

PRÁCTICA 2

Visualización e Interacción en bases de datos n-dimensionales

1. XMDV TOOL

Xmdv Tool es una herramienta gratuita para la visualización interactiva de bases de datos de multivariadas. Podemos descargarlo de <http://davis.wpi.edu/xmdv/> para diferentes plataformas.

Vamos a trabajar con esta herramienta para conocer en mejor detalle algunas de las técnicas de Visualización vistas en el Tema 2, y algunas técnicas de interacción vistas en el Tema 5.

1.1) Base de datos

a) Lo primero es descargar e instalar el programa.

b) Por defecto nos proporcionan la base de datos *Cars*. Descargamos *Detroit* desde la misma dirección que la herramienta, y abridla. Elimina la vista de *Cars*.

Detroit es una base de datos sobre homicidios anuales en Detroit en los años 1961 a 1973. Así, por cada año tenemos un registro, descrito por 7 variables:

Ft_police: policías a tiempo completo por 100.000 habitantes

Unemp: % de población en paro

Manu_wrkrs: miles de trabajadores

Handgun_lcs: licencias de pistolas por 100.000 habitantes

Gov_wrkrs: miles de funcionarios

Cleared: % homicidios que han finalizado con un arresto

Homicides: número de homicidios por 100.000 habitantes.

Abre el archivo con un procesador de textos para ver el formato. Xmdv Tool también permite importar archivos .csv.

1.2) Paralelas Coordenadas

a) Por defecto el color corresponde al número (orden) de cada caso en la base de datos. Lo primero es elegir el esquema de colorido en Tools→Color Strategy. Ahí indicad que el color dependa del valor de una variable. Probad con varias variables.

Para la variable *Homicides* fíjate en:

-¿Qué color corresponde a menos homicidios?

-Para homicidios bajos, ¿hay alguna variable que suela tener siempre el mismo valor?

b) Recuerda que en el Tema 2 dijimos que en los sistemas de visualización de muchas dimensiones, el orden de las variables es importante. Y, más aún, para las Paralelas Coordenadas.

-¿Crees que hay algún orden de los ejes verticales que nos permitan ver mejor las relaciones entre variables? Pruébalo con Tools→Manu Dimension On/Off Reorder. Además de reordenar, puedes ocultarlas.

-Podemos ordenarlos automáticamente con Tools→Automated Dimension Reorder, utilizando los métodos basados en similaridad (relación entre variables). ¿Algún método coincide con tu intuición?

-Los métodos basados en sobrelapamientos de casos (clutter-based) evitan o aumentan el sobrelapamiento o unión de casos. ¿Es útil maximizar este sobrelapamiento?

1.3) Matriz de dispersión

a) Haz click en Add View, y selecciona ScatterPlot Matrix.

Aquí puedes observar la relación entre variables 2 a 2. Puedes cambiar el Zoom utilizando los iconos, y cambiar la forma y tamaño de los puntos utilizando Tools→ScatterPlot Customization.

b) La diagonal principal no muestra ninguna información adicional por defecto. Personaliza la matriz para que muestre:

- Histograma de valores de cada variable
- Distribución de la variable dada otra.
- Si eliges 2 dimensiones y Dirigido por Datos, dibuja en todas las casillas la misma gráfica para las 2 variables seleccionadas.

c) El rectángulo rosa se utiliza para realizar Brushing y Brushing & Linking (ver Tema 5:técnicas de interacción).

-Selecciona los casos con mayor valor para la variable homicidios. Para seleccionar tienes que cubrir todas las dimensiones del caso con el rectángulo rosa, o bien pulsar la tecla SHIFT+botón izquierdo del ratón mientras pasas el puntero por los casos de interés. Sabrás si están seleccionados porque se colorean en rojo.

-Ahora vuelve a la vista de Coordenadas Paralelas, y observa los valores que toman los casos seleccionados para cada variable (también pueden verse en la otra vista).

-Practica un rato con el brushing.

1.4) Pictogramas de Estrella

Recuerda del tema 2 que cada estrella es un registro (en este caso, un año). Cada vértice de la estrella es una variable, y la longitud del tamaño del segmento que nace del centro al vértice es proporcional valor de la variable correspondiente.

a) Ordena las estrellas por año (elige en orden el criterio 'Original') en Tools → Glyph Customization →sorting mode: original

b) Fíjate que al pasar el ratón por una estrella, en la parte más baja de la aplicación te muestran los valores de cada variable.

c) Haz que el color de las estrellas dependa del número de homicidios. ¿En qué años hay menos y más homicidios?

d) Reordena las variables por correlación de Pearson.

