# Métodos no supervisados y detección de anomalías: medidas de validación

UCO ONLINE

# Medidas de validación

## 1. Introducción

En las lecciones anteriores se han estudiado diferentes algoritmos para resolver el problema de agrupamiento. Cada algoritmo tiene su propio funcionamiento y parámetros, por lo que es habitual que vayamos a obtener diferentes agrupamientos según el algoritmo y parámetros elegidos.
Llegados a este punto, cabe preguntarse cómo podemos comprobar que la asignación de grupos devuelta por un algoritmo es mejor que la obtenida por otro. También, si la estructura de grupos obtenida es representativa. Estos son los objetivos de la *validación de grupos*, una rama de estudio dentro del agrupamiento que proporciona métodos cuantitativos y cualitativos para evaluar la calidad del agrupamiento.
Dentro de las técnicas de validación, existe una clasificación de criterios en función de la información de la que se dispone a la hora de evaluar el resultado de un método de agrupamiento:

- **Criterios de validación externa**. Dado un conjunto de datos X y una asignación de grupos, C, obtenida por medio de un algoritmo ejecutado sobre X, los criterios de validación externa comparan C con información conocida a priori sobre la estructura de X. Por ejemplo, si el conjunto de datos estaba previamente etiquetado, podemos cuantificar cómo de bien la asignación de datos a grupos refleja la distribución de clases.

- **Criterios de validación interna**. Dados X y C, definidos como en caso anterior, los criterios de validación interna no utilizan información adicional sobre X para evaluar la calidad de C. En este caso, es la propia información sobre la proximidad de los puntos la que debe utilizarse para determinar la calidad de cada grupo obtenido, lo cual permite conocer propiedades como su densidad o tamaño.

- **Criterios de validación relativos**. Dada una agrupación C, los criterios de validación relativos permiten comparar C con otras estructuras (C', C'', C''') obtenidas tras la aplicación de otro algoritmo o de otra configuración de parámetros de un mismo algoritmo. A modo de ejemplo, un criterio relativo serviría para comparar los distintos agrupamientos obtenidos al variar el parámetro *k* del método k-medias.

Tanto los criterios de validación externa como interna se relacionan con los métodos estadísticos de prueba de hipótesis (*hypothesis tests*), pues la forma de validación en agrupamiento se puede plantear según la siguiente metodología: Dado un conjunto de datos X, la hipótesis nula establece que no existe una estructura en X, o que dicha estructura es aleatoria. Con un método de muestreo estadístico, como el análisis Monte Carlo o Bootsrapping, se puede estudiar la distribución de los datos que cumpliría dicha hipótesis. En base a dicha distribución, se obtendría un umbral que, dado un nivel de significancia estadística, indica cuándo se cumple dicha suposición. Este umbral es el que se utiliza para comparar frente a la medida de validación en cuestión, de forma que se determina si la medida de validación contradice o no la hipótesis de partida. La hipótesis nula antes expuesta es general, pero puede descomponerse en tres hipótesis más concretas:

- Hipótesis de posición aleatoria. Establece que todas las posiciones de los puntos del conjunto de datos son igualmente probables.
- Hipótesis de grafo aleatorio. Establece que todas las matrices de proximidad que ordenan los datos son igualmente probables.
- Hipótesis de etiquetas aleatorias. Establece que todas las etiquetas para los puntos son igualmente probables.

Este marco estadístico implica conocimientos más avanzados que no son el objetivo principal de la lección, aunque es conveniente conocerlo. Por tanto, las siguientes secciones se centran en describir los principales criterios de validación de cada tipo: externa, interna y relativos.

## 2. Criterios de validación externa

Los criterios de validación externa necesitan una partición "real" (*ground truth*) que sirva de referencia para comprobar si los puntos del conjunto de datos han sido asignados al grupo correcto o no. Dicha partición de referencia, a la que llamamos $P$, está prefijada y es conocida a priori. Dentro de los criterios de validación externa, se pueden diferenciar tres grupos de medidas: basadas en correspondencia (*matching*), basadas en comparaciones por pares (*pairwise*) y basadas en entropía. Para comprender mejor el significado de estas medidas, se utilizará un ejemplo con la distribución de puntos que se muestra en la Figura 1:

