Herramientas de Software Libre para el aprendizaje automático.

Ing. Corso, Cynthia Lorena, Ing. Gibellini Fabián
Departamento de Ingeniería en Sistemas de Información/Laboratorio de Sistemas de Información.
Facultad Regional Córdoba/Universidad Tecnológica Nacional.
Maestro M. López esq. Cruz Roja Argentina. Ciudad Universitaria
cynthia@bbs.frc.utn.edu.ar
speaker@bbs.frc.utn.edu.ar

RESUMEN

En este trabajo se ha realizado un estudio preliminar de las diversas opciones de software libre para el aprendizaje automático, seleccionando para su estudio, aquellas que tienen un mayor grado de popularidad. Se plantea como finalidad de este trabajo el establecimiento de parámetros que nos faciliten la comparación de las mismas; focalizando en aspectos como características, herramientas y técnicas implementadas para el aprendizaje automático; permitiendo establecer recomendaciones en que caso es más propicio el uso de las herramientas consideradas.

Palabras claves

Software Libre, Inteligencia Artificial, Aprendizaje Automático, Weka, Orange, RapidMiner, Knime.

INTRODUCCIÓN

[1] El avance importante que ha tenido el campo de la tecnología y el abaratamiento de costos ha traído como consecuencia un aumento significativo en la cantidad de datos que son almacenados en muchas ocasiones en diferentes formatos.

El aprendizaje automático es una rama de la inteligencia artificial cuyo propósito es la creación de técnicas que permitan a las computadoras aprender. Un aspecto vinculado con lo mencionado en el apartado anterior, se trata de la generación de programas que tengan la capacidad de generalizar comportamientos o tendencias a partir de una información proporcionada en forma de ejemplos.

El aprendizaje automático está relacionado con ramas como la estadística, pero se centra más en el estudio de la Complejidad computacional.

El ámbito de aplicación del aprendizaje automático es variado, como: motores de búsqueda, diagnósticos médicos, análisis de mercado de ventas, juegos y robótica entre otros.

Existen diferentes tipos de algoritmos en el aprendizaje automático y se agrupan teniendo en cuenta la salida de los mismos. A continuación se detallan algunos de ellos:

Aprendizaje supervisado: permite generar una función que hay una correspondencia entre la entrada y salidas deseadas. Un ejemplo que aplica este concepto son los algoritmos de clasificación, en la que el programa de aprendizaje intenta clasificar una serie de valores utilizando una entre varias categorías (clases).

Aprendizaje no supervisado: el procedimiento de modelado se realiza tomando como base un conjunto de ejemplos formado tan solo por entradas al sistema y no se cuenta con información sobre las categorías de las entradas.

Aprendizaje por refuerzo: el algoritmo aprende del mundo que lo rodea, es decir que su información de entrada es la retroalimentación que obtiene del mundo exterior como respuesta a sus acciones.

[2] Para concluir podemos decir que el Aprendizaje Automático está íntimamente relacionado el desarrollo de programas que sean capaces de mejorar su efectividad con la experiencia. Este tema es especialmente interesante por su aplicabilidad a todos los ámbitos que se nos ocurran.

La construcción de programas que sean capaces de aprender no es una tarea sencilla ni mucho menos simple, y de hecho, numerosos programas incorporan esta capacidad de una manera u otra. El problema reside en que muchos programas incorporan un algoritmo “ad-hoc”, es decir, diseñado específicamente para el problema que se trata y testeado de una manera limitada.

Esto es un problema porque existe literalmente una diversidad de algoritmos de aprendizaje, y no hay garantías de que en un programa concreto se esté utilizando el algoritmo más efectivo, es decir, el que garantiza mayor eficacia y rapidez.

[2] Para la experimentación con algoritmos de aprendizaje han surgido en los últimos años múltiples librerías de software que incorporan no sólo muchos de los algoritmos programados, sino además un entorno completo para evaluarlos tanto a nivel de eficacia como de rendimiento. Usando uno de estos entornos, el programador puede efectuar gran cantidad de pruebas que le permiten escoger el algoritmo más eficaz para su problema, con garantías de que será realmente útil en su aplicación.

METODOLOGÍA

La metodología usada para llevar a cabo este proyecto es la recopilación de diversas fuentes libros, publicaciones, páginas en internet para identificar cuáles son las alternativas de software más populares que se aplican en el campo del aprendizaje automático.

A continuación se detallan las alternativas que hemos considerado en este trabajo y se explica en forma breve cuales son las características de cada una de ellas:

Weka: (Waikato Enviroment for Knowledge Analysis) es una conocida suite de software para el aprendizaje que soporta varias tareas para el aprendizaje de automático, especialmente los datos del proceso previo (Prepocesamiento).

El agrupamiento, clasificación, regresión, visualización y selección de características son algunas de las funcionalidades incluidas en esta herramienta.

