Relevancia de variables de entrada

Marta Rodríguez Sampayo
Minería de datos
Universidad Nacional de Educación a Distancia
mrodrigue8255@alumno.uned.es

2 de septiembre de 2020

1. Definición de variables

En base a las sugerencias realizadas por el equipo docente sobre extender el ejemplo XOR de [1], el conjunto de datos empleado para llevar a cabo esta práctica consta de 100 instancias caracterizadas por las siguientes variables (todas ellas booleanas):

- X_1, X_2, X_3 y X_5: variables booleanas aleatorias
- $X_4 = X_2 \oplus X_3$
- Variable clase: $Y = X_1 \oplus X_2 \oplus X_3$

De esta forma, las variables predictoras X_2, X_3 y X_4 son relevantes en sentido débil; X_1 es una variable relevante en sentido fuerte y X_5 es completamente irrelevante.

Según las definiciones de relevancia recogidas en [1]:

- Relevancia fuerte: una variable X_i es relevante en sentido fuerte si existe algún x_i, y e s_i para los cuales $p(X_i = x_i, S_i = s_i) > 0$ tal que

$$p(Y = y|X_i = x_i, S_i = s_i) \neq p(Y = y|S_i = s_i)$$

- Relevancia débil: una variable X_i es relevante en sentido débil si no es relevante en sentido fuerte y existe un subconjunto de variables S_i' de S_i para el cual existe algún x_i, y e s_i' con $p(X_i = x_i, S_i' = s_i') > 0$ tal que

$$p(Y = y|X_i = x_i, S_i' = s_i') \neq p(Y = y|S_i' = s_i')$$

Además, una variable es irrelevante siempre que no sea relevante en sentido fuerte o débil. En términos de un clasificador bayesiano: una variable X es relevante en sentido fuerte si eliminarla resultase en un deterioro del rendimiento del clasificador óptimo; mientras que, X se considerará relevante en sentido débil, si no es relevante en sentido fuerte y existe un subconjunto de variables S, tal que el rendimiento del clasificador en S es peor que en $S \cup \{X\}$.

2. Técnicas empleadas

2.1. Técnicas de filtrado y ranking

Los métodos de filtrado suelen emplearse como una etapa de preprocesado, de forma independiente a los algoritmos de aprendizaje empleados. Las variables se seleccionan en base a sus puntuaciones, obtenidas, habitualmente, a partir de tests estadísticos.

RELIEF Este algoritmo considera una variable como fuertemente relevante si permite distinguir fácilmente entre dos instancias de diferentes clases, definiendo el peso de cada variable según esta lógica [2]. El algoritmo Relief original está limitado a problemas de clasificación binarios y no dispone de un mecanismo para manejar los casos de falta de datos, por lo que se han desarrollado estrategias para mejorarlo como Relief-f o ReliefF[3]. Este último se diferencia del algoritmo original en que calcula el promedio de las contribuciones de las k instancias más cercanas (Relief utiliza $k = 1$).

Información mutua Es una medida de la dependencia mutua entre dos variables, basándose en la información obtenida sobre una variable observando otra variable dada. Es decir, la información mutua mide cuánta información aporta la presencia o ausencia de una variable X a la hora de realizar predicciones correctas sobre Y. En el caso de variables discretas o nominales se estima según:

$$I(i) = \sum_{x_i} \sum_{y} P(X = x_i, Y = y) log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)}$$

Es, por tanto, una medida de dependencia entre la densidad de la variable x_i y la densidad del objetivo y, estimando las probabilidades por conteos de frecuencia.

Si X e Y son independientes, su información mutua es 0. Si X es determinista de Y, entonces la información mutua es la entropía de X, que mide o cuantifica la cantidad de información de una variable.

Chi cuadrado Es un test estadístico aplicado a grupos de variables para evaluar la probabilidad de correlación o asociación entre ellas usando su distribución de frecuencia. Se calcula entre cada variable predictora y la variable clase, seleccionando el número deseado de variables con la mejor puntuación. La puntuación se calcula según la siguiente fórmula:

$$\chi^2 = \frac{(\text{Frecuencia observada} - \text{Frecuencia esperada})^2}{\text{Frecuencia esperada}}$$

donde la frecuencia observada es el número de observaciones de clase y la frecuencia esperada es el número de observaciones de clase esperadas si no hubo relación entre la variable predictora y la variable clase. Es decir, se parte de la hipótesis nula de que las frecuencias observadas se corresponden con las esperadas y según los resultados se acepta o rechaza dicha hipótesis.