-¿Podríamos decir que cuanto más forma de cabeza de pájaro tiene la estrella (con pico a la derecha, y cuello recto), más homicidios hay?

-¿Y que conforme pasan los años, la forma de pájaro se va acentuando?

1.5) Apilado Dimensional

a) El Apilado Dimensional requiere que haya pocas variables para que sea interpretable, y también que las variables de mayor interés sea más externas. Así que elimina todas las variables menos *homicidios*, *cleared*, *handgun_lcs* y *ft_police*. Y colócalas en este mismo orden. Fíjate que quedan menos puntos coloreados que casos, eso es porque se superponen.

b) El color rosa para realizar *Brushing* puede despistarnos en este tipo de gráfico. Elimínalo dándole color blanco en Tools → Color Customization

c) Las variables son discretizadas automáticamente. Podemos elegir el número de posibles valores de cada variable en Tools → Dimension Stacking Customization. Cambia homicidios a 3 valores.

d) Como estrategia de color, elige la que quieras, por ejemplo número de homicidios, que será redundante pero te servirá para recordar en qué eje se encuentra esta variable.

e) Responde a estas preguntas:

-A más homicidios, ¿hay más o menos arrestos? (variable *cleared*, eje vertical externo).

Comprueba tu conclusión comparando esas 2 variables en la Matriz de Dispersión.

-Añade o reordena variables un rato.

1.5) Gráfico orientado a píxeles

Esta es una técnica recomendada para bases de datos con pocas variables y muchos registros. Así que descárgate de la página de Xmdv Tool la base de datos Cars (o cierra y abre el programa, ya que es la base de datos por defecto con la que se inicia).

Cars tiene 392 registros (1 por coche) descritos cada uno por 7 variables:

MPG: millas que puede recorrer por galón de gasolina consumido.

Cylinders: cantidad de cilindros

Horsepower: caballos

Weight: peso

Acceleration: aceleración

Year: año

Origin: país de origen

a) Al seleccionar que se genere el gráfico de píxeles, , cada píxel es un registro. Escala los píxeles en Tools-> Pixel-based Customization, para ver bien las gráfica.

b) Por defecto, en la primera gráfica se ordenan los casos por valor de MPG y en espiral. En el resto de gráficos, los casos ocupan la misma posición que en la gráfica de MPG, y el color corresponde al valor de la variable que ese gráfico representa.

Responde a estas preguntas:

- A valores pequeños de MPG, ¿Qué valores suele tomar el número de cilindros?
- ¿Se puede sacar alguna conclusión comparando MPG y el año del coche?

c) Haz ahora que la variable de referencia sea *Year*, y que los píxeles se dibujen en Horizontal.

-¿Puede decirse que a fechas más avanzadas, aumentan las millas recorridas por galón?

-¿Te da la impresión que la cilindrada y los caballos eran mayores en fechas más antiguas? ¿Será efecto del gráfico o defecto en la recogida de datos?

1.6) Técnicas jerárquicas

Xmdv Tool permite unir datos similares en clusters, y clusters similares en otros clusters mayores, formando así un árbol de clusters. Éste árbol puede visualizarse utilizando las mismas técnicas ya vistas, pero en este caso el valor mostrado por aun cluster es la media de sus valores, y la banda que le rodea es la varianza.

En *Coordenadas Paralelas*, una línea es la media en el cluster para la variable correspondiente en cada eje. Las bandas delimitan los valores del cluster.

En *Matriz de Dispersión*, un punto es la media de un cluster, y su banda también delimita el cluster. Es una forma muy visual de contemplar los clusters creados.

En *Pictogramas de Estrella*, cada estrella es la media de un cluster para cada variable, y también se delimita el cluster con las bandas.

En *Apilado Dimensional*, un punto es un cluster, y la banda se dibuja en el mismo punto, o en los puntos que lo rodean si sus valores caen en un valor discreto distinto.

a) ¿Te resulta más sencillo encontrar relaciones?

b) Puedes cambiar el nivel de clustering en Brushes→Structured-based Brush for Hier Displays
→Non-brushed cluster Radius