Sus técnicas se basan en la hipótesis de que los datos están disponibles en un único archivo plano o una relación, donde se etiqueta cada punto de datos por un número fijo de atributos.

WEKA proporciona acceso a bases de datos SQL utilizando Java Database Connectivity y puede procesar el resultado devuelto por una consulta de base de datos. Su interfaz de usuario principal es el Explorer, donde incluye todas las tareas esenciales para llevar a cabo cualquier proyecto de aprendizaje automático.

Orange: Es una biblioteca que está basada en componentes y escrita en Python y C++, desarrollada por el laboratorio de Inteligencia artificial de la Universidad de Liublania en Eslovenia.

[3] Proporciona una herramienta visual que facilita tareas para el análisis exploratorio de los datos y visualización. Además posee componentes que se encargan de realizar preprocesamiento de datos, filtrado de datos, modelado y evaluación de datos.

Su interfaz gráfica de usuario está basada en el marco de Qt multiplataforma.

Una de las fortalezas de esta herramienta es la gran diversidad de widgets implementados para la visualización de datos como para los modelos generados.

[8] **RapidMiner**: antes llamado YALE (Yet Another Learning Enviroment)

Esta herramienta visual se enfoca en la concepción de trabajo con nodos y dispone de una variedad de herramientas para el aprendizaje automático como Clasificación, Agrupamiento y Asociación.

Si bien esta herramienta dispone un volumen menor de algoritmos de Clasificación si lo comparamos con la herramienta Weka, pero como ventaja es que es su flexibilidad para integrarse con algoritmos implementados en Weka.

A la hora de crear un nuevo proyecto tenemos de opciones principales de interfaz:

- Repositorios [Repositories]: En esta sección permite la carga de la fuente de datos para el inicio del proyecto.
- Operadores [Operators]: Esta opción incluye diferente funciones, mencionamos alguna de ellas:
- Transformación de datos [Data Transformation], es decir herramienta de preprocesado de datos, Modelado [Modeling] en el que aparece una estructura jerárquica en forma de árbol las diferentes técnicas para el aprendizaje automático y Evaluación [Evaluation] que incluye funciones para determinar el nivel de confianza de los modelos obtenidos.

KNIME (o Konstanz Information Miner): Está construido bajo la plataforma Eclipse y programado esencialmente en Java. Dispone de una herramienta gráfica, conformada por un conjunto de nodos que encapsula los distintos algoritmos y fechas que representan el flujo de datos.

[7] KNIME fue desarrollado originalmente en el departamento de bioinformatica y minería de datos de la Universidad de Constanza, Alemania.

Esta herramienta se basa en el concepto del diseño de un flujo de ejecución en el que se reflejan las distintas etapas de un proyecto de minería de datos.

La interfaz principal de KNIME se presenta en paneles en forma de nodos, a continuación se detalla cuales son:

- Entrada de Datos [IO/Read]
- Salida de datos [IO/Read]
- Preprocesamiento [Data Manipulation]

  Las acciones que se pueden ejecutarse sobre una fuente de datos son:

  - Manipulación de filas, columnas: permite realizar operaciones de muestreos, transformaciones, agrupaciones etc.
  - Visualización: Histogramas.
  - Generación de modelos estáticos como: árboles de decisión, máquinas de vector de soporte y regresiones.

- Validación de modelos: utilización de métodos como curva ROC.
- Scoring: es decir aplicación de modelo de datos sobre conjuntos nuevos de datos.
- Generación de informe a medida gracias a su interacción con BIRT.

- Minería de datos [Mining]
- Salida de resultados [Data Views]

KNIME es utilizado desde el año 2006 utilizados en investigación farmacéutica, pero existen otras aéreas que lo han usado como: análisis de datos de cliente CRM, inteligencia de negocio y análisis de datos financieros.

Hasta ahora hemos señalado cuales son las principales características de cada herramienta que han sido consideradas para el estudio, en base a este conocimiento previo hemos detectado los siguientes parámetros, que se detallan en la tabla que se muestra a continuación:

Parámetro de comparación	**Significado**
Implementación	Especificación del lenguaje en que se implementó la herramienta.
Licencia	Especifica el tipo de licencia de software libre de la herramienta.
Interfaz	Determina si la herramienta tiene una interfaz amigable e intuitiva.
Formato de fuente de datos que soporta	Indica cuales son los formatos de la fuente de datos que son compatibles con la herramienta.
Herramienta de preprocesado de datos.	Señala si la herramienta tiene incorporado funciones para el preprocesamiento de datos y cuáles son.
Herramientas de visualización.	Indica que herramienta de visualización dispone el software para ilustrar el modelo de conocimiento obtenido de la aplicación de cualquier tipo de algoritmos.
Nivel de popularidad	Determina el nivel de usabilidad en proyectos.
Técnicas de clasificación.	Especifica cuales técnicas de clasificación dispone la herramienta.

