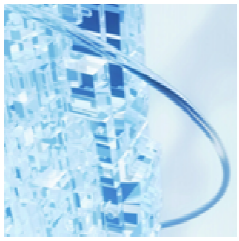


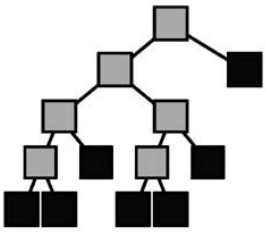
Prácticas de Minería de Datos

Grado en Ingeniería Informática

Curso 2013-14

PRÁCTICA 7

Clustering (Agrupamiento)



OBJETIVO

- Familiarizarse con el uso de las técnicas de *clustering*.
- Saber interpretar los resultados obtenidos tras el agrupamiento y evaluar la bondad del mismo.

1. Clustering con WEKA[1]

Los algoritmos de *clustering* permiten clasificar un conjunto de elementos de muestra en un determinado número de grupos basándose en las semejanzas y diferencias existentes entre los componentes de la muestra.

El agrupamiento es la tarea descriptiva más habitual. A diferencia de la clasificación, en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta, es decir, sirve para crear clases (es precisamente los grupos o clases y la pertenencia a los grupos lo que se quiere determinar a partir de un conjunto de datos de los que a priori no se sabe, ni cómo son, ni cuantos grupos hay).

Los datos son agrupados basándose en el principio de ***maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos*** (es decir, se forman grupos tales que los elementos de un mismo grupo deben ser homogéneos y muy similares entre sí y al mismo tiempo lo más diferente posible de los contenidos en los otros grupos).

Para ilustrar su funcionamiento en WEKA vamos a tratar de separar las 52 provincias españolas en *clusters* basándonos en algunas de sus características socio demográficas.

Partiremos del fichero *provincias.arff* que contiene una serie de datos para cada una de las 52 provincias. Los datos son:

- Nombre de la provincia
- Población
- Ratio varones/mujeres
- Ratio extranjeros/españoles
- Extensión de la provincia (en Km^2)
- Paro
- Número de teléfonos fijos registrados
- Número de vehículos de motor matriculados
- Número de oficinas bancarias
- Precio medio del m^2 de vivienda

