Diseño de un experimento con Pig

Miguel Ángel Roldán Carmona

1. Estructura y contenido de la base de datos

Para el diseño del experimento se ha escogido el conjunto de datos Wine Quality disponible aquí.

Este conjunto de datos contiene muestras de vino blanco del vino portugués "Vinho verde". El conjunto tiene 4899 instancias, 12 atributos y no presenta valores perdidos .

Es un conjunto muy versátil debido a que se puede utilizar para tareas de regresión y clasificación. Las clases están ordenadas y no balanceadas (e.g. hay más vinos normales que excelentes o pobres). Se pueden aplicar algoritmos de detección de outliers para detectar vinos excelentes o pobres. Además, no se conoce la importancia de las variables, por lo que podría ser interesante probar métodos para selección de características.

1.1. Información sobre los atributos

Variables de entrada (todas son numéricas):

- fixed acidity.
- volatile acidity.
- citric acid.
- residual sugar.
- chlorides.
- free sulfur dioxide.
- total sulfur dioxide.
- density.
- pH.

- sulphates.
- alcohol.

Variable de salida (numérica):

- quality (puntuación entre 0 y 10)

2. Consultas planteadas

Para el diseño de este experimento se ha tenido en cuenta la siguiente gráfica:

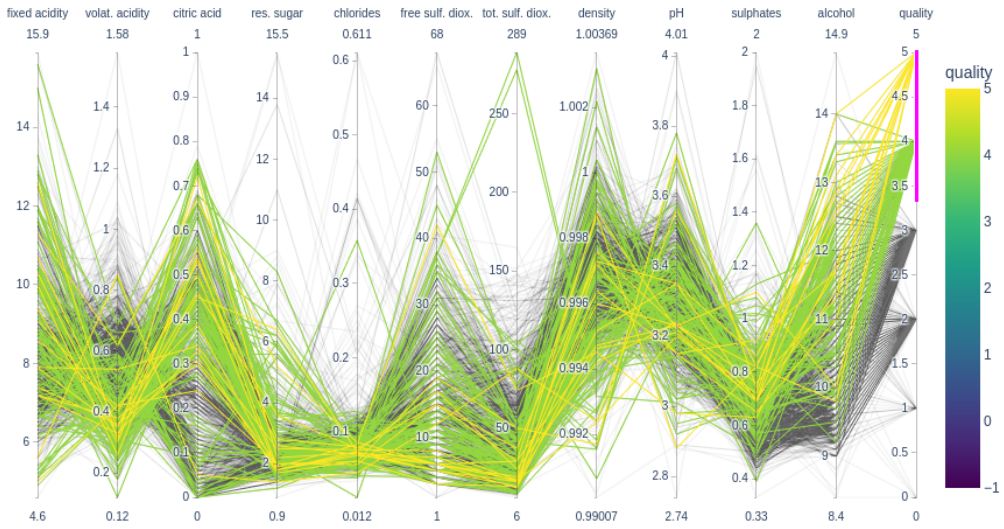

Figura 1: Distribución de valores para las variables. Valores normalizados para la variable calidad.

En este caso seleccionamos los valores con mayor calidad y vemos que la mayoría de vinos con mayor calidad tienen un valor para la variable sulfato menor a 1, por lo que podemos filtrar por calidad y sulfatos. A continuación los datos más dispersos son la acidez cítrica, densidad y alcohol así que agruparemos por calidad y calcularemos la media de esas características para cada grupo de calidad.

2.1. Carga de ficheros

Accedemos a la shell y cargamos el fichero winequality-white.csv al sistema de ficheros hdfs para utilizarlo a continuación.

```
wget https://archive.ics.uci.edu/ml/machine-learning-databases/
    wine-quality/winequality-white.csv
# mkdir /var/tmp/materialPig
mv winequality-white.csv /var/tmp/materialPig
# hdfs dfs -mkdir input
hdfs dfs -ls input
```

2.2. Creación de la tabla

Accedemos a pig, cargamos el archivo, lo mostramos por pantalla y guardamos el resultado en WineQuality.

```
pig

measure = load 'input/winequality-white.csv' using PigStorage(';') AS
(fixed_acidity:float, volatile_acidity:float, citric_acid:float,
residual_sugar:float, chlorides:float, free_sulfur:float,
total_sulfur:float, density:float, ph:float,
sulphates:float, alcohol:float, quality:float);

dump measure;
store measure into 'pigResults/WineQuality';
```

2.3. Consultas planteadas

```
# Proyección
Calidad = foreach filter_measure generate citric_acid, density,
sulphates, alcohol, quality;
# Selección
filter_measure = filter Calidad by (quality > 5) AND (sulphates < 1);
# Agrupamientos (group) y resúmenes de información (cálculo sobre grupos)
agrupados = GROUP filter_measure BY quality;
final = foreach agrupados generate group, AVG(filter_measure.citric_acid),
AVG(filter_measure.density), AVG(filter_measure.alcohol);

# Guardar resultado
store final into 'pigResults/WineQualityAvg';

# Salimos de pig
hdfs dfs -ls input
```