Capítulo 4

Clustering: Algoritmo K-Medias

4.1. Introducción

La tarea de clustering (agrupamiento) es una tarea muy frecuente en la minería de datos que trata de encontrar grupos dentro de un conjunto de individuos. El principal objetivo reside en que los individuos perteneciente a cada grupo encontrado presenten características similares entre sí y, a su vez, distintas a la de los individuos pertenecientes al resto de grupos. Por tanto, en este tipo de tarea, el concepto de distancia juega un papel importante ya que en la mayoría de los casos será el artificio matemático utilizado para medir el grado de similitud.

Aunque existe una amplia variedad de métodos de agrupamiento, aquí nos centraremos en el algoritmo K-medias por ser uno de los más populares dentro de los denominados métodos de agrupamiento por partición en contraposición a los denominados métodos por agrupación jerárquicos.

4.2. Requisitos previos

1. Haber estudiado el capítulo 16, "Métodos basados en casos y en vecindad", de [Hernández et al-04].

2. Haber estudiado el capítulo 17, "Técnicas de evaluación", de [Hernández et al-04].

4.3. Objetivos

1. Familiarizarse con *Weka* en el uso de algoritmos que resuelven tareas de *clustering*.

2. Experimentar con el algoritmo k-Medias.

3. Comprobar la dependencia del algoritmo k-Medias respecto de sus parámetros de configuración:

   - Semilla que permite generar los centroides iniciales.
   - Número de clusters.

4. Comprobar la dependencia del algoritmo k-Medias respecto del poder de predicción de los atributos utilizados.

4.4. Material

- Programa Weka.
- Archivo "iris.arff".

4.5. Actividades

1. Cargar el fichero ïris.arff"desde el panel *Preprocess* de *Explorer*. El contenido de este archivo es una colección de instancias que representan 3 variedades de lirio. Se compone de cinco atributos. Los cuatros primeros corresponden a diversas medidas morfológicas de la planta (anchura y longitud de sépalos y pétalos) y el quinto corresponde al valor de la clase[1] (tipo de lirio). Para más detalle consulte la información de cabecera de dicho archivo.

2. En la ventana *Explorer*, seleccionar el panel *Cluster* (ver figura 4.1).

3. Ignorar el atributo *class* para evitar que éste intervenga en el proceso de agrupamiento al ejecutar el algoritmo. Para hacer esto, pulsar el botón Ignore attributes y, en la nueva ventana emergente, seleccionar el atributo en cuestión (el atributo clase en nuestro caso). Finalmente, pulsar el botón Select.

4. Seleccionar el algoritmo *SimpleKMeans* (k-Medias) y configurar sus parámetros: el parámetro *numCluster* se pone a 3 puesto que éste es el número de grupos (3 tipos de lirios) que se espera obtener. Nótese que en muchas aplicaciones reales, el número de clusters no se sabe de antemano y, tal como se indicará más adelante, esto obligará a realizar varios análisis variando el valor de este parámetro y evaluando cada vez la bondad de los clusters obtenidos. El parámetro *seed,* cuya finalidad se comentará más adelante, se utilizará por ahora con su valor por defecto.

[1] Aunque la tarea de *clustering* tiene entre otras condiciones de contorno el no conocimiento de la clase a la que pertenecen las instancias, por motivos pedagógicos, se utilizará aquí este fichero de datos respetando las indicaciones que se hacen en las distintas actividades.

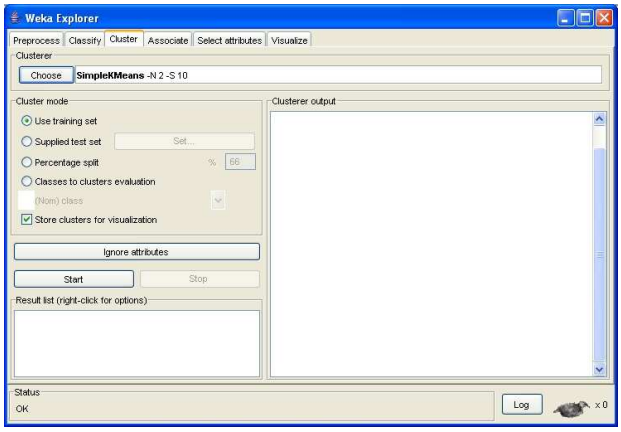

Figura 4.1: Panel *Cluster* de la ventan *Explorer*.

5. Seleccionar ejemplos entrenamiento/test: en la caja *Cluster Mode*, seleccionar la opción *Percentaje split* y rellenar el campo % con el valor 66. Esto hará que se utilice 2/3 (66 %) de la base de datos para entrenamiento y el resto (1/3) para evaluación. En esta misma caja del panel, activar también la casilla *Store clusters for visualization* (esto permitirá almacenar datos del análisis que podrán usarse posteriormente para tareas de visualización).

