MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

Análisis Inteligente de Datos

Trabajo Práctico

Clasificación supervisada y no supervisada

Alumno: MARIA JOSE FERREYRA VILLANUEVA. DNI 23297670

Fecha: Agosto 2020

INTRODUCCIÓN y OBJETIVOS

En el presente trabajo práctico se introducen los conceptos de clasificación supervisada y no supervisada. Se analizan dos conjuntos de datos provistos por la cátedra, aplicando técnicas de clasificación abordadas durante el curso de la materia.

Para el caso de la clasificación supervisada, se usó el conjunto de datos de *accidente*, tomando el 90% del conjunto de datos original, previa aplicación de una semilla random basada en el nro. de DNI del alumno. La tarea consiste en encontrar el algoritmo que mejor identifique los patrones de la clase objetivo. Para esto se prueban varios algoritmos vistos en clase y se analizan los resultados con dos métricas de performance elegidas.

Para el caso de la clasificación no supervisada, se seleccionaron 5 variables del conjunto de datos de *telecomunicaciones*. Se aplicaron dos de las técnicas de clustering vistas en clases con el objetivo de obtener una agrupación que resulte de las similitudes entre las variables originales seleccionadas.

Junto con este informe se adjuntan los conjuntos de datos utilizados, en dos archivos de formato *csv* y dos archivos formato *rmd*, uno por cada tipo de clasificación estudiada. Allí se encuentran los gráficos que aquí se mencionan y el código R que origina este informe.

A continuación, se describen ambos procesos con sus respectivas conclusiones.

CLASIFICACION SUPERVISADA

Selección de variables y análisis exploratorio

El conjunto de datos de accidentes viene conformado por 36 observaciones y 6 variables. Se enumeran los nombres y características de cada una:

- *Vehículo:* numérica, identificador de registro y sus valores arrancan desde el número 1 en adelante.
- *Antigüedad:* numérica, indica la antigüedad de vehículo en años. Los valores están entre 1 y 15 años.
- *Edad del conductor:* numérica, indica la edad del conductor en años. Los valores están entre 19 y 65 años.
- *Potencia*: numérica, se asume que indica la potencia en CV (caballos de vapor) del vehículo. Los valores que se observan están entre 7 y 130 CV.
- *Grave*: categórica, variable objetivo o de clase. Toma valores 0 para *leve* y 1 para *grave*. Hay 21 observaciones para la categoría *leve* y 15 para la *grave*.

La variable *vehículo* se eliminó del conjunto de datos, porque es un identificador de registro y toma un valor distinto para cada observación, por lo tanto no sirve para la clasificación.

Se exploraron los boxplots del resto de las variables en busca de valores atípicos o características especiales en su distribución. Se observó que la variable *potencia* presentaba un valor atípico. Se trata de una observación que tiene valor de 7 CV. El resto de las observaciones tiene valores entre 60 y 130. Al investigar un poco en la web sobre los valores posibles de CV, se identificó el valor como un error, por lo cual se decidió eliminar el registro del conjunto de datos.

El conjunto de datos quedó finalmente conformado por 35 observaciones y 4 variables.

Se buscaron valores atípicos multivariados sobre el conjunto de datos, aplicando la distancia de mahalanobis y, utilizando el método del determinante de mínima covarianza (MCD). Se encontraron 8 observaciones con las mayores distancias a partir un umbral definido en 7.

Se observó el dispersograma, identificando con color la clase objetivo. Las observaciones se ven muy dispersas y los colores que identifican las clases, muy mezcladas; a primera vista no se aprecia ninguna relación que discrimine de forma óptima la clase objetivo. Esto último se confirma observando el histograma por variable, identificando también la clase con color. En el dispersograma, tampoco se observa una relación lineal entre pares de variables, esto también se distingue en la matriz de correlación, donde el mayor índice de correlación es de 0.39 (positiva) entre *la edad del conductor* y la *antigüedad del vehículo*.

Por último, analizamos las distribuciones de todas las variables agrupándolas por la clase objetivo. Se testeó normalidad con la prueba de Shapiro y de Anderson, y se analizaron los gráficos de cuantiles e histogramas comparando con sus respectivas funciones de densidad normal correspondiente a su media y desvío. Estos gráficos pueden observarse en el archivo *rmd* adjunto. La variable *edad del conductor*, para la clase *leve (0)*, fue la única cuyo p-valor no rechazó la normalidad. El resto de los p-valores resultaron menores a 0.05, rechazando la hipótesis nula de normalidad. Se muestran a continuación, los p-valores para las pruebas mencionadas.

```
##   antigüedad edad.conductor potencia     Test Grave
## 1     0.0013         0.2109    1e-04  Shapiro     0
## 2     0.0082         0.0155   0.0019  Shapiro     1
## 3      5e-04         0.3155        0 Anderson     0
## 4     0.0039         0.0151    4e-04 Anderson     1
```

Métodos de clasificación y supuestos

Hay varios pasos que queremos completar previo a la aplicación de los algoritmos de clasificación.

