de esas otras alternativas, el usuario del paquete AMORE puede adaptar las estrategias de aprendizaje a sus necesidades mediante la sencilla programación en lenguaje R de los complementos necesarios. De este modo, resulta especialmente sencillo modificar las funciones de coste a utilizar durante el aprendizaje a fin de lograr satisfacer los requisitos impuestos por el usuario. Esta característica es fundamental, por ejemplo, para satisfacer algunos de los problemas reales ante los que los investigadores del grupo EDMANS se han enfrentado: modelización de procesos industriales con un enfoque basado en datos, siendo el grado de contaminación de estos datos suficientemente elevado como

para desestimar el uso de las funciones de coste tradicionales que otras alternativas tan sólo emplean.

La curva de aprendizaje del paquete AMORE hace cómodo también su uso para el usuario novel no iniciado en el lenguaje del entorno R, pudiendo desde un principio acceder a funciones de simulación y entrenamiento de redes neuronales típicas: creación de redes multicapas, algoritmos de aprendizaje tipo backprogation con y sin momentum, funciones de coste LMS, LMLS, TAO-robust, modo batch y adaptativo, etc.

3. R Software

R es un paquete de programas integrados para el manejo de datos, simulaciones, cálculos, y realización de gráficos. Es además un lenguaje de programación orientado a objetos. R es una implementación libre (open-source), e independiente del lenguaje de programación S, que en la actualidad es un producto comercial llamado S-PLUS, y que es distribuido por Insinhtful Corporation (Muenchen & Hilbe, 2010).

El lenguaje S, fue desarrollado a mediados de los años 70 en Bell Laboratories (antigua AT&T, y actual Lucent Technologies) por John Chambers y sus colaboradores. Originalmente fue un programa para el sistema operativo Unix, aunque actualmente se pueden obtener también versiones para Windows, Macintosh y Linux. A pesar de que existen diferencias entre R, y S-PLUS (principalmente en la interfaz gráfica), son esencialmente idénticos, y la mayoría del código escrito en S funciona en R.

El proyecto R fue iniciado por Robert Gentleman, y Ross Ihaka (de cuyas iniciales deriva "R") del Departamento de Estadística, de la Universidad de Auckland en 1995. Actualmente R es mantenido por una comunidad internacional de desarrolladores voluntarios, el R Core Development Team. La versión actual de R, es la 2.10.1. En la Figura 2, se puede observar el logo del Proyecto R.

Figura 2: Logo del proyecto R

3.1 Ventajas de S4

La característica fundamental de la nueva especificación de S (S4), en comparación con su versión anterior S3, es su capacidad para implementar funciones que permiten considerar a S como un lenguaje orientado a objetos, por lo tanto las ventajas de utilizar S4, serán las ventajas que aporta la programación orientada a objetos.

La programación orientada a objetos es lo que se conoce como un paradigma o modelo de programación, estos significa que no es un lenguaje específico, o una tecnología, sino una forma de programar, una manera de plantearse la programación. La programación orientada a objetos es una forma especial de programar, más cercana a como expresaríamos las cosas en la vida real que otros tipos de programación. En los siguientes puntos se pueden resumir las ventajas más importantes de este tipo de programación, y por tanto de la cuarta especificación del lenguaje S:

- Fomentar la reutilización y extensión del código.
- Permite crear sistemas más complejos.
- Relacionar el sistema al mundo real.
- Construcción de prototipos.
- Agiliza el desarrollo de software.
- Facilita el trabajo en equipo.
- Facilita el mantenimiento del software.

4. PMML

En la actualidad, el intercambio de información tiene tal interés, que nuestra era se ha denominado como la era de la información, y del conocimiento, pero para intercambiar toda esta información es necesario un marco común que facilite el entendiendo, objetivo que se alcanzará mediante la definición y la utilización de estándares.

La minería de datos obtiene conocimiento a partir de grandes volúmenes de datos, y costosos análisis de los mismos, por lo que la posibilidad de intercambiar modelos aprendidos entre distintos grupos resulta de un gran interés. Para que dos aplicaciones de minería de datos sepan intercambiar modleos, éstos han de ser definidos de algún modo común, mediante un interfaz que ambos sistemas sean capaces de entender (Grossman, Hornick, & Meyer, 2002).

Algunos estándares desarrollados para el intercambio de modelos entre aplicaciones de minería de datos son XML for Analysis, JSR 73, SQL/MM, y el PMML (Predictive Model Markup Language), que está basado en el estándar XML del W3C.