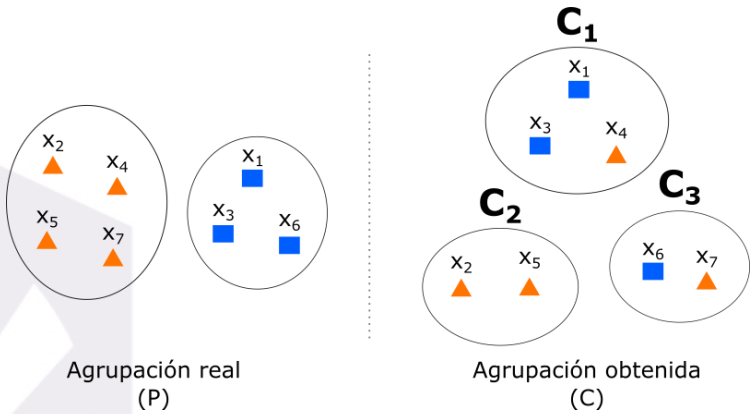


*Figura 1. Ejemplo de agrupación real y aproximada*

## 2.1. Medidas basadas en correspondencia

Este primer grupo se compone de medidas que analizan si, dentro de cada grupo, los puntos asignados concuerdan entre sí con respecto a la partición de referencia. A continuación, se explican dos medidas basadas en esta idea: *Purity* y *Maximum matching*.

### Purity

Esta medida evalúa en qué grado cada grupo de la partición obtenida contiene puntos de un único grupo de la partición de referencia. Una vez obtenido el grado de "pureza" de cada grupo, la "pureza" de la agrupación se obtiene como la media entre todos los grupos. Matemáticamente, la pureza de un grupo $i$ se calcula como:

$$Purity_i = \frac{1}{n_i}\max\{n_{ij}\} \quad j = [1,k] \tag{1}$$

Donde $n_{i,j}$ representa el número de puntos asignados al grupo $i$ que pertenecen al mismo grupo $j$ en la partición de referencia (formada por $k$ grupos). A partir de esta ecuación, la pureza de una agrupación formada por $c$ grupos se obtiene como:

$$Purity = \sum_{i=1}^{c}\frac{n_i}{n}Purity_i = \frac{1}{n}\sum_{i=1}^{c}\max\{n_{ij}\} \quad j = [1,k] \tag{2}$$

Si la partición obtenida contiene el mismo número de grupos que la partición de referencia y todos los puntos se han agrupado correctamente (esto es, cada grupo contiene solo puntos que comparten una etiqueta), entonces el valor de pureza alcanza el valor 1. Partiendo del ejemplo de la Figura 1, contabilizamos el número de puntos para cada etiqueta (azul o naranja) en cada uno de los tres grupos de la agrupación aproximada:

| Grupo | Etiqueta azul | Etiqueta naranja | Tamaño | Pureza |
|---|---|---|---|---|
| $C_1$ | 2 | 1 | 3 | 2/3=0.67 |
| $C_2$ | 0 | 2 | 2 | 2/2=1.00 |
| $C_3$ | 1 | 1 | 2 | 1/2=0.50 |

La pureza total será: $Purity = \frac{2+2+1}{7} = 0.71$

Un inconveniente de esta medida es que no captura adecuadamente el hecho de que un grupo de puntos "real" esté representado en un único grupo de la partición obtenida. Por ejemplo, si $C_1$ y $C_3$ solo contuvieran puntos de etiqueta azul, cada uno de ellos aporta un valor de pureza 1 al total. Sin embargo, una mejor agrupación sería aquella que uniera a todos los puntos de etiqueta azul en un único grupo. Para tener en cuenta esta posibilidad, se puede utilizar la siguiente medida.

### Maximum matching

Esta medida sigue la misma filosofía que la pureza, pero solo permite que un grupo de la partición de referencia sea "asignado" a un grupo de la partición obtenida por medio del algoritmo de agrupamiento. Matemáticamente, se formula como sigue:

$$Maximum\ matching = max\left\{\frac{w(M)}{n}\right\}, w(M) = \sum_{e-g \in M} n_{ij} \quad (3)$$

Donde $w(M)$ establece la correspondencia entre cada etiqueta real y un único grupo, de forma que la combinación de etiqueta-grupo (e-g) que obtenga el mejor "emparejamiento" es la que determina el valor de la medida. En la siguiente tabla se plantea un ejemplo hipotético con tres etiquetas y tres grupos:

| Grupo | Etiqueta 1 | Etiqueta 2 | Etiqueta 3 | Total |
|---|---|---|---|---|
| $C_1$ | 0 | 20 | 10 | **30** |
| $C_2$ | 0 | 10 | 5 | **15** |
| $C_3$ | 30 | 0 | 0 | **30** |
| **Total** | **30** | **30** | **15** | **75** |