Técnicas de asociación.	Detalla cuales son las técnicas de asociación dispone la herramienta.
Técnicas de agrupamiento.	Especifica las técnicas de agrupamiento incluidas en la herramienta.

CUADRO COMPARATIVO DE HERRAMIENTAS LIBRES PARA APRENDIZAJE AUTOMATICO.

Parámetro/Herramienta	Weka	Orange	RapidMiner	Knime
Implementación	Java	Python y C++.	Java.	Java
Licencia	GNU/GPL	GNU/GPL	La última versión solo viene en dos ediciones: Community edition: licencia GNU/GPL. Enterprise edition: licencia comercial.	GNU/GPLv3
Interfaz	Amigable. Poco intuitiva. Dispone distintas interfaces de usuario: • Knowledge Flow. • Explorer • Experimenter	Amigable e intuitiva similar a RapidMiner. Las interfaces de usuario que dispone: • Data • Visualize • Classiffy • Regression • Evaluate • Unsupervised • Associate	Amigable y más intuitiva que Weka, Orange y Knime. Las interfaces principales que incorpora son: • New ○ Repositories ○ Operators • Open recent • Open • Open template • Online tutorial	Interfaz amigable e intuitiva. Knime se organiza en diferentes paneles: • Workflow Editor: • Workflow Projects • Node repository. • Favorite Nodes • Node Description • OutLine • Console
Formato de fuente de	El formato de archivo	El formato de archivo	Los formatos soportados	La fuente de datos los

datos que soporta	compatible con esta herramienta es: *.arff, *.cvs, binarios y base de datos libre como MySQL.	soportado por Orange es de extensión: *.tab, *.xls (solo disponible en Windows),	de la fuente de datos son: *.arff, *.mdb, *.bibtex, *.dbase, *.xls entre otros.	formatos compatibles con esta herramienta son: *.xml, *.tst, *.trn, *. *.all, *.cvs, *.arff, *.xls.
Herramienta de preprocesado de datos.	Incluye operaciones como: Trabajo con Filtros: SI. Dispone de una variedad importante, ya sea a nivel de atributo y de instancias. Permite cargar y guardar el fichero de entrada, una vez filtrado los datos. Rellenado de valores faltantes: SI Generador de ruido: SI	Incluye operaciones como: Trabajo con Filtros: SI Discretizar:SI Selección de atributos: SI Missing values: SI	Incluye operaciones como: Filtros: SI. Incluye TFIDF, manejar serie de valores y otros. Rellenado de valores faltantes: SI Generador de ruido: SI	Incluye operaciones como: Filtros: SI Discretizar: SI Normalizar: SI Selección de variables: SI Missing values: SI
Herramienta de visualización de datos.	Incorpora herramientas de visualización para la representación en 2D de los datos (más específicamente de las	Orange ha sido diseñada para brindar una vasta variedad de formatos de visualización como:	Rapid Miner dispone de herramientas para la visualización de atributos o a nivel de tabla de datos. La variedad de	La representación de datos en Knime incluye desde gráficos de barras, circulares, gráficos de cajas e histogramas.

	relaciones existentes entre pares de atributos) y Filtrado "Gráfico" de los datos.	diagrama de barras, arboles, dendogramas, gráficos de redes y otros.	gráficos que se puede configurar es amplia, incorpora desde Histograma hasta gráficos circular.	
Nivel del popularidad	Muy alto.	Alto	Alto	Medio
Técnicas de clasificación	Esta herramienta dispone de una gran variedad de algoritmos de este tipo: Arboles de decisión: SI Vecino más cercano:SI Redes neuronales: SI Naive Bayes: SI Maquinas de vectores de soporte: NO. Análisis discriminante multivariante: NO	Las técnicas de clasificación disponible son: Arboles de decisión: SI Vecino más cercano: NO Redes neuronales:SI Naive Bayes: SI Máquinas de vectores de soporte: SI	Las técnicas que soporta se enumeran a continuación: Arboles de clasificación: SI Redes neuronales: SI Naive Bayes: SI Maquinas de vectores de soporte: SI. Análisis discriminante: SI Permite interactuar con algoritmos implementados en Weka.	Las técnicas disponibles son: Arboles de decisión: SI Vecinos más próximos: SI Redes neuronales: SI Naive Bayes: SI Maquina de vectores de soporte: SI Análisis discriminante multivariante: SI
Técnicas de asociación	Incluye las siguientes técnicas: Reglas de asociación: SI	Incluye las siguientes técnicas: Reglas de asociación:	Incluye las siguientes técnicas: Reglas de Asociación: SI	Incorpora las siguientes técnicas: Reglas de asociación: SI