El primero de ellos es aplicar el test de Hotelling para comparar las medias multivariadas. Para ello, primero analizamos los vectores de medias en forma gráfica, con el grafico de perfiles de medias. Se puede apreciar el grafico en el archivo adjunto, y allí vemos que los perfiles no son paralelos y las líneas de las clases se cruzan. El resultado del test de Hotelling arroja un p-valor de 0.0008714, estableciendo que hay suficiente evidencia para afirmar que las medias no son iguales. Este test nos da la pauta de que vale la pena utilizar los algoritmos de clasificación sobre los datos, ya que hay evidencia de que los grupos significativamente distintos.

El segundo paso es comprobar los supuestos de algunos de los algoritmos que vamos a aplicar, éstos son lo de: normalidad multivariada y homocedasticidad multivariada.

Aplicamos el test de Shapiro en su versión multivariada para ambos grupos. Los p-valores resultantes fueron: 0.03424 para la clase *leve* y 0.002708 para la clase *grave*. En ambos casos, tenemos suficiente evidencia para rechazar la hipótesis de normalidad. El test M de box arrojó un p-valor de 0.0317 para la prueba de homocedasticidad. Añadimos la prueba de Levene multivariada, como alternativa robusta al test M de Box que podría verse afectada por la falta de normalidad. El p-valor resultante fue de 0.6087, por lo tanto, concluimos, que no tenemos suficiente evidencia para rechazar la hipótesis nula de homocedasticidad.

Verificamos el balance de nuestras clases en el conjunto de datos. Comprobamos que la clase *leve* (0) representa un 60% del conjunto de datos y la clase *grave* (1), un 40%.

A continuación, veremos los resultados de aplicar distintos algoritmos de clasificación supervisada a nuestro conjunto de datos. En todos ellos, para estimar la probabilidad de clasificación correcta, lo hicimos con dos alternativas: muestra de entrenamiento y de validación y leave one out (LOO en adelante). Al ser un conjunto de datos con pocos valores, las muestras de entrenamiento y validación estuvieron confirmados por 25 y 10 observaciones cada uno (70%-30%). Es importante tener en cuenta que en el método LOO, se arma un modelo del algoritmo, por cada observación, por lo tanto, sus coeficientes pueden variar de un caso a otro.

Las medidas de performance que se evaluaron fueron, el accuracy y el recall. El primero nos da un indicador de la cantidad de observaciones que fueron predichas correctamente, y el segundo, hace hincapié sobre la cantidad de casos positivos (*graves*) que el modelo es capaz de identificar. Por las características del conjunto de datos, consideramos que es importante abarcar todos los casos en que un accidente sea *grave*, y minimizar la cantidad de falsos negativos, por eso, el recall, es un indicador importante a tener en cuenta a la hora de escoger el mejor algoritmo.

El primero algoritmo que aplicamos, fue el Análisis Discriminante Lineal (LDA). Al analizar los valores absolutos de los coeficientes del modelo resultante, la variable

antigüedad con un coeficiente lineal de 0.15710903, es la más importante a la hora de predecir la clase. La teoría nos dice que el LDA sólo es válido si se satisfacen los supuestos de independencia de las observaciones, y normalidad y homocedasticidad multivariante.

Luego aplicamos el Análisis Discriminante Cuadrático Robusto (QDAR), el cual surge como opción al LDA, cuando no existe homocedasticidad; aunque esta técnica asume normalidad multivariante también. Decidimos aplicar la versión robusta de este algoritmo, ya que utiliza el determinante mínimo de la matriz de varianzas-covarianzas (MCD) como alternativa ante la presencia de valores atípicos en el conjunto de datos. Al existir homocedasticidad, pero no la normalidad, la aplicación de este modelo tampoco sería válido.