Claramente, el grupo $C_3$ acierta al agrupar puntos de la etiqueta 1. Sin embargo, las otras dos etiquetas están distribuidas entre los grupos $C_1$ y $C_2$. Podemos hacer dos correspondencias: 1) $C_1$ con la etiqueta 2 y $C_3$ con la etiqueta 3; o 2) $C_1$ con la etiqueta 2 y $C_3$ con la etiqueta 3.

Caso 1: $w(M) = \sum_{e-g \in M} n_{ij} = 20 + 5 + 30 = 55$
Caso 2: $w(M) = \sum_{e-g \in M} n_{ij} = 10 + 10 + 30 = 50$

Por tanto: $Maximum\ matching = \max\left(\frac{55}{75}, \frac{50}{75}\right) = 0.73$

## 2.2. Medidas basadas en comparaciones por pares

El segundo grupo de medidas se basa en comparar la asignación de grupos para cada par de puntos que componen el conjunto de datos. Dada la parición obtenida por un método de agrupamiento ($C$) y $P$, pueden darse las siguientes circunstancias entre dos puntos $x_i$ y $x_j$:

1. Caso 1: $x_i$ y $x_j$ son asignados al mismo grupo en $C$ y además pertenecen a la misma categoría en $P$. El total de pares de puntos en este caso se denota $a$.
2. Caso 2: $x_i$ y $x_j$ son asignados al mismo grupo en $C$, pero pertenecen a distintas categorías en $P$. El total de pares de puntos en este caso se denota $b$.
3. Caso 3: $x_i$ y $x_j$ son asignados a distintos grupos en $C$, pero pertenecen a la misma categoría en $P$. El total de pares de puntos en este caso se denota $c$.
4. Caso 4: $x_i$ y $x_j$ son asignados a distintos grupos en $C$ y pertenecen a distintas categorías en $P$. El total de pares de puntos en este caso se denota $d$.

Dado un conjunto de datos con N instancias (puntos), el total de pares evaluados es: $M = \frac{N(N-1)}{2} = a + b + c + d$. Las siguientes medidas de validación, también llamados índices externos, se calculan en base a estos valores.

Para comprender mejor el cálculo de estos índices, utilizaremos un ejemplo con 7 puntos. En la Figura 1 se muestran las agrupaciones $C$ y $P$. En la Tabla 1 se contabiliza cuántos pares de puntos se asocian a cada uno de los cuatro pasos descritos anteriormente.

| Caso | Parejas de puntos | Total |
|---|---|---|
| 1 | $(x_1, x_3), (x_2, x_5)$ | a=2 |
| 2 | $(x_1, x_4), (x_3, x_4), (x_6, x_7)$ | b=3 |
| 3 | $(x_1, x_6), (x_2, x_4), (x_2, x_7), (x_3, x_6), (x_4, x_5), (x_4, x_7), (x_5, x_7)$ | c=7 |
| 4 | $(x_1, x_2), (x_1, x_5), (x_1, x_7), (x_2, x_3), (x_2, x_6), (x_3, x_5),$ $(x_3, x_7), (x_4, x_6), (x_5, x_6)$ | d=9 |

## Índice de Rand

El índice de Rand, propuesto por William D. Rand en 1971, es una medida de similitud de conjuntos que se puede aplicar al problema de agrupamiento. Está relacionado con la precisión en la asignación de puntos a grupos, pues se focaliza en los casos de acierto (casos 1 y 4). Se formula como el cociente entre el número de pares en los que ambas particiones concuerdan (a y d) entre el total de pares comparados (M):

$$R = (a + d)/M \tag{4}$$

Es una medida a maximizar que varía entre 0 (cuando las particiones no coinciden en la asignación de ningún par de puntos) y 1 (ambas particiones son iguales). Por tanto, representa la frecuencia con la que coinciden las dos agrupaciones. Otra forma de interpretarlo es como la probabilidad de que las dos agrupaciones coincidan dada una pareja de puntos aleatoria. Para el ejemplo propuesto, el resultado sería: $R = \frac{a+d}{M} = \frac{2+9}{21} = 0.5238$

Existe una variante del índice, llamada índice de Rand ajustado, que corrige un comportamiento no deseado en esta medida. Ante una asignación totalmente aleatoria de etiquetas a los puntos, el índice Rand no garantiza que vaya a obtenerse un valor próximo a cero, como cabría pensar. El índice ajustado modifica este comportamiento utilizando internamente como base un modelo aleatorio.

