

SISTEMAS Y HERRAMIENTAS DE MINERÍA DE DATOS. EJEMPLOS:

Lo que se pretende conseguir es proporcionar información detallada sobre algunos ejemplos de software diseñados para implementar la Minería de Datos.

☞ Librerías: las librerías de Minería de datos son un conjunto de métodos que implementan funcionalidades y utilidades básicas como el acceso a datos, modelos de redes neuronales, métodos bayesianos, exportación de resultados... Las librerías se encargan principalmente de facilitar el desarrollo de las tareas de Minería de Datos que son más complejas, como el diseño de experimentos. El problema de las librerías, es que es precisa la comprensión de conocimientos de programación.

Algunas de las Librerías más importantes son:

- **XELOPES** (extended library for prudys embedded solution): es una librería con licencia pública GNU para el desarrollo de aplicaciones de Minería de Datos. Esta librería está implementada para que sea eficiente para la mayoría de los algoritmos de aprendizaje, por eso, es importante destacar que el usuario puede desarrollar aplicaciones particulares de Minería de Datos. Sus principales características son:
 - Acceso a datos
 - Modelos de redes neuronales
 - Métodos de agrupamiento
 - Métodos de reglas de asociación
 - Árboles lineales
 - Árboles no lineales
 - Exportación de datos

Para más información sobre XELOPES:

<http://www.prudsys.com/Produkte/Algorithmen/Xelopes>

- **MLC++** (Machine learning library in C++): Es un conjunto de librerías que fueron desarrolladas por la Universidad de Standford. La mayoría de las versiones son bajo dominio de investigación, a excepción de la versión 1.3.x, que se distribuye bajo licencia de dominio público.

Las principales características son:

- Acceso a datos (archivos con formato plano)
- Transformaciones de datos

- Métodos de aprendizaje mediante objetos

Para más información sobre MLC++ : <http://www.sgi.com/tech/mic> o <http://www.purpleingit.com>

- **SUITES**: posee las mismas capacidades que el procesamiento de datos, los modelos de análisis, el diseño de experimentos o el soporte gráfico para la visualización de resultados. En este caso, Suites destaca porque existe una interfaz que facilita la interacción entre el usuario y la herramienta

- **SPSS CLEMENTINE**: es uno de los sistemas de Minería de Datos más conocidos. Posee una herramienta visual desarrollada por ISL que tiene una arquitectura cliente / servidor.

Este sistema se caracteriza por:

- Acceso a datos (fuentes de datos, archivos ASCII...)
- Procesamiento de Datos
- Técnicas de Aprendizaje (redes neuronales, reglas de asociación)
- Técnicas de evaluación de modelos
- Visualización de resultados (histogramas, diagramas de dispersión, gráficos en 3-D...)
- Exportación (informes en HTML o texto)

Para más información sobre SPSS CLEMENTINE:

<http://www.spss.com/spssbi/clementine>

- **WEKA** (Waikato environment for knowledge analysis): es una herramienta visual de libre distribución desarrollada por los investigadores de la Universidad de Waikato en Nueva Zelanda.

Sus principales características son:

- Acceso a los datos desde un archivo en formato ARFF (es un archivo de texto plano organizado en filas y columnas)
- Preprocesado de datos (selección, transformación de atributos...)
- Modelos de Aprendizaje (reglas de asociación, modelos de agrupamiento, modelos combinados...)
- Visualización del entorno

Para más información sobre WEKA:

<http://www.cs.waikato.ac.nz/~ml/WEKA>

- **KEPLER:** Sistema desarrollador y transformado en una herramienta comercial distribuida por Dialogis. Posee múltiples modelos de análisis. Sus principales herramientas de aprendizaje son:

- Árboles de decisión
- Redes neuronales
- Regresión no lineal
- Aplicaciones estadísticas

Así mismo permite el preprocesado de datos, la elección de un modelo o la manipulación de la representación gráfica de los modelos obtenidos.

Para más información sobre KEPLER: <http://www.dialogis.de>

- **ODMS** (oracle data mining suite): Está diseñado sobre una arquitectura cliente servidor; ofrece una gran versatilidad en cuanto al acceso a grandes volúmenes de información. Se caracteriza principalmente por:
 - Acceso a datos en diversos formatos: almacenes de datos, bases de datos relacionales como SQL, Oracle..., archivos planos...
 - Preprocesado de datos: muestreo de datos, patrones de datos...
 - Modelos de aprendizaje: redes neuronales, regresión lineal...
 - Herramientas de visualización: resultados estadísticos, importación de datos en Excel, Word o Power Point...