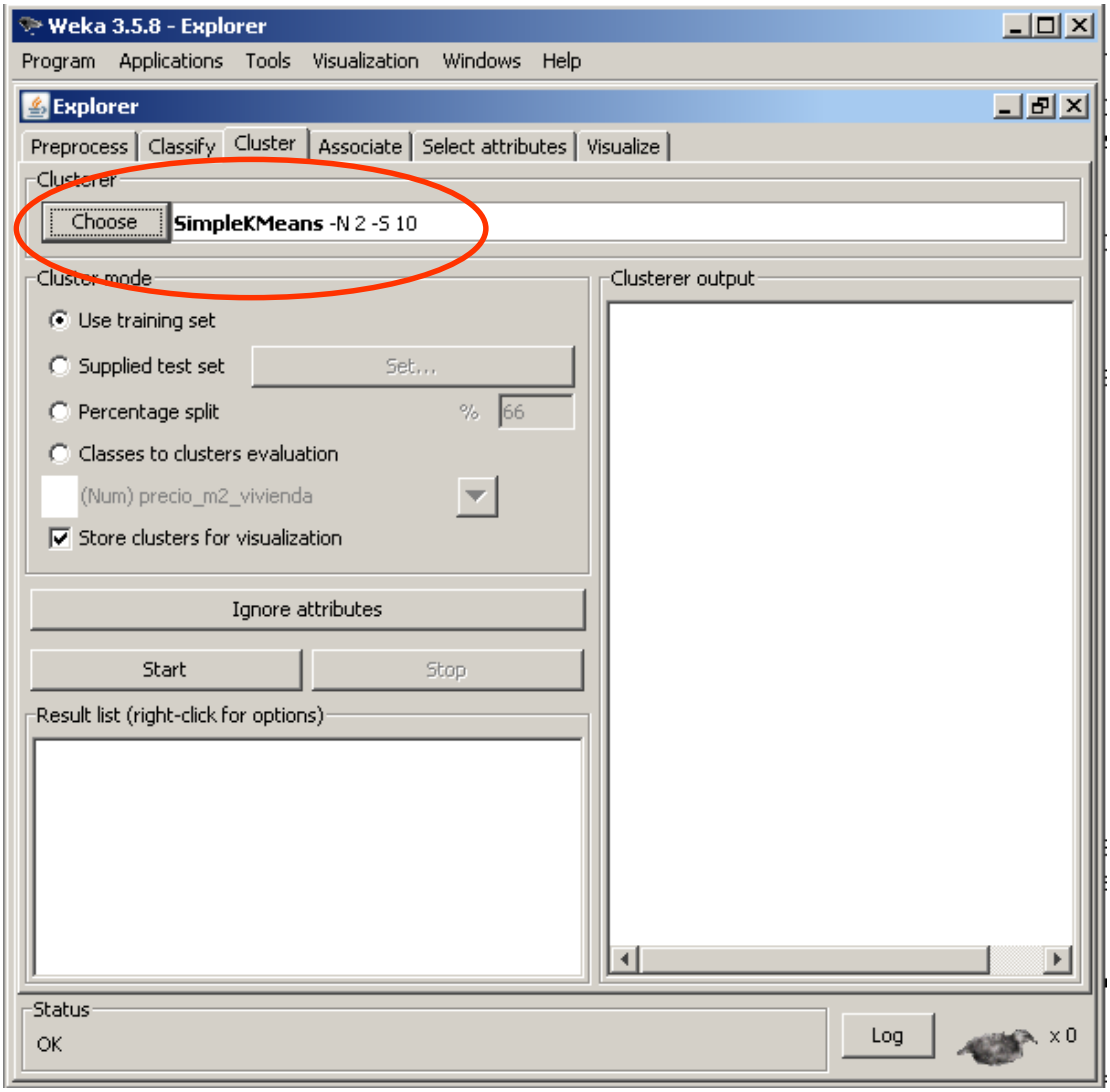
[1] Extraído de www.locualo.net

Todos los datos que contiene el fichero ARFF son reales y han sido extraídos del anuario económico de *La Caixa* del año 2006 y de la página Web *del Ministerio de Vivienda*.

Para empezar abrimos la ventana *Explorer* de Weka y desde la pestaña *Preprocess* abrimos el fichero *provincias.arff* que comentábamos antes.

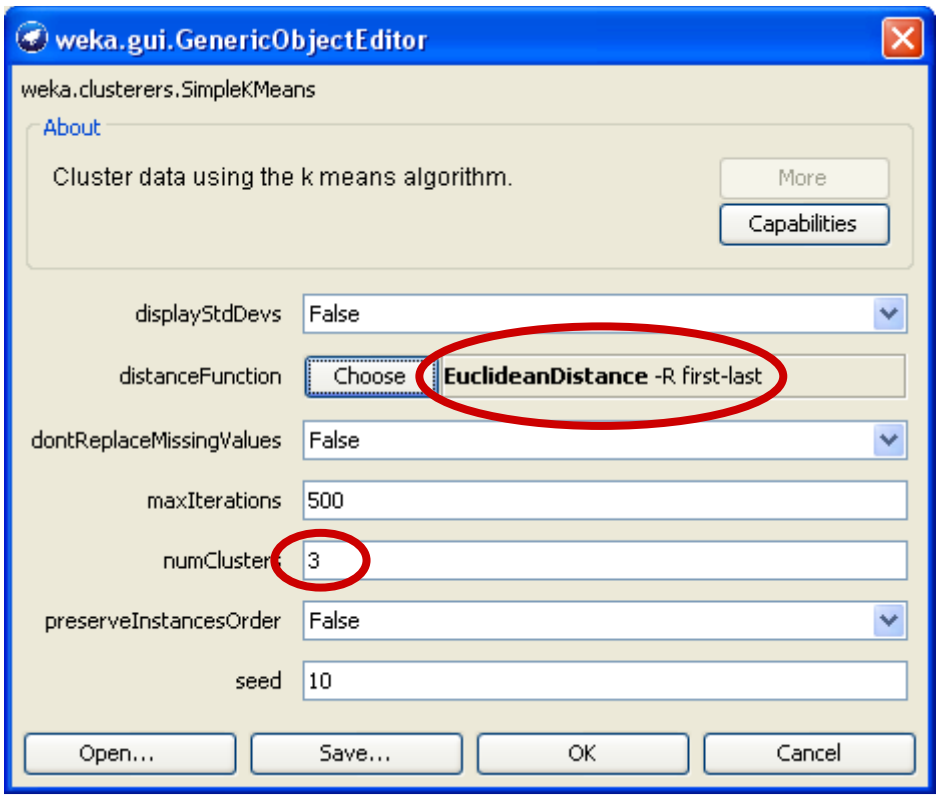
Ahora que ya hemos cargado los datos nos vamos directamente a la pestaña *Cluster*. El mecanismo de selección, configuración y ejecución es similar al de la clasificación: primero pinchamos en *Choose* y escogemos el algoritmo de *clustering* a utilizar y una vez seleccionamos configuramos los parámetros de dicho algoritmo.

