INFORME LOS COCHES DEL JEFE II (REPARTO)

Abstract

En este informe se emplean técnicas de análisis clúster para partiendo de una base de datos que almacena información sobre 125 vehículos de los que se acumula información en quince variables repartir estos modelos en una serie de puntos geográficos de manera por una parte óptima (coches similares en lugares próximos) y por otra parte consistente (la clasificación no debe ser puramente matemática si no que debe tener en cuenta las características más relevantes de cada vehículo).

Una vez realizado en el informe anterior (ver Informe los coches del jefe) el análisis exploratorio de datos, en este segundo informe se afronta el problema de la clasificación y reparto de los coches en las distintas propiedades del jefe que se encuentran distribuidas según las localizaciones marcadas en el mapa (**Figura 1**). Anteriormente se había construido un modelo con nueve clústers empleando todas las variables disponibles. En este informe se plantearán dos modelos más; uno construido a partir de las variables establecidas en el primer informe y un segundo refinado eliminando de manera justificada dos variables más.

Figura 1. Localización de las distintas propiedades.

En todos los modelos presentados a continuación se emplean para la clasificación tanto variables cuantitativas (precio, rpm, potencia…) como cualitativas (número de plazas, aceleración, cilindros…) por lo que es necesario emplear una métrica que permita trabajar con ambos tipos de variables. En este caso se empleará la métrica Gower. La gran desventaja de que esta métrica es que es muy intensiva computacionalmente pero como se trabaja con una muestra (relativamente) pequeña esto no es un problema para este estudio.

El procedimiento seguido para el cálculo de los modelos es el siguiente: tras calcular la matriz de distancias con la métrica Gower se comprueba cuáles son los elementos más próximos y más dispares. Esto sirve como comprobación del cálculo pues los modelos más próximos salen

prácticamente similares y los modelos dispares presentan diferencias radicales en cada una de las características valoradas.

Tras ello es necesario elegir un algoritmo que genere los clusters, para la elaboración de estos modelos se ha elegido el algoritmo PAM (Partitioning Around Medoids). Se elige dicho algoritmo por su robustez ante el ruido blanco y a los valores atípicos. Su mayor desventaja es que es de orden cuadrático tanto en tiempo como en memoria pero al trabajar con una base de datos de tamaño reducido esto no supone un problema.

Mediante este algoritmo se generan todos los clusters posibles entre dos y diez para ver cuál sería el número óptimo de clusters (no confundir con el número óptimo de garajes). A la hora de decidir cuál es el número adecuado de clusters se emplea una métrica de validación interna que mide como de similares son las observaciones a las de su propio clúster comparadas con las del clúster más próximo (silhouette width). La manera más sencilla de interpretar esta medida es mediante un gráfico en el que el punto con mayores valores en la ordenada se corresponde en la abscisa con el número óptimo de clusters.

Aclarada la metodología se procede a explicar los diferentes modelos construidos.

(**Nota:** Con el objetivo de no sobrecargar el informe con información innecesaria todos los resultados a los que se hace alusión están indicados y comentados en el código para el lector que deseé una mayor profundización.)

Versión general

En un primer momento se había planteado un análisis clúster que contemplara todas las variables presentes en la base de datos. Sobre la matriz de distancias construida se realizan las comprobaciones antes mencionadas (cálculo de los modelos más próximos y más lejanos) obteniendo resultados muy consistentes. Tras calcular la silhouette width y construir el gráfico se observar que el número óptimo de clusters es 9. En el código se muestra como se realizaría la asignación de los coches a los diferentes clusters.

Versión basada en el análisis exploratorio del informe anterior

En el informe antes citado se llega a la conclusión de que las variables marca, consumo a 90 km/h y aceleración de 0 a 100 son prescindibles por lo que se eliminan y se realiza un análisis idéntico al expuesto anteriormente. En este caso el Silhouette width nos indica que el número óptimo de clusters es dos.