2.2. Análisis de componentes principales

Este método, conocido por las siglas PCA del inglés *Principal Component Analysis*, no es exactamente una técnica de selección de variables, si no de transformación. Sin embargo, se suele emplear para reducir la dimensionalidad de grandes conjuntos de datos, procurando reducir el número de variables con la menor pérdida posible de información.

En primer lugar, es necesario realizar una estandarización de los datos, esto es una conversión de las variables a una escala común para evitar problemas derivados de la sensibilidad de PCA a las varianzas de las variables iniciales. Matemáticamente, este proceso se puede definir mediante la siguiente fórmula:

$$z = \frac{x - \bar{x}}{\sigma}$$

donde $\bar{x}$ y σ son la media y la desviación típica de las muestras.

Para analizar la relación existente entre las variables en base a la correlación entre ellas se construye una matriz $p \times p$ (siendo p el número de dimensiones) simétrica denominada matriz de covarianza. Por ejemplo, en el caso de $p = 3$:

$$\begin{pmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{pmatrix}$$

En cuanto a sus valores, se debe tener en cuenta que la covarianza es conmutativa y la covarianza de una variable consigo misma es su varianza. En caso de que el

valor obtenido sea positivo, ambas variables están correlacionadas, si es negativo, la correlación es inversa.

Las componentes principales se definen como variables construidas como combinaciones lineales de las variables iniciales, de forma que estas nuevas variables no están correlacionadas y se preserve la información de forma comprimida. Si se parte de 5 variables, se obtendrán 5 componentes principales, pero estas estarán ordenadas según la información que contengan de mayor a menor (partiendo de la hipótesis de que máxima varianza es equivalente a máxima información útil). Escogiendo las componentes principales según este criterio permite reducir la dimensionalidad de los datos. Para obtener estas componentes principales, se calculan los autovectores de la matriz de covarianza, es decir, las direcciones de los ejes de máxima varianza y los autovectores que se corresponden con el valor de la varianza proporcionada por cada componente principal.

Finalmente, se emplea el vector de características (matriz formada por los autovectores de las componentes que se decide conservar) para reorientar los datos según los ejes que representan las componentes principales. Por ejemplo, multiplicando el conjunto de datos original traspuesto por la traspuesta del vector de características.

2.3. Técnica de envoltura

Las técnicas de envoltura se emplean para puntuar subconjuntos de variables según su poder de predicción, utilizando la máquina de aprendizaje en cuestión como una caja negra [4]. Para poner en práctica este tipo de técnicas es necesario definir el mecanismo de búsqueda de los posibles subconjuntos de variables (algoritmos genéticos, *best-first*, *branch-and-bound*, etc.), el método de evaluación de la predicción (validación cruzada o conjunto de validación) y el predictor a utilizar (árboles de decisión, *naïve Bayes*, SVM, etc.).

Las estrategias de búsqueda codiciosas (*greedy*) suelen resultar eficientes y robustas contra el fénomeno del sobreajuste y se dividen en: *forward selection*, las variables se añaden a subconjuntos cada vez mayores; y *backward elimination*, se comienza con un conjunto de todas las variables y se eliminan progresivamente las menos prometedoras. Dentro de esta última categoría se encuentra la técnica empleada para el desarrollo de esta práctica, descrita a continuación.

Recursive Feature Elimination (RFE) El subconjunto final de variables se obtiene ajustando el modelo de aprendizaje, ordenando las variables por importancia

(empleando la precisión obtenida), descartando aquellas menos importantes y volviendo a ajustar el modelo. Este proceso se repite de forma recursiva hasta alcanzar un número de variables previamente especificado [5]. Este método de selección de características es uno de los más populares y, por tanto, ampliamente empleado.

3. Selección de técnicas de filtrado

El problema planteado consiste en una tarea de clasificación sobre el conjunto de datos definido en la Sección 1. La variable clase y las variables predictoras definidas como booleanas son variables categóricas, es decir, toman un número limitado y fijo de valores, habitualmente asociados a una etiqueta. Los métodos de filtrado+ranking más empleados en este tipo de escenarios son los test estadísticos χ^2 e información mutua y el algoritmo Relief.

En el caso de χ^2 con este tipo de variables, resulta sencillo obtener la tabla de contingencia con las frecuencias observadas y compararlas con las esperadas. De esta forma, los valores altos de χ^2 indican que los valores observados y esperados difieren considerablemente (variables dependientes) y al contrario si el valor es bajo (variables independientes), dando una medida de la distancia entre estas frecuencias. El cálculo de la información mutua funciona de forma similar, siendo preferible escoger las variables con un valor mayor.

En el conjunto de datos utilizado, la variable clase (Y) depende en diferente medida de algunas de las variables predictoras (X_1 y X_2, X_3 o X_4), no obstante esta relación no es lineal ni directa, por lo que los resultados obtenidos en cuanto a la relevancia probablemente no se correspondan con los reales de forma exacta.