En este caso vamos a seleccionar ***SimpleKMeans*** como algoritmo de *clustering*, que se caracteriza por su sencillez. Se trata de un algoritmo clasificado como método de particionado o recolocación porque representa cada uno de los *clusters* por la media (o media ponderada) de sus puntos, es decir, por su centroide. Este método sólo se puede aplicar a atributos numéricos, y los *outliers* (datos anómalos) le pueden afectar muy negativamente**.**



En este método las opciones más importantes son *numClusters* donde marcamos el número de grupos que deseamos que cree y *distanceFunction* donde indicamos que métrica de distancia vamos a utilizar (por defecto Weka utiliza la distancia Euclídea).

Para configurar el método, una vez seleccionado (*Choose*) **SimpleKMeans** como algoritmo de *clustering*, pinchamos sobre el nombre del algoritmo para configurar sus propiedades. En este ejemplo vamos a querer obtener 3 *clusters* de provincias, así que configuramos el atributo *numClusters* con valor 3 (como distancia usamos la Euclídea) y pulsamos en *OK*.



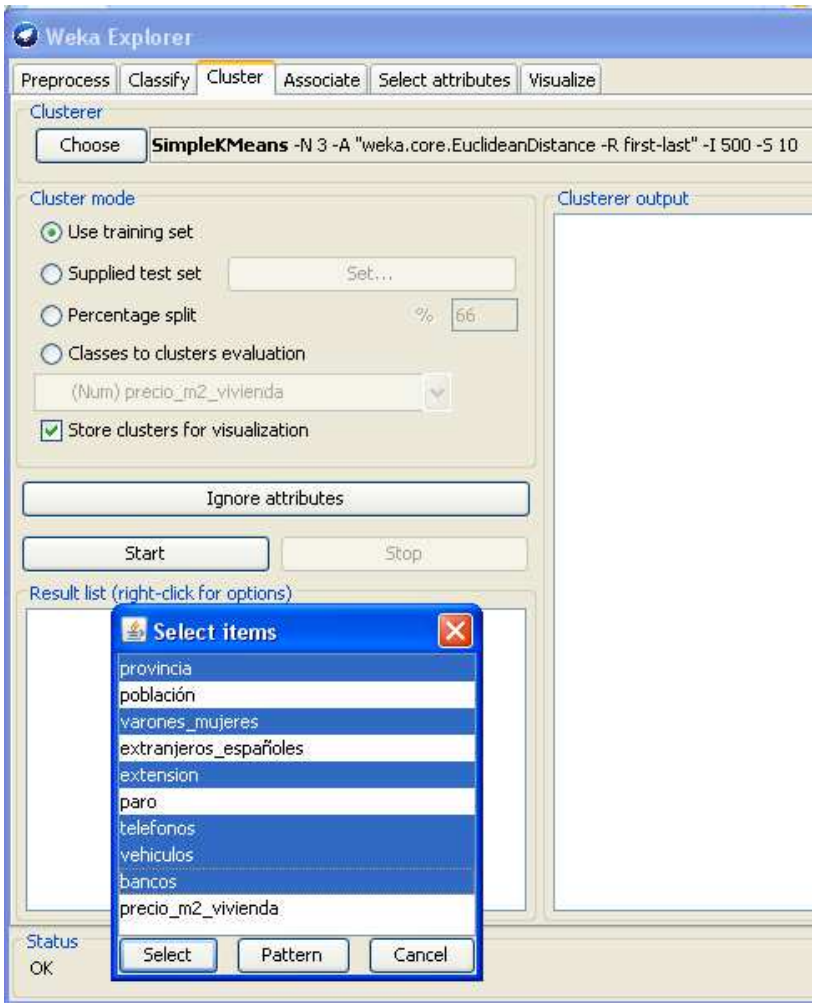
Antes de ejecutar nuestro primer clustering tenemos que seleccionar los atributos que NO queremos usar en el proceso de entre los que contiene el fichero ARFF inicial de datos.