Como siguiente opción, aplicamos Regresión Logística (LG). Este modelo no necesita la premisa de normalidad multivariante y es específico para casos de clasificación binaria. En la tabla de coeficientes del modelo entrenado, la *edad del conductor* y la *potencia*, tienen un p-valor de 0.0211 y 0.0426 respectivamente, con lo cual resultan estadísticamente significativas para la predicción de la gravedad del accidente. En cambio el coeficiente de *antigüedad* tiene un p-valor de 0.0578, y está en el límite del valor que la define como significativa.

El test de Hosmer-Lemeshow del modelo de regresión logística nos arroja un el p-valor igual a 1, por lo tanto no rechazamos la hipótesis nula que sostiene que el modelo se ajusta a los datos.

Por ultimo aplicamos los algoritmos de Maquina de Soporte Vectorial (SVM) y Random Forest (RF) (visto en la materia de Aprendizaje automático).

Las matrices de confusión de cada modelo pueden verse en el archivo adjunto. Abajo se muestran dos tablas con los valores de accuray y recall resultantes para todos los modelos.

Accuracy

```
##      Validation  LOO
## LDA         0.7 0.74
## QDAR        0.7 0.74
## LG          0.6 0.74
## SVM         0.7 0.69
## RF          0.7 0.74
```

Recall

```
##      Validation  LOO
## LDA        0.50 0.57
## QDAR       0.75 0.50
## LG         0.50 0.57
## SVM        0.75 0.57
## RF         0.50 0.71
```

Conclusión

Todos los modelos tuvieron un accuracy similar, con excepción de LG y SVM para los casos de Validacion y Loo respectivamente, donde su performance fue un poco menor. En la métrica de recall, los algoritmos que se destacaron, fueron QDAR y SVM para el caso de Validacion y RF para Loo.

Nuestros candidatos al mejor modelo, se reducen a aquellos con mayor valor de recall: SVM y QDAR. Sabemos que el conjunto de datos no cumple con el supuesto de normalidad multivariante, el cual es necesario para que el modelo de QDAR sea válido. Por lo tanto nos inclinamos por SVM, como mejor opción, dándole mayor importancia a capturar la mayor cantidad de accidentes graves.

CLASIFICACION NO SUPERVISADA

Selección de variables

El conjunto de datos de telecomunicaciones tiene 36 variables, equitativamente repartidas entre categóricas y numéricas, y 1000 observaciones, donde cada una representa un individuo. Ante la oportunidad de utilizar las técnicas de clustering, la intención fue agrupar a los usuarios teniendo en cuenta características sociales y del servicio contratado; y obtener clases que sean suficientemente separables, con el objetivo de descubrir asociaciones útiles en los datos. Teniendo en cuenta que los algoritmos van a tratar de establecer similitudes entre las variables utilizando distancias o medidas de similitud, es conveniente trabajar con variable que sean del mismo tipo.

Se seleccionaron las siguientes variables:

- *permanencia*: meses con servicio. Toma valores entre 1 y 72 meses.
- *llamadas_gratuitas_mes*: llamadas gratuitas en el último mes. Toma valores entre 0 y 173. Se asume que los valores están expresados en minutos.
- *inhalámbrico_mes*: Inalámbrico en el último mes, Toma valores entre 0 y 111,95. Se asume que los valores están expresados en minutos.
- *n_pers_hogar*: Número de personas en el hogar. Toma valores entre 1 y 8.
- *edad*: edad en años del usuario que contrata. Los valores van desde 18 a 77 años.

En el archivo *rmd* adjunto se pueden apreciar los boxplots de las variables elegidas. Las variables *inhalámbrico_mes llamadas_gratuitas_mes y n_pers_hogar* presentan algunos valores atípicos.

Métodos de clustering

Los agrupamientos que logremos en este análisis, dependen en gran medida de la distancia seleccionada, la cantidad de grupos deseados y la condición de estandarización o no de las variables. Estas características serán explicadas, a continuación, en el proceso de selección de los algoritmos.

Con el fin de poner en práctica distintos métodos, seleccionamos dos algoritmos, uno jerárquico y otro no jerárquico. Comparamos los agrupamientos logrados y finalmente elegimos el que consideramos más apropiado.

Algoritmo no jerárquico

Como representación de esta técnica, elegimos el algoritmo Kmeans, el cual es muy utilizado debido a su sencillez; aunque el número de grupos (K) es un requerimiento previo a la ejecución el algoritmo.

Para aplicar esta técnica se estandarizaron los datos tomando la media y el desvío estándar del conjunto de datos. Luego se realizaron dos gráficos con los indicadores de Silhouette y Suma de Cuadrados Dentro para poder elegir correctamente el número de clústers antes de ejecutar el algoritmo. En ambos casos, los indicadores coinciden e indican que el número óptimo de clústers es 3. Se pueden ver los gráficos en el archivo adjunto.