PMML es un lenguaje de marcas basado en el estándar XML que sirve para describir modelos estadísticos y de mineria de datos, para lo que se definen los datos de entrada al modelo, las transformaciones realizadas sobre los mimos, y los parámetros propios que lo definen.

PMML se define sobre la base de un esquema de documento XML, de manera que se establece la estructura necesaria a informar para definir el modelo. En su última versión, la 4.0 el PMML define un esquema XML para cada tipo de problema de minería de datos, utilizándose en este caso el esquema XML de las redes neuronales, representado en la Figura 3 (DMG, 2010) (Pechter, 2009).

Figura 3: Esquema XML (PMML) de una red neuronal

```
<xs:element name="NeuralNetwork">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Extension" minOccurs="0" maxOccurs="unbounded" />
      <xs:element ref="MiningSchema"/>
      <xs:element ref="Output" minOccurs="0" />
      <xs:element ref="ModelStats" minOccurs="0"/>
      <xs:element ref="ModelExplanation" minOccurs="0"/>
      <xs:element ref="Targets" minOccurs="0" />
      <xs:element ref="LocalTransformations" minOccurs="0" />
      <xs:element ref="NeuralInputs" />
      <xs:element maxOccurs="unbounded" ref="NeuralLayer" />
      <xs:element minOccurs="0" ref="NeuralOutputs" />
      <xs:element ref="ModelVerification" minOccurs="0"/>
      <xs:element ref="Extension" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="modelName" type="xs:string" />
    <xs:attribute name="functionName" type="MINING-FUNCTION" use="required" />
    <xs:attribute name="algorithmName" type="xs:string" />
    <xs:attribute name="activationFunction" type="ACTIVATION-FUNCTION" use="required" />
    <xs:attribute name="normalizationMethod" type="NN-NORMALIZATION-METHOD" default="none"/>
    <xs:attribute name="threshold" type="REAL-NUMBER" default="0" />
    <xs:attribute name="width" type="REAL-NUMBER" />
    <xs:attribute name="altitude" type="REAL-NUMBER" default="1.0" />
    <xs:attribute name="numberOfLayers" type="xs:nonNegativeInteger" />
  </xs:complexType>
 </xs:element>
```

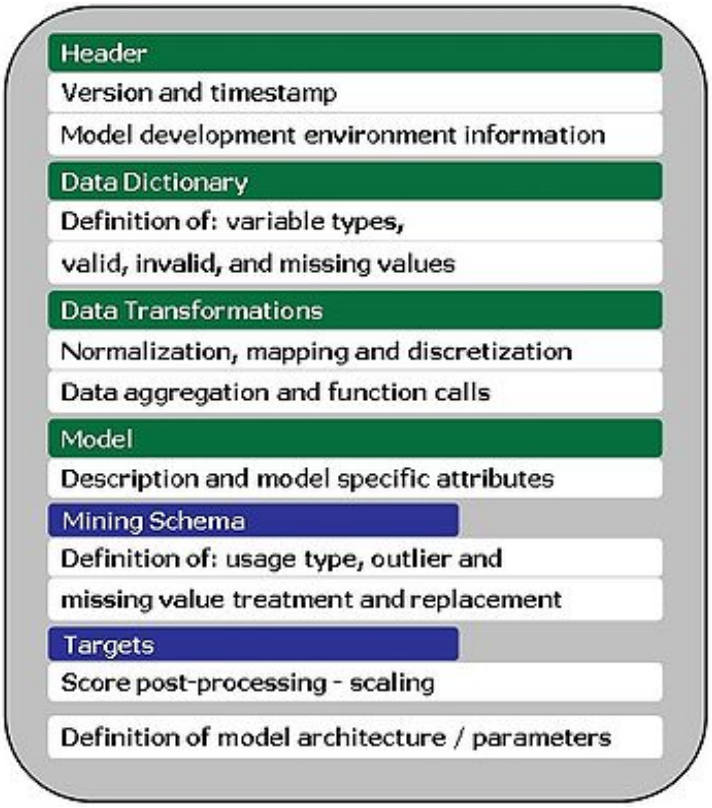
El esquema de documento PMML como se puede observar en la Figura 4 está compuesto por cuatro secciones, la cabecera, el diccionario de datos, las transformaciones de datos, y la definición del modelo. En la cabecera “<header>” se define la información sobre el propio archivo PMML, como el copyright y la descripción del modelo. El diccionario de datos “<datadictionary>" incluye los atributos o características del modelo, así como su tipo y posibles valores. Las transformaciones de datos “<datatransformations>” contienen las diferentes transformaciones que se aplicarán a los datos, y por último el modelo “<model>” donde se definen dos secciones, el esquema de minería que define los atributos que se utilizarán en el aprendizaje, indicando cuál de ellos hará de clase, y la propia definición del modelo que se realizará de manera recursiva mediante marcas que contienen información sobre las condiciones que debe satisfacer el ejemplo para entrar en ese nodo, y la predicción de la clase realizada en ese punto.

