UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

FACULTAD DE INGENIERIA

Comparación de particiones en aprendizaje automático no supervisado

Meliza González

Programa de Posgrado en Ingeniería Matemática
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay
Noviembre de 2018

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

FACULTAD DE INGENIERIA

Comparación de particiones en aprendizaje automático no supervisado

Meliza González

Tesis de Maestría presentada al Programa de Posgrado en Ingeniería Matemática, Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Magister en Ingeniería Matemática.

Director de tesis:
Ph.D. Prof. Mathias Bourel

Codirector:
Ph.D. Prof. Badih Ghattas

Director académico:
Ph.D. Prof. Franco Robledo

Montevideo – Uruguay
Noviembre de 2018

González, Meliza

Comparación de particiones en aprendizaje automático no supervisado / Meliza González. - Montevideo: Universidad de la República, Facultad de Ingeniería, 2018.

VIII, 56 p. 29, 7cm.

Director de tesis:

Mathias Bourel

Codirector:

Badih Ghattas

Director académico:

Franco Robledo

Tesis de Maestría – Universidad de la República, Programa de Ingeniería Matemática, 2018.

Referencias bibliográficas: p. 53 – 55.

1. Comparación de particiones, 2. Índices de validación externa, 3. Error de clasificación, 4. Análisis de cluster, 5. Aprendizaje no supervisado. I. Bourel, Mathias *et al.* II. Universidad de la República, Programa de Posgrado en Ingeniería Matemática. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Ph.D. Prof. Marcelo Fiori

Ph.D. Prof. Gustavo Guerberoff

Ph.D. Prof. José Rafael León

Montevideo – Uruguay
Noviembre de 2018

RESUMEN

En esta tesis se presenta un estudio sobre índices de comparación de particiones de un mismo conjunto de datos, utilizados para la evaluación externa de los resultados de métodos de clasificación no supervisada. Se elabora un estado del arte en cuanto a los índices existentes y sus propiedades y se calculan algunos de los más conocidos sobre datos simulados a modo de ejemplo.

Este trabajo se centra en el índice Mínimo Error de Clasificación (MCE, por su sigla en inglés), medida basada en la tabla de contingencia de dos particiones. Se estudia y profundiza sobre sus propiedades y en especial su distribución. Se establece la expresión analítica de la función de distribución teórica para el caso de la comparación de dos particiones independientes, con dos clases balanceadas. Algunas propiedades demostradas pudieron extenderse para el caso de tres clases y para el caso general. También se estudian las propiedades de la distribución empírica sobre datos simulados, variando algunos parámetros experimentales, y mostramos una aplicación sobre un conjunto de datos supervisados reales de imágenes de dígitos escritos a manos, conocido como MNIST. En este último caso, planteamos el problema de clasificación no supervisada y la validación externa de los resultados basada en nuestro índice se realiza comparándolos con la verdadera etiqueta de los datos. Los resultados del MCE se comparan con otros índices de validación externa mediante correlaciones y en distintos escenarios.

Finalmente, a partir de la distribución del índice, se diseña un test de hipótesis que permite contrastar si dos particiones son independientes. El desempeño de la prueba se evalúa calculando los errores de tipo I y II obtenidos con datos simulados artificialmente.

Palabras claves:
Comparación de particiones, Índices de validación externa, Error de clasificación, Análisis de cluster, Aprendizaje no supervisado.

ABSTRACT

This paper carries out a study on comparison indexes over the same data set, used for external evaluation of the results of unsupervised classification methods. It presents a state of the art review of indexes and their properties and calculates some of the better known ones over simulated data as examples.

Throughout the text, we focus on the definition of the Minimum Classification Error (MCE) index, which is based on the contingency table of two data partitions. Its properties and, especially, distribution function are analyzed. The analytic expression of the theoretical distribution function is derived for the case of two independent partitions, with two balanced classes. Some of the demonstrated properties can be extended for three and more classes. We also study the properties of the empirical distribution on simulated data, by manipulating some experimental parameters. An application is shown on a set of real supervised data of images of handwritten digits, known as MNIST. We pose the problem of unsupervised classification and carry out an external validation of the results obtained by the developed index by comparing them with the true data label.The results of the MCE are compared with other validation indexes through correlations.