	Regresión logística: NO Algoritmos genéticos: evolutivos: NO Redes bayesianas: NO	SI Regresión Logística: NO Algoritmos genéticos: NO Redes Bayesianas: NO	Regresión Logística: NO Algoritmos genéticos: NO Redes Bayesianas: NO	Regresión Logística: SI Algoritmos genéticos evolutivos: SI Redes bayesianas: NO
Técnicas de Agrupamiento	Incluye las siguientes técnicas: KMeans (Clustering Numéricos): SI Cobweb (Clustering conceptual): SI EM (Clustering probablistico): SI Algoritmos genéticos y evolutivos: NO Máquinas de vectores de soporte: NO Redes neuronales: NO	Incluye las siguientes técnicas: KMeans (Clustering numérico): SI EM (Clustering probablistico): NO Clustering jerárquico: SI Algoritmos genéticos evolutivos: NO Maquinas de vectores de soporte: NO Redes neuronales: NO	Incluye las siguientes técnicas: KMeans (Clustering numérico): SI Cobweb (Clustering conceptual): NO EM (Clustering probabilístico): NO Algoritmos genéticos evolutivos: NO Maquinas de vectores de soporte: NO Redes Neuronales: NO	Incluye las siguientes técnicas: KMeans (Clustering numérico): SI Twoset, Cobweb (Clustering conceptual): SI EM (Clustering probabilistico): NO Maquinas vectores de soporte: SI Redes neuronales: SI Algoritmos genéticos evolutivos: NO

CONCLUSIONES

Este trabajo ha destacado las bondades y principales características de herramientas software libre para el aprendizaje automático.

Una vez finalizado la confección el cuadro comparativo, podemos obtener algunas inferencias respectos a las herramientas consideradas en este estudio:

- Rapid Miner, Orange, KNIME en ese orden disponen de una interfaz amigable y muy intuitiva.
- Weka tiene una interfaz amigable pero poco intuitiva.
- En el caso de necesitar realizar un proyecto de aprendizaje automático, en el que es necesario el uso de técnicas de clasificación Weka es la mejor alternativa ya que dispone una diversidad significativa de algoritmos implementados en forma nativa, respecto a las demás herramientas estudiadas.
- Para personas principiantes en el manejo de software para aprendizaje automático es muy importante y recomendable que la interfaz de la herramienta sea lo más amigable e intuitiva. La alternativa recomendable para esta situación es RapidMiner. Además la documentación que ofrece la página oficial de esta herramienta es muy clara y fácil de comprender.
- Si se necesita llevar a cabo un proyecto de aprendizaje automático, en la que el aspecto de visualización ya sea de los datos como del él/los modelo/s obtenido/s es un parámetro significativo para el proyecto, Orange es la alternativa más adecuada por la gran diversidad de herramientas de visualización que dispone.
- Si tenemos en cuenta la flexibilidad de formato de la fuente de datos que soporta la herramienta, las mejores alternativas son RapidMiner, Knime.
- De acuerdo a las investigaciones realizadas de proyectos de aprendizaje automático implementados con herramientas libres, Weka encabeza a nivel de popularidad.

REFERENCIAS

[1] “Aprendizaje Automático: conceptos básicos y avanzados. Aspectos prácticos utilizando el software Weka”, Basilio Sierra Araujo, Año: 2006.

[2] “Programando con inteligencia artificial”, Beata Lackoga, Revista Linux.net, 2011 http://revistalinux.net/articulos/programando-con-inteligencia-artificial-2/

[3] “5 Programas libre para la Minería de Datos”, Fuente: TechSource, http://fraterneo.blogspot.com/2010/11/5-programas-libres-para-data-mining.html

[4] “Análisis de Datos en Weka- Pruebas de Selectividad”, María García Jiménez, Aránzazu Álvarez Sierra, Universidad Carlos III; http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf

[5] “Weka: Waikato Environment for Knowledge Analysis: Introducción básica, Explorer”, Carlos J. Gonzalez. (Departamento de Informática), Universidad Valladolid, http://www.infor.uva.es/~calonso/IAII/Aprendizaje/Practica1/IntroduccionWeka.pdf

[6] Sitio oficial KIME, http://www.knime.org

[7] “Sistemas de Inteligencia de Gestión. Práctica 1. Herramienta de Datos: KNIME”, Juan Carlos Cubero, Fernando Berzal, http://elvex.ugr.es/decsai/intelligent/workbook/D1%20KNIME.pdf

[8] Sito oficial RapidMiner, http://rapid-i.com

[9] “Técnicas de análisis de datos. Aplicaciones prácticas usando Microsoft Excel y Weka”, José Manuel Molina López, José García Herrero, 2006.

[10] Sitio oficial de Orange, http://orange.biolab.si/