Por otra parte, una de las técnicas más populares para estimar la calidad de las variables en relación con otras variables es Relief. Además, este algoritmo no tiene en cuenta la distribución de la población o el tamaño de la muestra, lo cual puede resultar beneficioso en el contexto de esta actividad. AL tratarse de un problema de clasificación binaria con una dimensionalidad reducida, este método de filtrado, aunque básico, debería conseguir identificar la relevancia de las variables empleadas.

Por tanto, se emplean dos técnicas de filtrado que, aunque acordes al tipo de datos, se espera que obtengan resultados mediocres al identificar la dependencia de las variables y una tercera que resulta, a priori, apropiada para este conjunto de variables. El objetivo es comparar los resultados de las tres técnicas para contrastar las hipótesis de idoneidad formuladas en la literatura consultada y resumidas

anteriormente, en base al conocimiento previo que se posee sobre el conjunto de variables utilizado.

4. Experimentos

4.1. Configuración

- χ^2 e Información mutua: se evalúan las 5 variables para observar las puntuaciones obtenidas en cada caso según el estadístico correspondiente.
- ReliefF: se varía el valor de k, probando los valores 50, 40, 30, 20 y 10. Con $k = 10$, valor indicado como suficiente en la documentación consultada, se asigna la relevancia adecuada a las variables, con $k > 10$ las diferencias entre las puntuaciones asignadas son más evidentes .
- Eliminación recursiva de características: se emplean 3 algoritmos de clasificación diferentes para comprobar la eficacia de esta técnica (Bosque aleatorio, regresión logística y árbol de decisión) con sus valores por defecto para no favorecer el rendimiento de ninguno. Además se elige seleccionar 2 variables del conjunto completo. En los resultados se presenta el orden que adjudica este algoritmo a las 5 variables.
- Análisis de componentes principales: se dejan los valores por defecto, de forma que se mantienen todas los componentes para mostrar la gráfica del ratio de varianza explicada por componente. Se debe tener en cuenta que más que una técnica de selección de variables, PCA es un método de reducción de dimensionalidad del conjunto de datos, de forma que no se muestra un *ranking* de las variables del conjunto de datos original.

4.2. Resultados

Puntuaciones / Variables	Chi2	Información mutua	ReliefF ($k = 10$)
X_1	0.008637	0	**1000**
X_2	0.496962	**0.010734**	692
X_3	0.158202	0.005279	692
X_4	0.189155	0	692
X_5	**0.637573**	0	240

Tabla 1: Resultados de los métodos de filtrado.

Clasificador / Variables	Bosque aleatorio	Regresión logística	Árbol de decisión
X_1	**1**	4	**1**
X_2	3	**1**	4
X_3	2	3	3
X_4	**1**	2	**1**
X_5	4	**1**	2

Tabla 2: Resultados del método de envoltura RFE para distintos métodos de clasificación.

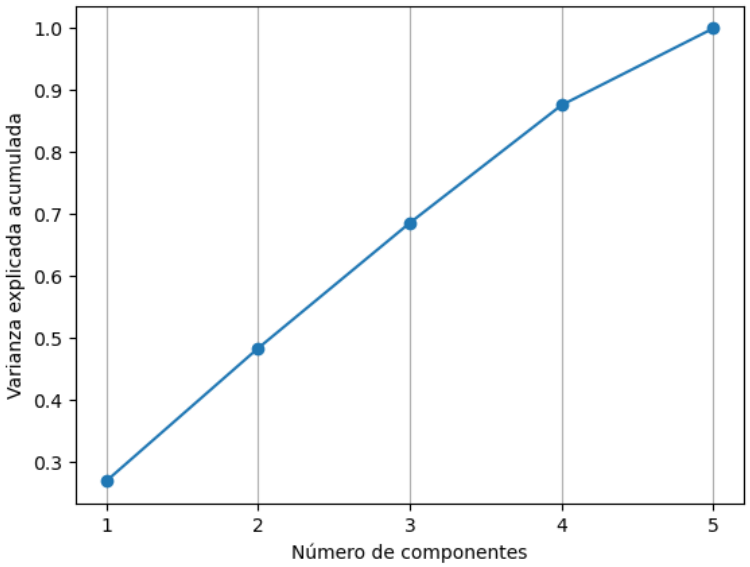

Figura 1: Varianza explicada PCA: [0.27 0.214 0.202 0.19 0.124]