Aunque estas conclusiones son perfectamente válidas a partir del análisis exploratorio, no parece muy satisfactorio a nivel operativo distinguir solo dos tipos de coches luego aunque esta sea una posibilidad se descarta en busca de modelos que permitan distinguir de manera más precisa los automóviles.

Versión mejorada partiendo del análisis exploratorio

En esta versión además de descartarse las variables anteriormente mencionadas se eliminan los cilindros pues no aportan una información muy diferente a la que aportan las cilindradas y la aceleración pues realmente esta no es importante a la hora de clasificar vehículos.

Empleando de nuevo la métrica Gower se construye la matriz de distancias y se realizan las comprobaciones pertinentes (que ofrecen de nuevo resultados consistentes). A continuación se emplean PAM y el silhouette width para decidir el número óptimo de clusters para la clasificación.

La gráfica (**Figura 2**) muestra que en este caso el número óptimo de clusters es cinco:

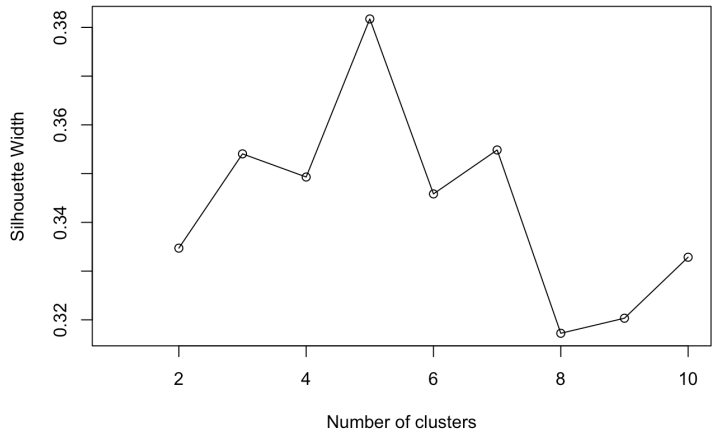

Figura 2. Calidad de la clasificación según el número de clusters.

Una vez obtenido el número óptimo de clusters se procede a la construcción de estos mismos mediante el algoritmo PAM antes mencionado. En el código adjunto se presenta un pequeño resumen de las características de los coches de cada clúster (mínimo, máximo, media, mediana y cuartiles). Tras ello se visualiza la distribución de los vehículos en cada clúster (**Figura 3**) donde se aprecian tres grupos claramente diferenciados (1, 2 y 5) estando el 3 y el 4 algo más mezclados con los demás grupos.

