DETECCIÓN DE DATOS CON RUIDO EN BASE DE DATOS UTILIZANDO HERRAMIENTAS OLAP

Pautsch Germán A, Kuna Horacio D., Caballero Sergio D., Rambo Alice R., Meinl Evaldo, Steinhilber Andrés

Departamento de Informática, Facultad de Ciencias Exactas Químicas y Naturales (F.C.E.Q. y N) , Universidad Nacional de Misiones (U.Na.M), Félix de Azara 1552 - C.P. N3300LQH - Posadas - Misiones - Argentina.
E-mail : gpautsch@hotmail.com

Experimentación

1. Selección de una fuente de datos.

Para esto hacemos click secundario en Root / new operador / io / examples (Figura 1). Aqui definimos la ruta donde se encuentra la funte de datos.

Figura 1. Selección de una fuente de datos.

2. Generación de una matriz de dispersión (Scatter Matrix)

El segundo paso es verificar si un atributo de la base presenta ruido. Debido a que esto a priori se desconoce, lo que hacemos es generar una matriz dispersión que grafica un atributo con respecto a todos los demás. Esto se genera en forma automática para cada atributo en un mismo gráfico (Figura 2).

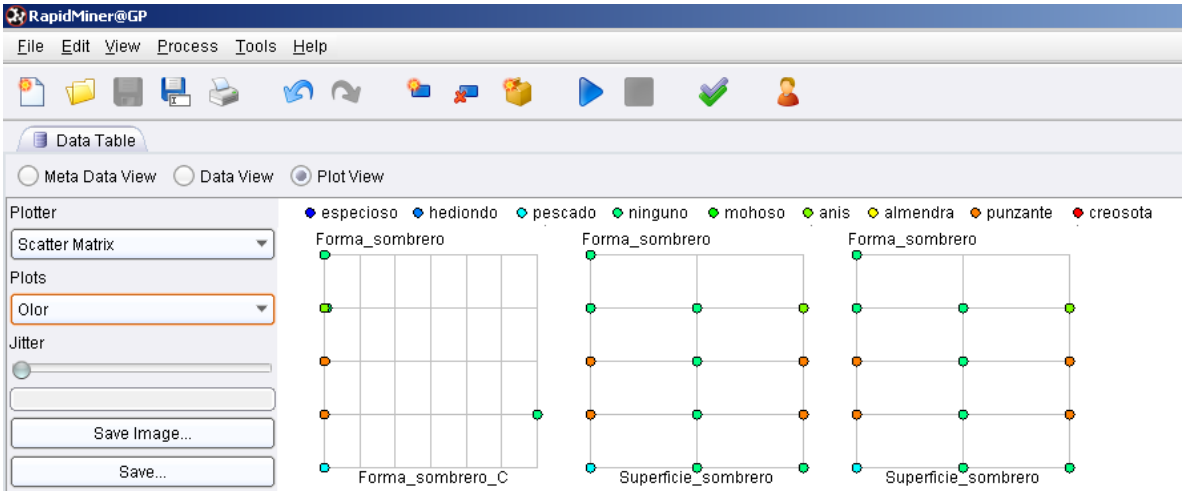

Figura 2. Matriz de Dispersión.

3. *Aplicación de un desplazamiento aleatorio sobre los valores de x e y (Jitter), para lograr una clara visualización de como se agrupan los datos*

Realizar un Jitter refiere a la aplicación de un desplazamiento aleatorio sobre los valores de x e y (Jitter), para lograr una clara visualización de como se agrupan los datos.

Esta opción esta en el marco superior derecho izquierdo de la pantalla. Figura 3

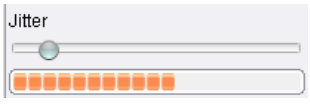

Figura 3. Jitter

4. *Identificación de los atributos cuya dispersión represente la probabilidad de ruido*

Luego de la aplicación del Jitter, y como se puede ver en la Figura 4, evidentemente en cada gráfico donde esta presente el atributo "Forma_sombrero_C", existe una dispersión muy acentuada (probable ruido).

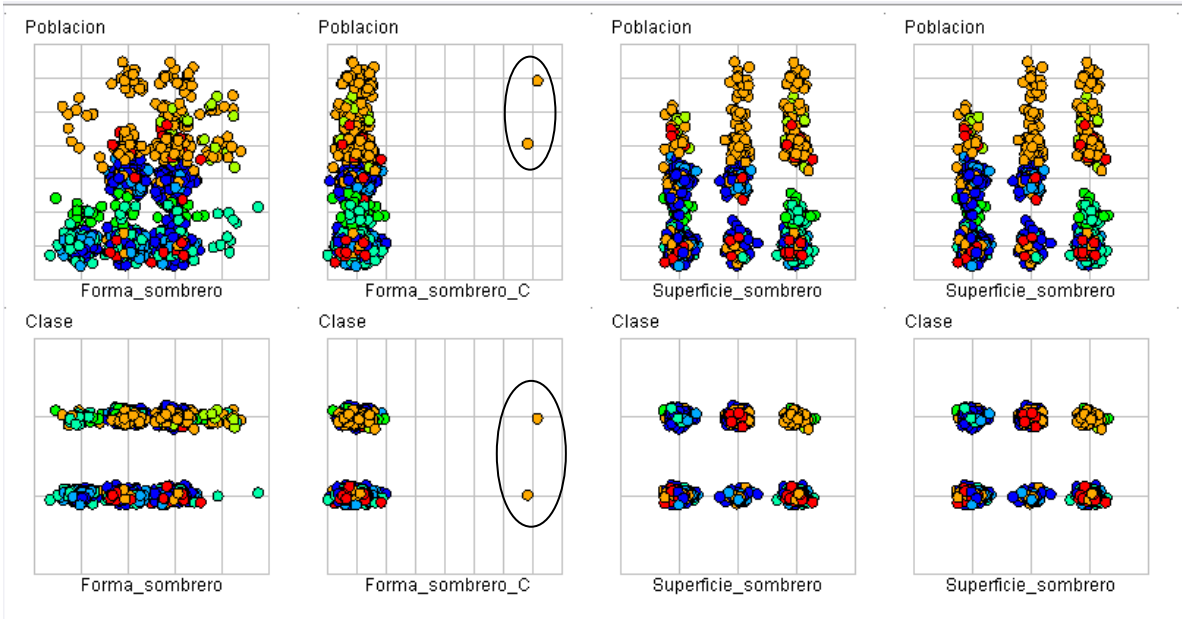

Figura 4. Matriz de Dispersión.