Se realiza la clasificación con al algoritmo kmeans, con el parámetro k=3, y se obtiene la siguiente composición de grupos:

```
##
##   1   2   3
## 364 186 450
```

El análisis de componentes principales es un gran aliado a la hora de caracterizar los clúster encontrados, ya que no sólo nos permite ver gráficamente todas las variables en un solo plano, sino que contamos con el valor agregado de saber la relación existente entre las variables. Entonces, utilizamos el gráfico de biplot del análisis de componentes principales y se colorean las observaciones con los clústers asignados cada observacion. De esta forma pueden identificarse e interpretar las características de cada grupo.

Los grupos obtenidos son:

- Grupo 1 (*rosa*): individuos con mayor nro. de personas en el hogar, y a la vez poca permanencia en el servicio y poca edad. También presentan poco uso de llamadas gratuitas y del servicio inalámbrico.
- Grupo 2 (*verde*): individuos con mayor uso de servicio inalámbrico y de llamadas gratuitas. Ambas variables se ven estrechamente relacionadas entre sí, de forma positiva. Sin embargo se ve ausencia de relación entre las de permanencia y edad; y muy poca relación con el nro. de personas en el hogar.
- Grupo 3 (*celeste*): Individuos con mayor permanencia en el servicio, mayor edad (estas dos variables están muy relacionadas entre sí, en forma positiva) y la cantidad de personas en el hogar es poca o reducida (esta variable se relaciona de forma inversa con las primeras dos)

Algoritmos jerárquicos

Como primer paso estandarizamos nuestro conjunto de datos con la técnica de *mínimos y máximos del rango de las variables,* luego, les calculamos la distancia de *manhattan* y la aplicamos a distintas técnicas aglomerativas:

- *Enlace simple*
- *Enlace completo*
- *Promedio entre grupos*
- *Ward*

Nos basamos en el valor del coeficiente de correlación cofenética para la selección del algoritmo óptimo:

Algoritmo	**C. cofenético**
complete linkage (enlace completo)	0.4889359
average (promedio entre grupos)	0,6136286
single linkage(enlace simple)	0,4684972
Ward	0,5183545

Si bien el mayor coeficiente corresponde al método de average, al momento de estudiar el dendrograma, se ve gráficamente que la cantidad óptima de clústers serían 2. Y al implementar la clasificación, los grupos quedan excesivamente desbalanceados:

```
## grupos
##   1   2
## 998   2
```

Por lo tanto se optó por seguir con el método Ward, que fue el segundo mejor valor del coeficiente cofenético.

El dendrograma de Ward, no sólo es más claro visualmente, sino que se ve bastante balanceado. La mayor de las alturas de los enlaces se da para 2 clúster, como optima elección. Al implementar la clasificación, los grupos quedan con la siguiente composición:

```
## grupos
##   1   2
## 521 479
```

Se plasman en el biplot los clústers obtenidos, tal cual lo hicimos con el algoritmo Kmeans. Se identifican los 2 clúster con color y se observan las siguientes características:

- Grupo 1(*rosa*): individuos con mayor permanencia en el servicio, mayor edad y poca cantidad de personas en el hogar. Se agregan también algunos individuos con mayor uso de llamadas gratuitas y del servicio inalámbrico. También se consideran algunos individuos con menor permanencia.
- Grupo 2 (*celeste*): individuos con alto nro. de personas en el hogar, baja permanencia y más jóvenes. Al igual que en el Grupo 1, se agregan algunos individuos con mayor uso de llamadas gratuitas y del servicio inalámbrico.

Conclusión

El algoritmo de Ward, agrupa el conjunto de datos en 2 clúster. Más allá de que, intuitivamente, pareciera ser poca la cantidad de clústers, el problema está en que los grupos formados se superponen mucho entre sí, siendo difícil establecer el límite entre ellos y caracterizarlos de forma concreta.

El algoritmo que mejor agrupa los datos es Kmeans. Las características de los grupos son más identificables y casi no se superponen entre sí. Seguramente esta clusterización será el puntapié para seguir descubriendo más características de los agrupamientos obtenidos, mediantes otras técnicas de exploración.

El presente informe, los archivos de datos y código R, están disponibles en https://github.com/mjfvillanueva/TP-Analisis-Inteligente-de-Datos