Finally, based on the distribution function of the index, we design a hypothesis test for determining whether two partitions are independent. The performance of the test is evaluated by calculating Type I and II errors obtained with artificially simulated data.

Keywords:
Comparing partitions, External validation measures, Classification error, Clusterings, Unsupervised learning.

Tabla de contenidos

Introducción

La evaluación de los resultados de un análisis de *cluster* se compone básicamente de dos etapas: la validación interna y externa de los grupos encontrados. La validación interna mide la calidad del agrupamiento, basándose en el cálculo de las propiedades de la partición resultante como la cohesión de los *clusters* y la separación entre los mismos, utilizando información intrínseca de los datos. La validación externa, a diferencia de la anterior, utiliza información externa dada de antemano. En este trabajo nos focalizaremos en la validación externa y en los índices utilizados para su valoración.

Los índices de validación externa o comparación de particiones son activamente utilizados para encontrar una buena solución a la clasificación. Se utilizan en general para ver qué tan bien el algoritmo de *clustering* reproduce la verdadera estructura de los datos. Idealmente el procedimiento consiste en la comparación de resultados de un método de *clustering* y la verdadera partición de los datos [7], por lo cual un problema que inicialmente es de clasificación no supervisada se traslada a un problema de clasificación supervisada.

Entender las propiedades, sus limitaciones y los supuestos asumidos resulta clave para la correcta utilización e interpretación de los índices.

Esta investigación se centra en el estudio de las propiedades teóricas y empíricas de un índice de comparación de particiones llamado Mínimo Error de Clasificación (MCE), propuesto por Meila en [17]. Se estudian sus propiedades, en diversas condiciones experimentales y haciendo foco en la función de distribución teórica, con el fin de avanzar en el desarrollo de técnicas de análisis de aplicación más general, levantando supuestos y restricciones.

El informe se estructura de la siguiente forma. En el primer capítulo se

presenta una elaboración del estado del arte en cuanto a índices de comparación de particiones y sus propiedades. En el segundo capítulo planteamos la definición del MCE y se demuestran propiedades, en particular su distribución teórica. La expresión analítica de la función de distribución del índice es obtenida para el caso de comparación de particiones independientes, con dos poblaciones balanceadas. Algunas propiedades son demostradas para el caso general. Se muestran resultados sobre simulaciones con variaciones en los parámetros experimentales tamaño de muestra, cantidad de grupos y dependencia o superposición de las particiones. El capítulo finaliza con la aplicación de la metodología a un problema de clasificación de imágenes con datos reales. Los resultados se comparan con otros índices de validación. El tercer capítulo presenta una prueba de hipótesis para la independencia de particiones, diseñada a partir de los resultados teóricos sobre el MCE. Se estudia el desempeño del test a partir de las particiones simuladas artificialmente. Finalmente se exponen consideraciones finales sobre el trabajo y posibles líneas de investigación a futuro.

Capítulo 1

Medidas de comparación de particiones

En este capítulo introducimos la notación básica común para luego repasar varios índices de comparación de dos particiones que encontramos en la literatura. Asimismo se presenta para cada uno un resumen de sus propiedades.

1.1. Agrupamiento de datos (*clustering*)

En un problema de clasificación no supervisada se trabaja a partir de un conjunto de n observaciones $\mathscr{L} = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^p$ provenientes de manera independiente de una variable aleatoria multidimensional $X = (X_1, \ldots, X_p)$, donde $X_1, \ldots, X_p$ son variables aleatorias (continuas o categóricas) unidimensionales. El método de agrupamiento de datos (o *clustering*) busca formar subgrupos, particionando el conjunto de datos en clases o grupos tal que los individuos de una misma clase son similares entre sí y relativamente diferentes a los de otra clase.