5. Generación de un gráfico de dispersión (Scatter) sobre cada atributo identificado en el paso previo.

Una vez que el atributo con ruido es identificado, realizamos un gráfico de dispersión del mismo. Esto nos va a permite identificar el rango de valores en el que el ruido se hace presente. (Figura 5)

Figura5. Gráfico de Dispersión

6. Jitter.

Debido a que el gráfico anterior no nos aporta mucha información a cerca de el ruido presente en el atributo "Forma_sombrero_C", identificado en el punto 4, vamos a introducir manualmente una pequeña cantidad de ruido repitiendo el paso mencionado en el punto 3 de esta experimentación.

7. Identificación del rango (del valor del atributo) en el que el ruido se hace presente.

En la figura 6 un experto en el estudio de hongos puede apreciar a simple vista que el atributo "Forma_sombrero_C" presenta ruido alrededor del valor 3000.

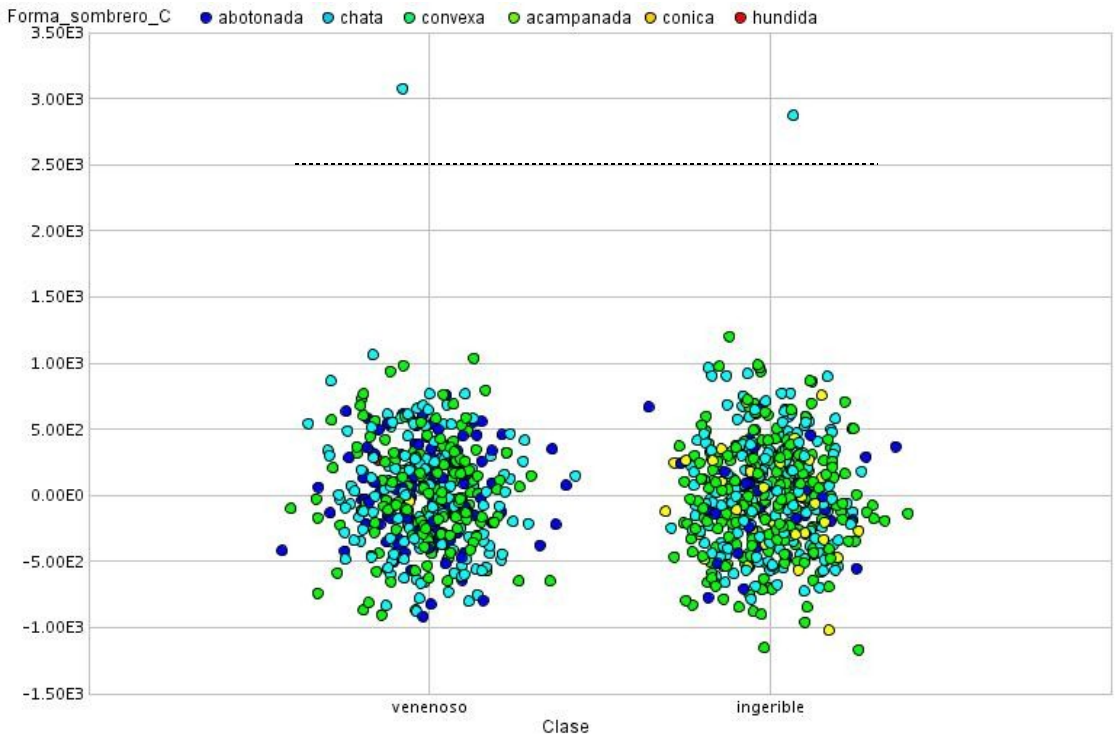

Figura 6. Gráfico de dispersión

Con este ultimo paso ya hemos identificado claramente el ruido presente en el atributo "Forma_sombrero_C".

Opcionalmete con el rango del valor de cada atributo se puede realiza una consulta a la base de datos para aislar el ruido. Esto se describe en el siguiente punto.

8. Aislar el Ruido Identificado

Una vez identificado el rango en el que el ruido se hace presente, se lo puede aislar de la base de datos utilizando Lenguaje de Manipulación de Datos (DML) de SQL (Structured Query Lenguage - Lenguaje de Consulta Estructurado) (Figura 7)

row no.	Forma_sombrero	Forma_sombrero_C	Superficie_...	Superficie_...	Magulladur...	Olor	Tipo_mem...	Espaciado_...	Tamaño_m...	Color_mem...
5132	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	chocolate
5133	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	negra
5134	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	rosa
5135	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	chocolate
5136	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	negra
5137	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	chocolate
5138	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	marron
5139	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	chocolate
5140	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	chocolate
5141	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	chocolate
5142	chata	3000	suave	suave	no	ninguno	libre	poblado	ancha	chocolate
5143	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	rosa
5144	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	negra
5145	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	marron
5146	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	negra
5147	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	negra
5148	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	marron
5149	chata	3	suave	suave	no	ninguno	libre	poblado	ancha	rosa

Figura 7.