Los atributos que seleccionemos dependen del tipo de agrupamiento o estudio que queramos realizar, aunque hay algunos atributos que, independientemente del agrupamiento a realizar, no tiene sentido utilizar.

Por ejemplo, no tiene sentido utilizar el nombre de la provincia ya que no aporta ninguna información útil para la separación en *clusters* de las provincias.

Tampoco tiene sentido utilizar la extensión de la provincia (a menos que queramos hacer un estudio por densidad de población).

En el ejemplo vamos a quedarnos con las siguientes columnas de datos: **población, ratio de extranjeros/españoles, paro y precio medio del m2 de vivienda**. Para ello pinchamos en *Ignore attributes* y seleccionamos los demás atributos.



Ahora ya estamos preparados para ejecutar el clustering así que pinchamos en *Start* y en un momento estaremos viendo los resultados.

De momento vamos a fijarnos en la siguiente sección que aparece al final del documento que Weka ha generado:

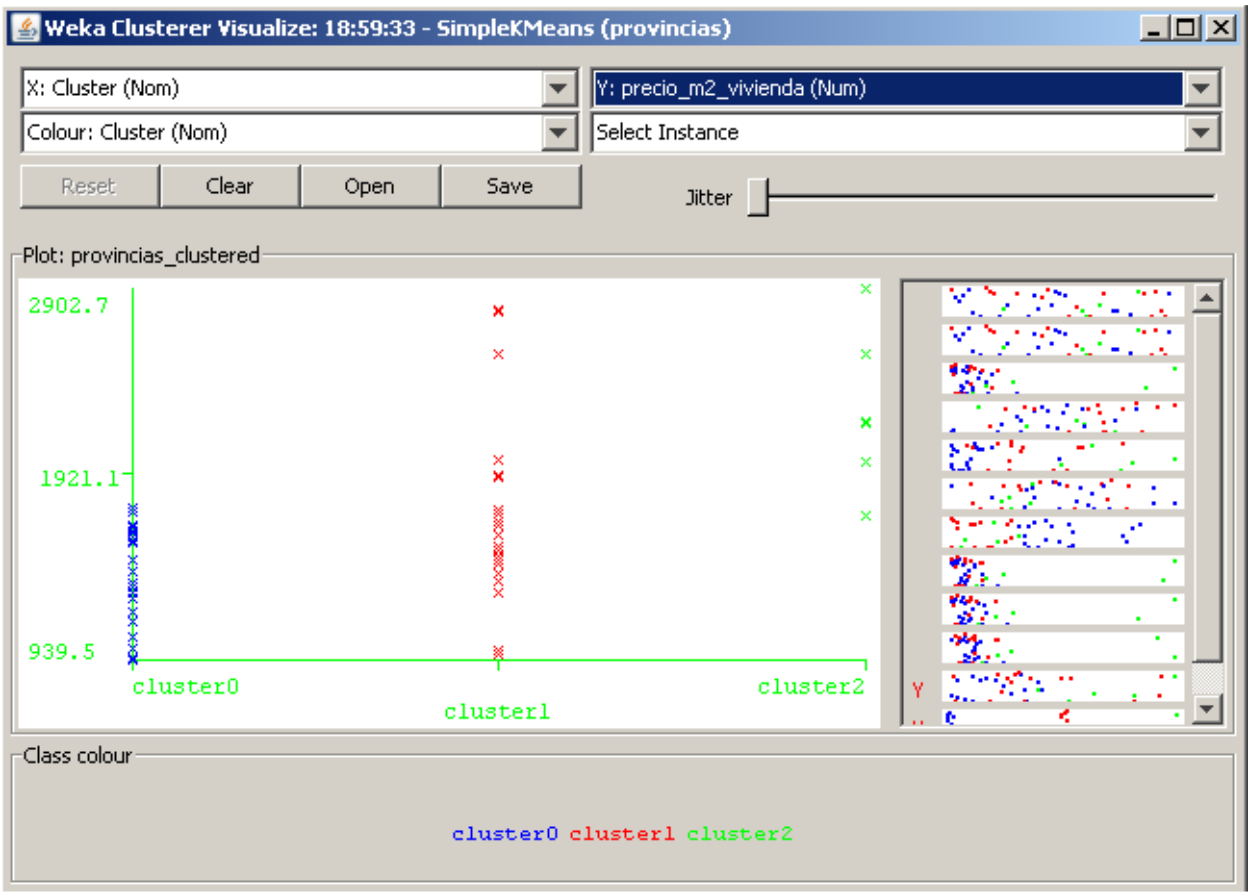
```
Clustered Instances

0       25 ( 48%)
1       21 ( 40%)
2        6 ( 12%)
```