5. Discusión y conclusiones

En cuanto a los métodos de filtrado, ReliefF es el que puntúa las variables de forma acorde a la relevancia asignada al generar el conjunto de datos. Como se observa en la última columna de la Tabla 1, la variable X_1 tiene mayor puntuación, seguida de X_2, X_3 y $X4$ que tienen la misma y, finalmente, la variable generada aleatoriamente y sin ninguna relación con la variable clase X_5. Estos resultados confirman la hipótesis de la idoneidad de esta técnica para el conjunto de datos, ya que las variables que dan valores diferentes a instancias vecinas de la misma clase

se penalizan y a aquellas que dan diferentes valores a vecinos de diferentes clases se les asigna mayor puntuación. De esta forma, se detecta la relación directa entre la variable clase Y y la variable X_1, con mayor puntuación que la relación de las variables X_2, X_3 y X_4, ya que esta última es una combinación de las dos primeras. También cabe mencionar que al aumentar el número de vecinos la diferencia de puntuación entre X_4, X_2 y X_3 aumenta, identificándose la primera como segunda variable más relevante, probablemente debido a que la relación de cada instancia de esta variable con la clase es más evidente.

Tanto χ^2 como información mutua adjudican puntuaciones bajas a todas las variables predictoras, siendo incapaces de valorar su relevancia de forma correcta. El motivo de estos resultados, presentados en las dos primeras columnas de la Tabla 1, es que, en el caso del estadístico χ^2 se calcula la probabilidad de que cada una de las predictoras sea independiente de la variable clase de forma individual, sin tener en cuenta las relaciones entre estas y, por tanto, no identificando la relevancia real. De forma análoga, la información mutua mide la dependencia entre dos variables, mientras que la variable clase de este conjunto de datos está formada por la asociación de varias variables, dependientes al mismo tiempo entre ellas. Aunque ambas medidas se adecúan a los *datasets* con variables predictoras y clase categóricas, como es el caso, no detectan el tipo de relación existente entre estas variables, puntuándolas de forma prácticamente aleatoria.

Por otro lado, los resultados obtenidos para el método de envoltura RFE, mostrados en la Tabla 2 también son bastante acertados teniendo en cuenta las características de las variables predictoras, aunque estos dependen del algoritmo de aprendizaje empleado. Ya que en la implementación se especifica que se conserven 2 de las variables predictoras, estas tienen un valor 1 en el ránking, considerándose las de mayor importancia. En el caso de los clasificadores basados en Bosque aleatorio y Árbol de decisión, las dos variables seleccionadas son X_1 y X_4, coincidiendo con la relevancia predefinida. Esto se debe a la relación entre estos dos tipos de clasificadores y al cálculo que realizan de la importancia de cada variable basada en la impureza de la misma, es decir, miden la frecuencia con la que una instancia escogida aleatoriamente del conjunto se etiquetaría incorrectamente si se etiquetase al azar de acuerdo con la distribución de etiquetas en el subconjunto. Mientras que el algoritmo de Regresión logística no obtiene buenos resultados ya que la relación entre la variable clase y las predictoras no es lineal, dificultando la asignación de importancia.

Finalmente, la Figura 1 muestra la cantidad de varianza explicada acumulada según aumenta el número de componentes principales generadas mediante la técnica

PCA. Puesto que PCA procura incluir en la primera componente la mayor cantidad posible de información, esta debería explicar la mayor parte de la varianza, de forma que con las primeras componentes se explicase la mayor parte de la varianza y se pudiese prescindir de las últimas. Sin embargo, como se puede observar, son necesarias mínimo 4 componentes para explicar aproximadamente el 88 % de la varianza, se puede entender la varianza explicada por una componente como la cantidad de información que contiene, de forma que no se mantendría ni el 90 % de información reduciendo una dimensión el conjunto de datos. La generación de las componentes se realiza a través de combinaciones lineales de las variables, empleando la covarianza como medida de la relación entre ellas, de forma que la dependencia no lineal no se ve reflejada al calcular esta correlación.

Referencias

[1] R. Kohavi and G. John, "Wrappers for feature selection," *Artificial Intelligence - AI*, vol. 1, 01 1997.

[2] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proceedings of the Tenth National Conference on Artificial Intelligence*, ser. AAAI'92. AAAI Press, 1992, p. 129–134.

[3] I. Kononenko, E. Šimec, and M. Robnik-Sikonja, "Overcoming the myopia of inductive learning algorithms with relieff," *Applied Intelligence*, vol. 7, pp. 39–55, 01 1997.

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 1157–1182, Mar. 2003.

[5] K. Johnson and M. Kuhn, *Recursive Feature Elimination.* Chapman and Hall/CRC, 2019. [Online]. Available: https://bookdown.org/max/FES/