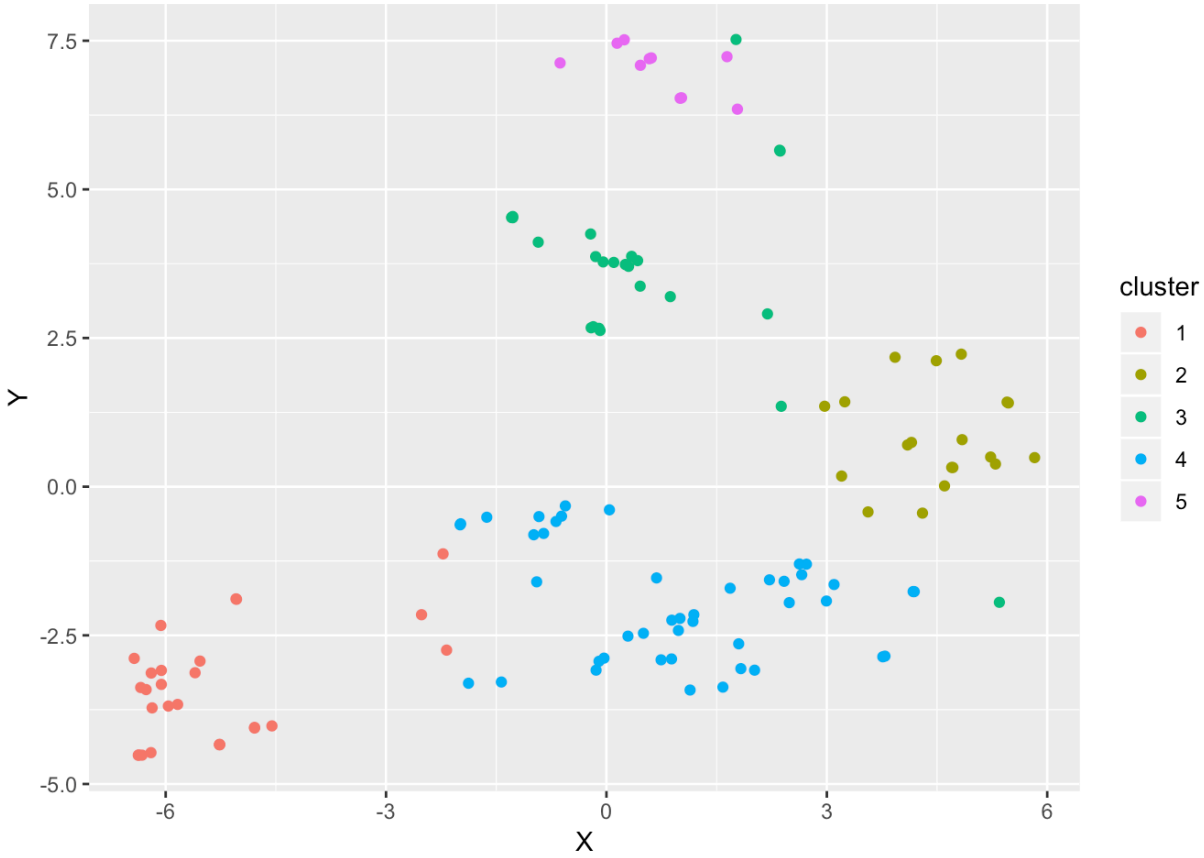

Figura 3. Distribución de los clusters.

Una vez se dispone del análisis clúster resulta más sencillo realizar el reparto entre las distintas propiedades así se calcula el número de elementos por clúster. En el clúster 1 se tienen 27, en el 2: 19, en el 3: 25, en el 4: 44 y en el 5: 10.

El punto más inaccesible de los propuestos en el mapa es el situado en Córcega por lo que en el se situarán los coches del clúster 5. El clúster 4 por el contrario requiere tres propiedades muy próximas por lo que se asignan sus vehículos a las propiedades de la Costa Azul. Los clústeres restantes requieren todos ellos dos garajes por lo que esto se otorgarán en pares más próximos por número de coches así los autos del clúster 1 irán a París, los del 3 a Suiza y los del 2 a La Rochelle y Andorra.

Una vez realizado el reparto y para completar el informe se presenta un pequeño resumen de los coches almacenados en cada propiedad:

En las propiedades de La Rochelle y Andorra se almacenan los coches con mayor precio, velocidad, potencia y cilindrada. Estas cuatro variables están muy asociadas (de manera directa) entre ellas. Son además los vehículos que mayor consumo tienen. Se podría decir que en términos generales estas propiedades se reparten los vehículos de mayor nivel.

En Córcega se han situado los coches de menor velocidad, entre ellos se cuentan los modelos con datos “atípicos” para el número de plaza, es decir, los modelos pick-up o biplaza. Son modelos muy eficientes presentando los menores consumos de la tabla a 120 km/h.

En París se almacenan los vehículos con menor cilindrada y mayor número de revoluciones por minuto. Además son vehículos con un consumo urbano óptimo lo cual es un punto más a favor para ser almacenados en la capital.

Todos los modelos almacenados en la Costa Azul tienen como característica común que son vehículos de cinco plazas. Por lo demás en esta propiedad se almacenan aquellos vehículos que no tienen características destacables respecto a los demás. Se podría decir que es el cajón de sastre en esta clasificación donde se introducen los coches de gama media.

Código

El código sobre el que se fundamentó tanto el análisis exploratorio previo como la creación de los distintos modelos clusters se encuentra disponible en el siguiente enlace.