Como se observa, se han conseguido 3 *clusters* con 25, 21 y 6 provincias respectivamente.

Podemos ver la distribución de una manera más gráfica pinchando con el botón derecho sobre la entrada correspondiente del listado de la izquierda (*Result list)* y después en *Visualize cluster assignments*. Para poder hacer esto, es necesario que antes de pulsar *Start* se haya marcado la opción *Store Clusters for Visualization*.

Desde esta pantalla podemos generar múltiples gráficas eligiendo cualquier combinación de atributos para los ejes.



Una vez obtenidos los *clusters* probablemente necesitemos guardar los resultados del agrupamiento para procesarlos o utilizarlos posteriormente. Esto se hace desde la ventana de visualización anterior (la de las gráficas) pinchando sobre el botón *Save*.

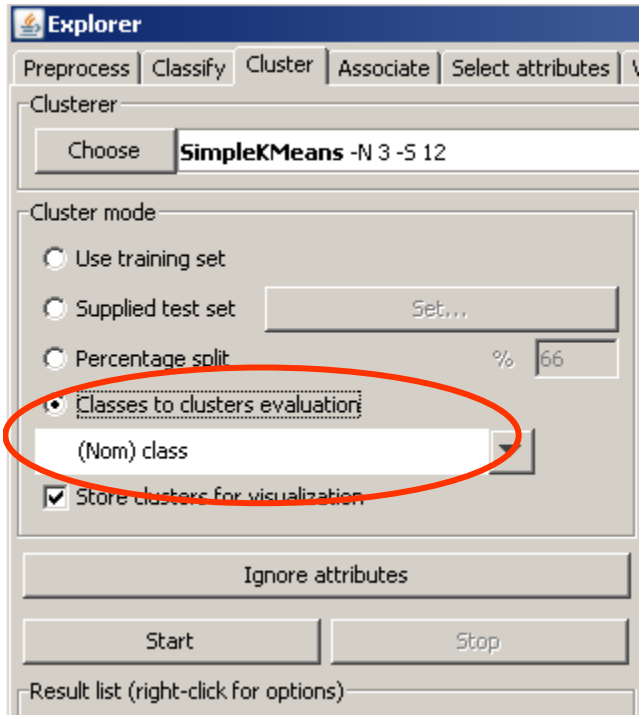
Al pulsar el botón *Save* podremos salvar en disco el agrupamiento obtenido ya que lo que hace es generar otro fichero ARFF que simplemente añade una columna con el cluster al fichero inicial de datos. Dicho fichero contendrá el cluster asociado a cada una de las instancias.