Denotamos por $\mathscr{C}$ a una partición o *clustering* que resulta de un análisis de *cluster* (o, en caso que corresponda, a la verdadera partición de los datos). $\mathscr{C}$ es una colección $\{C_1, \ldots, C_J\}$ de subconjuntos disjuntos dos a dos, no vacíos y cuya unión es igual al conjunto de datos originales $\mathscr{L}$. El conjunto de todos los *clustering* de $\mathscr{L}$ es denotado como $\mathcal{P}(\mathscr{L})$, ya que coincide con el conjunto de partes de $\mathscr{L}$. Denotamos $\mathscr{C}' = \{C'_1, \ldots, C'_L\} \in \mathcal{P}(\mathscr{L})$ un segundo *clustering* de

$\mathscr{L}$. Las cantidades J y L refieren a la cantidad de *clusters* (clases o grupos) de las particiones $\mathscr{C}$ y $\mathscr{C}'$, respectivamente y pueden ser diferentes en principio.

Observar que una vez hecho el *clustering*, cada observación $x_1, \ldots, x_n$ de $\mathscr{L}$ queda identificada por una etiqueta $y_1, \ldots, y_n$ de pertenencia a un grupo.

A través de los índices de comparación de particiones se busca saber si las particiones $\mathscr{C}$ y $\mathscr{C}'$ son similares y cuán diferentes son. En el contexto supervisado, cuando una de las particiones se corresponde con la verdadera etiqueta de los datos, el cálculo de índices de validación externa permiten validar los resultados de una clasificación no supervisada.

1.2. Índices de comparación de particiones

Los índices de validación externa se definen como medidas de similaridad o disimilaridad. Pueden clasificarse en tres tipos, cuyo orden de presentación se corresponde con la cronología de sus desarrollos [26]:

1. Conteo de pares (desarrollada en los años 1970's y 1980's): las medidas cuentan la cantidad de pares de observaciones que son clasificados de la misma forma por las diferentes particiones.
2. Superposición de conjuntos (tipología desarrollada en los años 1990's): las medidas basan su cálculo en la tabla o matriz de contingencia de las particiones.
3. Información mutua (2002, 2003): son medidas basadas en la teoría de la información y noción de entropía.

En las siguientes subsecciones se describen dichas tipologías y se presentan algunos índices de comparación pertenecientes a las mismas, siguiendo el artículo de Wagner [26], con un ejemplo ilustrativo.

1.2.1. Índices basados en conteo de pares

Una forma intuitiva de comparar particiones es contando los pares de observaciones que son clasificados de la misma forma en ambas particiones, es decir, elementos que están en el mismo *cluster* (o diferentes *clusters*, según

corresponda) en las particiones que se comparan.

El conjunto de todos los pares (no ordenados) de $\mathscr{L}$ es la unión disjunta de los siguientes conjuntos:

- S_a={pares que están en el mismo *cluster* en $\mathscr{C}$ y $\mathscr{C}'$}
- S_b={pares que están en distintos *clusters* en $\mathscr{C}$ y $\mathscr{C}'$}
- S_c={pares que están en el mismo *cluster* en $\mathscr{C}$ y distinto *cluster* en $\mathscr{C}'$}
- S_d={pares que están en distintos *clusters* en $\mathscr{C}$ y en el mismo *cluster* en $\mathscr{C}'$}

Se denotan como a, b, c y d a los cardinales de S_a, S_b, S_c y S_d, respectivamente y es claro que:

- S_a, S_b, S_c y S_d son disjuntos dos a dos
- $a + b + c + d = \frac{n(n-1)}{2} = \binom{n}{2}$.