## Coeficiente de Jaccard

El coeficiente de Jaccard (por Paul Jaccard), que ya fue estudiado como medida de similitud en la primera semana del curso, puede utilizarse también para comparar dos agrupaciones. A diferencia del índice de Rand, el índice de Jaccard no considera los aciertos del caso 4 (valor d), sino que se centra en los aciertos del caso 1 (valor a). Por tanto, se calcula como sigue:

$$J = a/(a + b + c) \tag{5}$$

Esta medida también varía entre 0 y 1, siendo mejor cuanto mayor sea el valor obtenido. Dadas las agrupaciones del ejemplo, el valor del índice es: $J = \frac{a}{a+b+c} = \frac{2}{2+3+7} = 0.1667$

### Índice de Fowlkes-Mallows

El índice Fowlkes-Mallows fue propuesto por E.B. Fowlkes y C.L. Mallows en 1983 como una medida para comparar agrupamientos jerárquicos. No obstante, también se puede utilizar para comparar dos agrupaciones cualesquiera, donde una sea la agrupación real. Su formulación se muestra a continuación:

$$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} \tag{6}$$

Al igual que las medidas anteriores, su resultado varía entre 0 y 1. El valor 0 es el peor caso posible, donde ningún par de puntos de la misma etiqueta ha sido agrupado junto (a=0). El mejor valor posible es 1, que indica que todos los puntos que comparten etiqueta han sido agrupados juntos (b=0 y c=0). Para el ejemplo propuesto:

$$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} = \sqrt{\frac{2}{2+3}\frac{2}{2+7}} = 0.2981$$

Este índice es considerado más preciso que el índice de Rand cuando se están comparando dos particiones que tienen poca relación entre sí y el conjunto de datos crece en tamaño. A medida que más puntos hay en el conjunto de datos, el índice Fowlkes-Mallows tiende a cero, mientras que el índice de Rand se aproxima a 1, dando una falsa sensación de similitud que es en realidad debida a un valor muy alto de *d*. De forma similar, si el conjunto de datos presenta ruido, el índice de Fowlkes-Mallows se verá menos afectado y, de hecho, tiende a reducir su valor a medida que la presencia de ruido aumenta.

## 2.3. Medidas basadas en entropía

La entropía es un concepto propio de la teoría de la información, pero muy utilizado en aprendizaje automático. Este concepto trata de medir la cantidad de "orden" que hay en los datos, en nuestro caso, en una partición obtenida por medio de un algoritmo de agrupamiento. La *entropía* se calcula en base a la probabilidad de que los puntos pertenezcan a cada grupo:

$$H(C) = -\sum_{i=1}^{c} p_{C_i} log\, p_{C_i}\,,\ p_{C_i} = \frac{n_i}{n} \tag{7}$$

No obstante, esta expresión se aplica a una partición, sin tener en consideración la partición de referencia. Para considerarla una medida de validación externa, la entropía se formula como una entropía condicional con respecto a la partición de referencia. Si $C$ es la partición obtenida por el algoritmo y $P$ es la partición real o de referencia (con $k$ grupos), entonces la entropía de $C$ con respecto a un grupo en $P$ se define como:

$$H(C|P_i) = -\sum_{j=1}^{k} \left(\frac{n_{ij}}{n_i}\right) \log \left(\frac{n_{ij}}{n_i}\right) \quad (8)$$

Y la entropía condicional de $C$ con respecto a toda la partición $P$ se calcula como:

$$H(C|P) = -\sum_{i=1}^{c} \left(\frac{n_{ij}}{n}\right) H(C|P_i) \quad (9)$$

La entropía se verá afectada por cómo se distribuyan los puntos que pertenecen a una misma categoría (mismo grupo en $P$) en los grupos de la partición $C$. A mayor dispersión de los datos de una misma categoría, mayor será la entropía condicional. Si la partición $C$ se corresponde perfectamente con la partición de referencia $P$, entonces la entropía será cero (valor óptimo).

