

Práctica 3 de Minería de Datos

Ejercicios Libres con Clementine



Curso de
Almacenes de Datos y Minería de Datos

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia

en colaboración con

Universidad de El Salvador

José Hernández Orallo. (jorallo@dsic.upv.es). Julio 2003

Índice

1.	Consejos para trabajar con grandes volúmenes de datos.....	3
2.	Modo Batch.....	3
3.	Ejercicios Libres.....	4
3.1	Tamaño Pequeño / Medio.....	4
3.2	Gran Tamaño.....	5

En esta tercera parte, se plantea un conjunto de problemas sobre los cuales podrás elegir. Se deja completa libertad para extraer modelos, es decir para hacer una minería de datos autónoma.

El tamaño y complejidad de los datos en esta tercera parte es mucho mayor, de hecho algunos de los problemas son concursos de conferencias del campo. Por tanto ¡competiremos a un alto nivel!

1. Consejos para trabajar con grandes volúmenes de datos

En primer lugar, para trabajar con datos de gran tamaño es posible que necesites cambiar la memoria disponible del Clementine. Esto se realiza en el menú Options → Memory Limit.

También puede ser recomendable en algún caso abrir más de una ventana del Clementine, para ir trabajando en un stream mientras en otro se está calculando un modelo.

No obstante, para grandes volúmenes los siguientes consejos son fundamentales:

- Realiza muestras de los datos para trabajar con pequeños volúmenes. Una vez ya hayas clarificado atributos relevantes y modelos a extraer ya puedes trabajar con el volumen inicial.
- Agrega datos. Si tienes datos muy detallados (por día o por hora), intenta agregar los datos en semanas o meses. Lo mismo puedes hacer con productos y familias. Reducirás el tamaño de una manera drástica y tendrás mayor manejo de los datos. Si las agregaciones son complejas realízalas en un Sistema de Gestión de Bases de Datos o en un Almacén de Datos y no en el Clementine.
- Elimina atributos redundantes o irrelevantes.
- Reserva la mitad de datos para validación.
- Intenta analizar los datos y comprender sus características básicas antes de intentar extraer modelos.
- Prueba los tipos de modelos más rápidos (C5.0, regresión o Kmeans) antes que otros modelos más lentos.

Finalmente, si todo lo anterior no te ayuda a reducir el tiempo de extracción de los modelos puedes optar por la opción del modo Batch, que se describe en la siguiente sección:

2. Modo Batch

La ejecución de algunos modelos puede ser bastante larga para trabajar interactivamente, o simplemente se desea que se realice a unas determinadas horas (por la noche, p.ej.). Para este tipo de situaciones, la ejecución de Clementine en modo Batch puede ser útil.

Vamos a ver un ejemplo. Para simplificar el desarrollo vamos a suponer que existe un directorio "D:\BATCH" donde se va a trabajar. Tú puedes crearte un directorio similar en esa u otra unidad o en el temporal.

Copia en ese directorio el fichero "customer.txt" que tienes en "..\LabKDD\customer.txt". Crea un nuevo stream en el directorio "D:\BATCH" como se muestra en la siguiente figura (puedes elegir otro stream, siempre que no tenga nodos gráficos o tablas y la salida vaya a fichero):

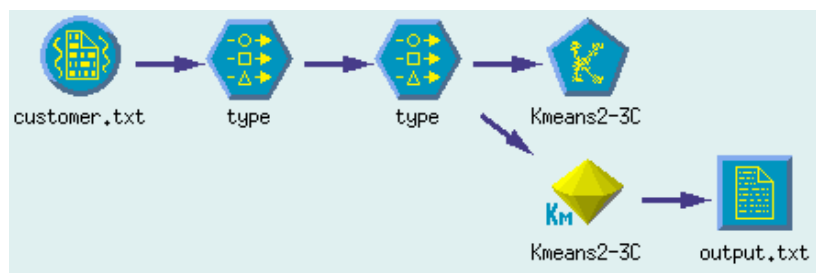


Figura 2.1. Stream para realizar en modo batch

Grábalo con el nombre “customer-batch.str”. Ahora, en el mismo directorio “D:\BATCH” crea un fichero “batch.bat” y edítalo para que contenga las siguientes 4 instrucciones:

```
c:
cd "c:\archivos de programa\clementine\5.2.1\bin"
call clementine.exe -batch -stream "d:\batch\customer-batch.str" -execute
-log d:\batch\resultados.log
pause
```

Nótese que la tercera entrada es una sola línea. Si ahora ejecutas este fichero “bat” deberá generar dos ficheros: “output.txt” y “resultados.log”. El primero contendrá la salida del nodo “output.txt” conectado al modelo Kmeans2-3C del stream. El segundo fichero “resultados.log” contiene los mensajes de error que se pudieran haber producido durante la ejecución del stream.