Ejemplo:

El siguiente ejemplo nos servirá para ilustrar los diferentes índices que incluiremos en este capítulo. Para un conjunto de datos con $n = 8$ observaciones, se definen las particiones, con $J = L = 3$ clases:

$$\mathscr{C} = \{C_1 = \{x_1, x_2, x_3\}, C_2 = \{x_4, x_5\}, C_3 = \{x_6, x_7, x_8\}\}$$

$$\mathscr{C}' = \{C_1' = \{x_6, x_7\}, C_2' = \{x_1, x_2, x_3, x_4, x_5\}, C_3' = \{x_8\}\}$$

Los pares de observaciones a comparar en las particiones son los siguientes:
$\{(x_1, x_2), (x_1, x_3), (x_1, x_4), (x_1, x_5), (x_1, x_6), (x_1, x_7), (x_1, x_8), (x_2, x_3), (x_2, x_4), (x_2, x_5),$
$(x_2, x_6), (x_2, x_7), (x_2, x_8), (x_3, x_4), (x_3, x_5), (x_3, x_6), (x_3, x_7), (x_3, x_8), (x_4, x_5), (x_4, x_6),$
$(x_4, x_7), (x_4, x_8), (x_5, x_6), (x_5, x_7), (x_5, x_8), (x_6, x_7), (x_6, x_8), (x_7, x_8)\}$

A partir de la comparación, dichos elementos se clasifican en los siguientes conjuntos definidos anteriormente:

- $S_a = \{(x_1, x_2), (x_1, x_3), (x_2, x_3), (x_4, x_5), (x_6, x_7)\}$
- $S_b = \{(x_1, x_6), (x_1, x_7), (x_1, x_8), (x_2, x_6), (x_2, x_7), (x_2, x_8), (x_3, x_6),$
  $(x_3, x_7), (x_3, x_8), (x_4, x_6), (x_4, x_7), (x_4, x_8), (x_5, x_6), (x_5, x_7), (x_5, x_8)\}$

- $S_c = \{(x_6, x_8), (x_7, x_8)\}$
- $S_d = \{(x_1, x_4), (x_2, x_4), (x_3, x_4), (x_1, x_5), (x_2, x_5), (x_3, x_5)\}$

Por lo cual el conteo de los $n(n-1)/2 = 28$ pares tiene como resultado las cantidades $a = 5, b = 15, c = 2$ y $d = 6$.

Dentro de las medidas basadas en conteo de pares se encuentran las siguientes:

Índice de Rand [20] Es una medida de similaridad y se define como:

$$R(\mathscr{C}, \mathscr{C}') = \frac{a+b}{a+b+c+d} = \frac{2(a+b)}{n(n-1)}.$$

Cuenta la proporción de pares de elementos correctamente clasificados de la segunda partición respecto a la primera. Es nulo cuando ningún par de observaciones es clasificado de la misma forma por ambos *clusterings* y vale uno cuando dos *clusterings* son idénticos.

En el ejemplo, este índice vale $R(\mathscr{C}, \mathscr{C}') = 0.72$.

Índice de Rand Ajustado Se muestra que el valor esperado del índice de Rand de dos particiones aleatorias no es cero. Hubert y Arabie proponen en [8] un ajuste, basado en el "modelo de permutaciones", en el cual los *clusterings* son generados aleatoriamente sujeto a un número fijo de cantidad de grupos y de elementos en cada *cluster* [25]. El índice de Rand Ajustado se determina como la diferencia normalizada del índice de Rand y su valor esperado cuando compara dos particiones independientes y se define como:

$$R_{adj}(\mathscr{C}, \mathscr{C}') = \frac{R(\mathscr{C}, \mathscr{C}') - \mathbb{E}(\mathbb{R}(\mathscr{C}, \mathscr{C}'))}{\text{máx}(R(\mathscr{C}, \mathscr{C}')) - \mathbb{E}(R(\mathscr{C}, \mathscr{C}'))} =$$

$$\frac{\sum_{i=1}^{J}\sum_{j=1}^{L}\binom{m_{ij}}{2} - \left[\sum_{i=1}^{J}\binom{|C_i|}{2}\sum_{j=1}^{L}\binom{|C'_j|}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i=1}^{J}\binom{|C_i|}{2}\sum_{j=1}^{L}\binom{|C'_j|}{2}\right] - \left[\sum_{i=1}^{J}\binom{|C_i|}{2}\sum_{j=1}^{L}\binom{|C'_j|}{2}\right]/\binom{n}{2}}$$

donde $m_{ij} = |C_i \cap C'_j|$.