## 3. Criterios de validación interna

Las medidas de validación interna se fundamentan en las dos características que definen el problema de agrupamiento. La primera es que los puntos asignados a un mismo grupo sean próximos entre sí, lo que denominaremos *cohesión*. La segunda es que los puntos asignados en distintos grupos estén alejados, lo que denominaremos *separación*. Una medida de cohesión es WSS (*within cluster sum of squares*), la cual recuerda a la medida de error SSE. Tal y como se muestra en la siguiente ecuación, consiste en sumar las distancias de cada punto al centroide, $m_i$, del grupo al que ha sido asignado:

$$WSS = \sum_{i=1}^{c} \sum_{x \in C_i} (x - m_i)^2 \quad (10)$$

Una medida de separación es BSS (*between cluster sum of squares*), que sigue una filosofía similar pero se calcula entre los centroides de cada grupo:

$$BSS = \sum_{i=1}^{c} |C_i| (m - m_i)^2 \quad (11)$$

Donde $|C_i|$ representa el tamaño del grupo $i$, y $m$ es el punto medio en el conjunto de datos.

WSS y BSS se centran cada una en un criterio deseable. Por el contrario, existe una medida que tiene en cuenta ambos aspectos a la vez: *Silhouette Coefficient*. Esta medida se basa en calcular, para cada punto del conjunto de datos, las siguientes dos distancias:

- Distancia "a": La distancia media del punto a todos los demás puntos con lo que comparte grupo.
- Distancia "b": La distancia media a todos los puntos del grupo más cercano.

En base a estas distancias, el coeficiente de silueta para un punto es:

$$s(x_i) = \frac{b-a}{\max(b,a)} \tag{12}$$

El coeficiente de silueta para la partición completa se obtiene como la media del coeficiente para todos los puntos, donde valores próximos a 1 indican un buen agrupamiento:

$$SC = \frac{1}{n}\sum_{i=1}^{N} s(x_i) \tag{13}$$

## 4. Criterios de validación relativos

Los criterios de validación externos son utilizados con el propósito de comparar dos agrupaciones obtenidas por algoritmos diferentes, o por un mismo algoritmo con distintos parámetros. Un ejemplo que ya se mencionó anteriormente es el estudio de $k$ para el algoritmo k-medias. Una primera posibilidad es calcular una medida de validación interna o externa para cada resultado y, sabiendo en qué rangos se mueven estas medidas, comparar sus valores. Es habitual utilizar el coeficiente de silueta con este fin. Existen otras dos medidas de tipo interno que consideran el compromiso entre cohesión y separación, y que han demostrado ser eficaces para la comparación de resultados: el índice de *Calinski-Harabasz* y el índice de *Davies-Bouldin*.
El primero se calcula en base al cociente entre la dispersión entre grupos y la dispersión dentro de cada grupo. La dispersión se define como la suma de las distancias entre puntos al cuadrado, requiriendo computar una matriz de dispersión en cada caso (intra e inter-grupo). Valores más altos del índice son deseables, pues indican que los grupos son densos y están muy alejados entre sí.
El segundo se obtiene como la similitud media de cada grupo con respecto al más cercano, donde la similitud se basa en las distancias entre puntos y centroides. En este caso, es una medida a minimizar cuyo mejor valor es 0.

## Referencias

C. C. Aggarwal, C. K. Reddy (eds.). "Data Clustering: Algorithms and Applications". Chapman & Hall / CRC Press, 1ª edición, 652 páginas. 2014.
T. Hastie, R. Tibshirani, J. Friedman. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Springer Series in Statistics, 2ª edición, 745 páginas. 2017.
G. James, D. Witten, R. Tibshirani, T. Hastie. "An Introduction to Statistical Learning with Applications in R". Springer Texts in Statistics, 1ª edición (7ª impresión), 426 páginas. 2017. Disponible en: https://www.statlearning.com/
A. Kassambara. "Practical Guide to Cluster Analysis in R". STHDA, 187 páginas. 2017.
S. Shalev-Shwartz, S. Ben-David. "Understanding Machine Learning: From Theory to Algorithms". Cambridge University Press, 1ª edición, 449 páginas. 2014.
R. Xu, D.C. Wunsch II. "Clustering". Wiley/IEEE Press, 1ª edición, 363 páginas. 2009.

## Referencias en línea

Medidas de rendimiento para agrupación en scikit-learn v1.0.1: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation (Último acceso: 04/12/2021)