Para más información sobre ODMS:

<http://otn.oracle.com/products/datamining>

- **DBMINER:** Sistema interactivo desarrollado por la Universidad de Simon Fraser de Canadá. Su licencia es pública a excepción de la empresarial (que es comercializada) Está concebido para la extracción del conocimiento de bases de datos relacionales, almacenes de datos y Web. Dentro de su arquitectura de diseño es importante destacar:

- OLAP (online analytic processing)
- OLAM (online analytic mining)

La herramienta de DBMiner posee dos modos de trabajo:

- Vía interfaz gráfica
- Vía interfaz de script

Para más información sobre DBMINER: <http://www.dbminer.com>

- **YALE:** herramienta de aprendizaje automático implementado en Java por la Universidad de Dortmund. El sistema incluye operaciones para:
 - Importación y pre-procesamiento de datos
 - Aprendizaje automático
 - Validación de modelos

Para más información sobre YALE: <http://yale.cs.uni-dormund.de>

- **DB2 INTELLIGENT MINER:** herramienta comercial distribuida según la arquitectura cliente / servidor y distribuida por IBM. Posee una serie de paquetes de implementación para Minería de Datos entre los que destacan:
 - Db 2 intelligent miner for data: destinado para aplicar tareas de minería en bases de datos. Soporta tareas de agrupamiento, asociaciones, patrones, clasificación...
 - Db 2 intelligent miner scoring: utiliza la funcionalidad de la base de datos para aplicar las técnicas de Minería de Datos
 - Db 2 intelligent miner modelling: orientado al descubrimiento de las relaciones entre los datos, como las asociaciones o el agrupamiento de dichos datos
 - Db 2 intelligent miner visualization: para visualizar los resultados procedentes de los modelos de análisis

Para más información sobre DB2 INTELLIGENT MINER: <http://www-3.ibm.com/software/data/iminer>

- **SAS ENTERPRISE MINER:** es uno de los sitios comerciales que proporciona el SAS Institute para tareas de Minería de Datos. Ésta herramienta se centra en la Minería de Datos (de forma tradicional) y no en su funcionalidad (es el caso de SAS Text Miner)
Posee una arquitectura distribuida, es decir, tiene una potente interfaz gráfica de usuario. Las tareas que realiza esta herramienta son:
 - Preprocesado de datos: tratamiento estadístico, filtros, tareas de muestreo...
 - Modelos: árboles de decisión, regresión lineal, redes neuronales, construcción de métodos de ensamblaje...
 - Evaluación: para comprobar la eficacia y eficiencia de dicha herramienta

- Visualización de resultados: a través de gráficos, diagramas, informes en formato html...

Para más información sobre SAS ENTERPRISE MINER:

<http://www.sas.com>

- **STATISTICA DATA MINER:** es una potente herramienta con un sistema visual desarrollado y comercializado, en la que destaca:
 - Base de datos: permite trabajar con un gran volumen de información, así como importar los datos en formatos Excel, Oracle, SQL...
 - Preprocesado de datos: a través del cual seleccionamos las características, el muestreo de datos, realizamos operaciones de filtrado, tratamiento de datos...
 - Modelos de análisis: como reglas de asociación, redes neuronales, modelos lineales de regresión, modelos no lineales de regresión, regresión múltiple...
 - Visualización: desarrollada a través de una interfaz gráfica que facilita las diversas tareas que pueda realizar el usuario. Destacan los diagramas de barras, diagramas de sectores, árboles de asociación, redes neuronales...

Para más información sobre STATISTICA DATA MINER:

<http://www.statsoft.com/dataminer.html>

☞ Herramientas Específicas: caracterizadas por centrarse en un único modelo (redes neuronales, modelos estadísticos...) o en una determinada tarea de Minería de Datos (clasificación o agrupamiento). Sin embargo, destacan porque permiten realizar todo el proceso de Minería de Datos y porque no es necesario poseer grandes conocimientos de programación.

Algunas de estas herramientas específicas son:

- **CART:** herramienta desarrollada y comercializada por Salford System y orientada a tareas de clasificación o regresión de Minería de Datos. Destaca principalmente por su accesibilidad, capacidad de visualización o información estadística relativa al modelo

Para más información sobre CART: <http://www.salford-systems.com/index.html>

- **AUTO CLASS:** son tres herramientas (Auto Class III, Auto Class X y Auto Class C) asociadas bajo el mismo nombre y desarrolladas por la NASA. Todas ellas tratan de solucionar los problemas de agrupamiento de la Minería de Datos utilizando métodos bayesianos. Sus características son:

- Formato propio de archivo de entradas
- Generación de informe
- Extracción de test

Para más información sobre AUTO CLASS:

<http://ic.arc.nasa.gov/ic/project/bayes-group/autoclass/>

- **NEURAL PLANNER, NEURODIET, EASY NN-PLUS:** son tres herramientas desarrolladas para trabajar con modelos de aprendizaje basados en redes neuronales

Para más información sobre NEURAL PLANNER, NEURODIET, EASY NN-PLUS:

<http://www.tropheus.demon.co.uk>

- **NEUROHELL:** son un conjunto de herramientas gráficas para trabajar con modelos de aprendizaje basados en redes neuronales

Para más información sobre NEUROHELL:

<http://www.wardsystem.com/products.asp>

- **SEE 5 / C5.0:** herramienta de fácil manejo, centrada en la construcción de modelos de clasificación basados en árboles de decisión y conjuntos de reglas. El inconveniente es que ha sido diseñada para trabajar con grandes volúmenes de datos

Para más información sobre SEE 5 / C5.0:

<http://www.rulequest.com/download.html>