2. Evaluación de Clusters con *Clases*

La tarea de agrupamiento suele realizarse con conjuntos de datos de los cuales se desconoce la clase (aprendizaje no supervisado) a la que pertenecen las instancias, y es precisamente la tarea del clustering asignarles una clase por semejanzas entre conjuntos de instancias.

No obstante, es posible buscar agrupaciones en conjuntos de datos con instancias de clases conocidas. En estas circunstancias, es posible evaluar la calidad del agrupamiento comprobando el porcentaje de acierto de instancias de cada una de las clases en el cluster asignado.

Para poder hacer esto, necesitaremos un conjunto de datos de los cuales se conozca la clase, y seleccionar como *cluster mode* la opción *Classes to cluster evaluation*, y de entre los atributos disponibles, aquel que corresponda a la clase.



Otra de las grandes ventajas de esta forma de agrupar es que conocemos de antemano el número de clusters existentes en el conjunto de datos, que se corresponderá con la cantidad de clases distintas existentes.

Una vez seleccionada esta forma de evaluación, tras el cálculo de los clusters, se nos proporciona una matriz de confusión con las instancias de cada clase y el cluster asignado, así como el porcentaje de instancias agrupadas de forma incorrecta.

A modo de ejemplo, la salida generada con el fichero *iris.arff*, seleccionando como algoritmo de *clustering* ***SimpleKMeans*** (como distancia usamos la Euclídea) y como semilla el valor 12 (el número de *clustering* a obtener debe ser el mismo que el número de clases tenga el fichero iris.arff, en este caso 3) es la siguiente:

```
Cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2
                   (150)       (50)       (39)       (61)
=============================================================
sepallength       5.8433      5.006     6.8462     5.8885
sepalwidth         3.054      3.418     3.0821     2.7377
petallength       3.7587      1.464     5.7026     4.3967
petalwidth        1.1987      0.244     2.0795      1.418


Clustered Instances

0       50 ( 33%)
1       39 ( 26%)
2       61 ( 41%)


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 50  0  0 | Iris-setosa
  0  3 47 | Iris-versicolor
  0 36 14 | Iris-virginica

Cluster 0 <-- Iris-setosa
Cluster 1 <-- Iris-virginica
Cluster 2 <-- Iris-versicolor

Incorrectly clustered instances :       17.0      11.3333 %
```

Matriz de confusión

Instancias agrupadas de forma incorrecta

Ejercicios

1. Usando la base de datos de provincias presentada en el apartado anterior, extraiga 3 clusters que agrupen dichas provincias. Seleccione aquellos atributos que le parezcan más relevantes según el criterio de agrupación que pretenda conseguir.
    a. ¿Cuáles son los atributos que ha seleccionado y con qué intención?
    b. Describa las propiedades de los clusters encontrados.
    c. Indique 5 de las provincias asignadas a cada uno de los clusters. Según su conocimiento, ¿están correctamente asignadas?.
    d. Represente gráficamente cada uno de los atributos seleccionados en función del cluster asignado, ¿le parecen homogéneos?. ¿Qué atributo representa mejor los clusters encontrados?. Incluya el gráfico en la memoria.
    e. Utilice diferentes números de cluster para intentar ajustar los resultados. ¿Qué número de clusters le parece más apropiado?.

2. Use la base de datos *vote*, que contiene las preferencias sobre ciertas cuestiones de los votantes estadounidenses, para localizar dos clusters, cuyo porcentaje de error sea inferior al 14% sobre la evaluación de las clases conocidas. Pruebe distintas semillas del algoritmo *k-means.* ¿Qué características distintivas tienen los votantes demócratas? ¿y los republicanos?

¿Cómo entregar la práctica?

- Utilizar un documento de texto para responder a las cuestiones y subirlo a través de la plataforma Web