6. Ejecutar el análisis (botón Start).

7. Interpretación de resultados alfanuméricos. En la ventana de resultados (*Clusterer Output*) la salida del análisis se muestra agrupada bajo dos epígrafes:

   - En el epígrafe *Clustering model* (*full training set*) se muestra el valor de las coordenadas de los distintos centroides (prototipos) asociados a cada cluster, considerando el conjunto de datos completo y en términos del valor medio y de la desviación estándar de cada atributo (si éste es numérico) o sólo de la moda (si es nominal).
   - Bajo el epígrafe *Evaluation on test split* se muestra las coordenadas nuevamente de los distintos centroides del modelo pero esta vez calculados sólo a partir del conjunto de entrenamiento (66 % de los datos originales en nuestro caso). Obsérvese que los centroides obtenidos para cada *cluster* son ligeramente diferentes al del caso anterior. Finalmente, incluido dentro de este epígrafe, también se muestra el número y porcentaje de ejemplos del conjunto de test asignados a cada cluster según el modelo construido a partir del conjunto de entrenamiento.

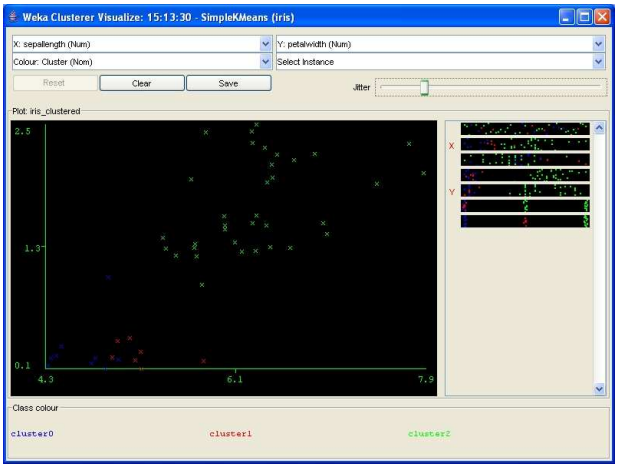

Figura 4.2: Ventana gráfica que permite representar la información obtenida del análisis de clusters.

- **Documentación a entregar**: Mostrar en una tabla y comparar las coordenadas de los centroides obtenidos utilizando todos los ejemplos de la base de datos frente a los obtenidos sólo utilizando los ejemplos de entrenamiento.

8. Interpretación de resultados gráficos: dentro de la caja *Result List* seleccionar con el botón derecho del ratón la última entrada producida (corresponde a la última ejecución realizada) y, sobre el menú desplegable, seleccionar la opción *Visualiza cluster assignments* que abrirá una ventana gráfica (ver figura 4.2). Esta ventana permite representar la distribución de los clusters obtenidos en función de los distintos atributos tomados de dos en dos: uno en el eje X, otro en el eje Y y utilizando la variable *cluster* como opción de color (*colour*). Otra representación interesante podría ser la visualización de la variable *cluster* en el eje Y en función de la variable *Instance_number* (N^o de instancia) en el eje X. De esta manera se podría saber qué instancia se asocia a cada cluster. Se puede visualizar la información asociada a cada instancia (valores de sus atributos) sin más que pulsar sobre el símbolo correspondiente con el botón izquierdo del ratón[2]. Para obtener información más detallada del resto de controles de esta ventana consultar el documento '*Weka: Guía de usuario de Explorer*'.

   - **Documentación a entregar**: representar, por ejemplo, cuatro gráficas utilizando distintos pares de atributos (eje Y-eje X) y la variable *cluster* como color. Finalmente, representar también la variable *cluster* (eje Y) frente a la variable *Instance_number* (eje X). Comentar

[2]No olvide utilizar la barra *Jitter* para visualizar posibles instancias solapadas.

los resultados.

9. Interpretación de resultados gráficos conocido el valor de clase: en los casos en los que se conoce el valor de clase, Weka proporciona una forma alternativa de evaluar la bondad de los cluster obtenidos. Obsérvese que ésta sería una situación atípica en un problema de clustering puesto que, en este tipo de problemas, lo que se pretende es precisamente encontrar agrupaciones de instancias similares cuando se desconoce la clase a la que pertenece cada una de ellas. No obstante, podría ser una opción útil si se dispusiese de un conjunto de datos clasificado y se quisiese investigar la tendencia de los distintos valores del atributo clase a formar grupos disjuntos. La forma en que Weka permite utilizar el valor de clase es la siguiente:

   - Desde el panel *Cluster* de *Explorer* se selecciona la opción *Classes to clusters evaluation* (dentro del la caja *Cluster Mode*) y, en el campo con menú desplegable asociado, se selecciona el atributo *class*. Tras ejecutar el análisis[3] y lanzar nuevamente la ventana gráfica asociada a dicho análisis, se visualiza la variable *cluster* (eje Y) respecto de la variable *Instance_number* (eje X) y se utiliza la variable *class* como color. En la gráfica correspondiente, mediante un cuadrado como símbolo, aparecerán todas aquellas instancias que el modelo asigna a clusters erróneos (desde el punto de vista del valor de clase). En la caja *Clusterer Output* del panel *Cluster* también se muestra una matriz de confusión con los porcentajes de acierto y error obtenidos por cada cluster.
   - **Documentación a entregar**: realizar un análisis con las opciones aquí indicadas y representar la gráfica correspondiente comentando la bondad de las agrupaciones obtenidas. Mostrar la matriz de confusión asociada al análisis realizado.

10. El método de las k-Medias es sensible a la inicialización aleatoria de centroides (prototipos) que se realiza en el primer paso del algoritmo. De hecho, el modelo de centroides aprendido puede variar significativamente según sea esta inicialización. En Weka, esto se controla mediante el parámetro *seed* (semilla), que representa un número entero en función del cual se generan los prototipos iniciales. Si este valor no se modifica entre la realización de dos análisis consecutivos iguales, los clusters obtenidos en ambos análisis también serán los mismos.

    - Para comprobar este hecho, ejecutar el análisis que se hizo en el apartado 9 modificando cuatro veces (por ejemplo: 51, 88, 99, 139) el valor del citado parámetro.

[3] Asegurese que el atributo *class* no interviene en al análisis habiéndolo descartado previamente con el botón Ignore attributes.

- **Documentación a entregar**: Muestre en una tabla y compare las coordenadas de los clusters obtenidos en cada análisis. Represente, para cada análisis, el tipo de gráfica indicada en el apartado 9 y comente y compare la bondad de las agrupaciones obtenidas en cada caso.

11. El algoritmo k-Medias es también sensible al poder de predicción de los atributos utilizados. Así, el uso conjunto de atributos de bajo valor predictivo (incluso nulo) con otros de alta predicción puede resultar en una disminución de la bondad de los clusters obtenidos.

    - Repita el análisis realizado en el apartado 9 pero ahora, además de descartar el atributo de clase, tal y como allí se hizo utilizando el botón "Ignore attributes", realice, por ejemplo, cuatro análisis con distintos descartes de atributos[4].
    - **Documentación a entregar**: Represente, para cada análisis, el tipo de gráfica indicada en el apartado 9 y comente y compare la bondad de las agrupaciones obtenidas en cada caso. ¿Qué ocurre si sólo se seleccionan para el análisis los atributos "*petallength*" y "*petalwidth*"?.

12. Igualmente podría ser interesante actuar sobre el número de clusters iniciales k (en Weka: *numCluster*) si se sospecha que no se está obteniendo una partición de clusters adecuada. La evaluación certera del número de agrupamientos obtenidos sólo puede realizarse si se conoce de antemano dicho número. En otro caso, no hay una forma rotunda de evaluar el número de grupos obtenidos. Para solucionar este inconveniente, se suele utilizar (ver [Hernández et al-04], pp. 433-434) algún otro algoritmo de clustering. Entonces, para un mismo valor de k, se lanzan sendos algoritmos. Este proceso se repite variando el valor de k. Aquel caso que arroje grupos con un número de individuos similar en el modelo producido por cada algoritmo es el que determinará el valor de k más adecuado.

    - Repetir el tipo de análisis realizado en el apartado 9 para tres valores (por ejemplo: 2, 3 y 4) distintos del parámetro *numCluster*.
    - **Documentación a entregar**: Represente, para cada análisis, el tipo de gráfica indicada en el apartado 9 y comente y compare la bondad de las agrupaciones obtenidas en cada caso.

4.6. Conclusiones

En la documentación a entregar, enumere finalmente de forma esquemática las conclusiones derivadas de los resultados prácticos obtenidos con la realización

[4] Para hacer una selección múltiple de atributos en la ventana resultante de aplicar el botón "*Ignore attributes*", presionar simultáneamente la tecla *Control* y el botón izquierdo del ratón sobre cada uno de los atributos a descartar.

de la prácticas. Hágalo tanto desde un punto de vista general relativo a la tarea de clustering como desde un punto de vista particular referido al algoritmo k-Medias.

4.7. Referencias

[Hernández et al-04] J. Hernández, M.J. Ramírez y C. Ferri, "*Introducción a la minería de datos*", Pearson-Prentice Hall, 2004.