De la definición del PMML se encarga el Data Mining Group (DMG), grupo independiente de la industria, cuyo principal objetivo es el de definir y desarrollar estándares para minería de datos, y en concreto el estándar PMML. El grupo DMG está formado por un conjunto bastante elevado de compañías importantes en el campo de la minería de datos como Angoss Software Corporation, Fair Isaac Corporation, Insightful Corporation, KXEN, Magnify Inc., Microsoft, MicroStrategy Inc., MINEit Software Ltd., Nacional Center for Data Minino, NCR Corp., Oracle Corp., Prudsys AG, Quadstone, SAP, SAS Inc., Salford Systems, SPSS Inc., StatSoft Inc. y Xchange Inc (DMG - Current Members, 2010).

La incorporación de PMML en sistemas comerciales todavía no ha sido masiva debido a su reciente desarrollo, pero siendo las empresas más importantes de la actualidad en el sector de la minería de datos las propulsoras de este estándar el grado de aceptación y su implantación a corto plazo está asegurado.

Es de prever que la implantación de la opción de intercambio mediante PMML sea un hito de todo sistema de minería de datos y que en un futuro próximo todos los sistemas de minería lo incluyan, aumentando las posibilidades del tratamiento de la información y consiguiendo crear una sensación total de control del conocimiento del negocio.

Figura 4: Estructura de un documento PMML



5. Conclusiones

La re-implementación del paquete AMORE utilizando la nueva especificación S4 supone una mejora importante en el rendimiento del mismo, permitiendo también realizar un mantenimiento mucho más eficiente, y efectivo, favoreciendo la incorporación de nuevas funcionalidades, y agilizando su desarrollo.

La utilización del estándar PMML para redes neuronales, permite la importación de modelos ya desarrollados utilizando aplicaciones externas al paquete AMORE, así como exportar la información de los modelos y de las simulaciones realizadas con el mismo, favoreciendo así la interconexión entre sistemas, y por tanto la estandarización del sector.

6. Referencias

Castejón Limas, M., Ordieres Meré, J. B., Martínez de Pisón Ascacibar, F. J., Pernía Espinoza, A. V., Alba Elías, F., González Marcos, A., y otros. (19 de febrero de 2009). *The AMORE package: A MORE flexible neural network package.* Recuperado el 20 de 03 de 2010, de http://rwiki.sciviews.org/doku.php?id=packages:cran:amore

DMG - Current Members. (2010). Recuperado el 20 de marzo de 2010, de Data Mining Grop: http://www.dmg.org/about.html

DMG. (s.f.). *PMML 4.0 - Neural Network Models*. Recuperado el 20 de marzo de 2010, de http://www.dmg.org/v4-0/NeuralNetwork.html

Grossman, R. L., Hornick, M. F., & Meyer, G. (2002). Data mining standards initiatives. *Communications of the ACM*, 59-61.

Grupo EDMANS. (s.f.). *EDMANS - Engineering Datas Mining And Numerical Simulations*. Recuperado el 20 de marzo de 2010, de http://www.mineriadatos.org

Kabacoff, R. (2010). *R in Action.* Manning.

Muenchen, R. A., & Hilbe, J. M. (2010). *R for Stata Users. Statistics and Computing.* Springer.

Pechter, R. (2009). What's PMML and what's new in PMML 4.0? *ACM SIGKDD Explorations Newsletter* (págs. 19-25). New York: ACM.

Agradecimientos

Los autores agradecen el apoyo financiero del Ministerio de Ciencia y Tecnología a través del proyecto DPI2009-08424 perteneciente al programa Nacional de Investigación Fundamental de I+D+i 2008-2011, en el área temática de Diseño y Producción Industrial.

Correspondencia (Para más información contacte con):

Javier Alfonso Cendón
Escuela de Ingenierías Industrial e Informática
Universidad de León
Campus de Vegazana s/n
24071 León
Phone: +34 9871779
Fax: +34987291779
E-mail: javier.alfonso@unileon.es
URL: http://www.mineriadatos.com