La ejecución de streams en modo “batch” dará error si existe algún nodo de visualización en el stream (tablas, gráficas, etc.). Tampoco se puede utilizar directamente para generar modelos, ya que no hay manera directa de recuperarlos de la zona de modelos una vez generado. Para realizar acciones en modo batch de este estilo o más sofisticadas, hay que hacer uso de los “scripts”, aspecto que no vamos a tratar aquí y para el cual se puede consultar el manual de Clementine.

3. Ejercicios Libres

Los ejercicios libres son los siguientes y se encuentran en el directorio “_LabKDD2” (no hace falta bajárselos de la página web). Algunos son tan libres que simplemente se trata de sacar cuantos más patrones útiles mejor (no hay tareas predeterminadas). Cuando elijas uno, entonces sí que es conveniente que lo copies a tu máquina, p.ej. en el directorio temporal.

Aparte del fichero “.data” que suele contener los datos, normalmente hay un fichero “.names” que contiene una breve descripción del problema (en inglés). A veces el problema también está separado en datos “.train” y “.test”.

En la carpeta “_LabKDD2” tienes dos subdirectorios:

- “BIG PROBLEMS”
- “SMALL AND MEDIUM PROBLEMS”

Pasemos a ver qué problemas hay en cada uno de estos directorios:

3.1 Tamaño Pequeño / Medio

Como hemos dicho, en el directorio “_LabKDD2\Small and medium problems”, se encuentran los siguientes problemas:

- UCI (varios problemas, 82Mb). Directorio “UCI repository”:
Es un conjunto con un centenar de problemas de diferente tipo, tamaño y complejidad. Casi todos son de carácter predictivo, aunque se pueden usar para otros fines. Puedes elegir cualquiera de ellos.
Se han descargado de la siguiente dirección.
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
<http://www.ics.uci.edu/pub/machine-learning-databases/>
- ML-DATASETS (varios problemas sintéticos, 3Mb). Directorio “ML-Synthetic-datasets”:
Es un conjunto de problemas sintéticos de clasificación. Se obtuvieron de <http://www.recurisive-partitioning.com/mv.html>, dirección que ahora ya no se encuentra disponible.
- IVE (datos de la Comunidad Valenciana). Directorio “IVE”:
Se han descargado unas pocas tablas de datos demográficos de la Comunidad de la web <http://ive.infocentre.gva.es/inicio-cifras.htm> aunque puedes descargarte muchas más si te interesan estos datos.

- Neptuno (base de datos relacional, 4Mbs). Directorio “Neptuno”:
El ejemplo de la empresa que viene que viene con el Ms Office. Permite trabajar con datos relacionales y probar conexiones ODBC.

3.2 Gran Tamaño

Como hemos dicho, en el directorio “_LabKDD2\Big Problems” se encuentran problemas de mayor dificultad. Aquí los problemas son todavía menos dirigidos que los anteriores, aunque en algún caso hay unas metas claras.

- DELVE (varios problemas, 93Mbs). Directorio “delve”:
Contiene varios problemas, algunos de gran tamaño y otros más discretos. Se ha descargado completamente de: <http://www.cs.utoronto.ca/~delve/index.html>
- UCI KDD Archive. (varios problemas, 14,5Mbs). Directorio “UCI KDD Archive (VERY BIG PROBLEMS)”
Extensión y continuación del archivo UCI ML visto en el apartado anterior pero incluye problemas más gordos. Existen problemas para clasificación, clustering y regresión.
Se han descargado de la siguiente dirección:
<http://kdd.ics.uci.edu/>
- SWISSLIFE DATASET (Ámbito: financiero, 4Mbs). Directorio “swiss-life”:
<http://research.swisslife.ch/kdd-sisyphus/>
- PKDD’99 Discovery Challenge (Ámbito financiero y médico, 19Mbs). Directorio “pkdd’99”:
<http://lisp.vse.cz/pkdd99/Challenge/chall.htm>
- KDD Cup 2000 (Ámbito comercial: 1148MB). Directorio “KDD-CUP2000”
Debido a su gran volumen, para este problema sólo se ha descargado la descripción. Si tienes ganas y te atreves, la página web es: <http://www.ecn.purdue.edu/KDDCUP/> con el siguiente username y password para obtener los datos:
UserName: kddcup
Password: legcare4KDD
- The COIL Challenge 2000 (Ámbito de seguros, 7,5Mbs). Directorio “cc2000”
An Insurance Company.
<http://www.liacs.nl/~putten/library/cc2000/>
- PKDD 2001 Challenges (Ámbito médico, 4Mbs). Directorio “Pkdd2001”
http://www.informatik.uni-freiburg.de/~ml/ecmlpkdd/discovery_challenges.html
The Predictive Toxicology Challenge for 2000-2001 (MUY COMPLEJO)
The Discovery Challenge on Thrombosis Data (Thrombosis)
- KDD 2001 (Ámbito médico, 3Mbs). Directorio “kddcup2001”
<http://www.cs.wisc.edu/~dpage/kddcup2001/>
Dataset 1 (Thrombin) : Prediction of Molecular Bioactivity for Drug Design -- Binding to Thrombin
Dataset 2 (Genes): Prediction of Gene/Protein Function and Localization