En términos de conteo de pares puede expresarse como:

$$R_{adj}\left(\mathscr{C},\mathscr{C}'\right)=\frac{a-((a+d)(a+c)/(a+b+c+d))}{\frac{(a+d)+(a+c)}{2}-\frac{(a+d)(a+c)}{a+b+c+d}}.$$

El índice toma valor cero para particiones independientes y uno para particiones idénticas, aunque el ajuste también puede producir valores negativos del índice. En el ejemplo el valor del índice es $R_{adj}(\mathscr{C},\mathscr{C}')=0.36$.

Índice de Fowlkes-Mallows [4] Es una medida de similaridad que toma valores entre cero y uno y se define por:

$$FM\left(\mathscr{C},\mathscr{C}'\right)=\frac{a}{\sqrt{(a+c)(a+d)}}.$$

Esta medida puede ser interpretada como la media geométrica de la precisión $\frac{a}{a+c}$ y exhaustividad (*recall* en inglés) $\frac{a}{a+d}$. Al igual que el índice de Rand, la similaridad de dos *clustering* se corresponde con la desviación del valor esperado bajo la hipótesis nula de dos particiones independientes con clases de tamaño fijo. En el ejemplo ilustrativo $FM(\mathscr{C},\mathscr{C}')=0.57$.

Métrica de Mirkin [3] Es una medida de disimilaridad. Vale cero para particiones idénticas y es mayor que cero si no. Se define por:

$$Mirkin\left(\mathscr{C},\mathscr{C}'\right)=2(d+c).$$

La métrica es una variación del índice del Rand ya que puede escribirse como [14]:

$$Mirkin\left(\mathscr{C},\mathscr{C}'\right)=n(n-1)\left(1-R\left(\mathscr{C},\mathscr{C}'\right)\right),$$

y en el ejemplo el índice toma el valor $Mirkin(\mathscr{C},\mathscr{C}')=16.2$.

Índice de Jaccard Introducido en [9] citado por [26]. Mide la similitud entre dos particiones con valores entre cero y uno. Es muy similar al índice de Rand, pero desestima los pares de elementos que están en diferentes *clusters* en las particiones comparadas. Se define como:

$$J(\mathscr{C},\mathscr{C}')=\frac{a}{a+c+d},$$

y con los datos del ejemplo vale $J(\mathscr{C}, \mathscr{C}') = 0.38$.

Indice de Diferencia de particiones (*Partition Difference* en inglés). Presentado en [13] citado por [26]. Es una medida de similaridad que cuenta los pares de elementos que pertenecen a diferentes *clusters* en ambas particiones.

$$PD(\mathscr{C}, \mathscr{C}') = b.$$

En el ejemplo $PD(\mathscr{C}, \mathscr{C}') = 15$.

Muchas de estas medidas tienen propiedades no deseables como sensibilidad al número de *clusters*, al número de observaciones y al tamaño relativo de los *clusters*. En la sección 1.3 se mencionan antecedentes al respecto.

1.2.2. Índices basados en superposición de conjuntos

Las medidas basadas en superposición de conjuntos son aquellas que basan su cálculo en una tabla de contingencia o matriz de confusión entre las particiones $\mathscr{C}$ y $\mathscr{C}'$ denotada como $M = (m_{ij})$. Dicha matriz tiene dimensión $J \times L$ y contiene en la ij-ésima entrada al número de elementos comunes de los *clusters* C_i y C'_j, por lo tanto, $m_{ij} = \left|C_i \cap C'_j\right|$, $1 \leq i \leq J$, $1 \leq j \leq L$.

Como las etiquetas de las particiones a comparar son arbitrarias, las clases de dos particiones son pareadas cuando tienen máxima superposición de casos, absoluta o relativa. A partir del emparejamiento se computan distintos índices de comparación.

Ejemplo:

Retomado el ejemplo práctico planteado en la subsección 1.2.1, podemos calcular la tabla de contingencia de $\mathscr{C}$ y $\mathscr{C}'$:

$$M = \begin{bmatrix} 0 & 3 & 0 \\ 0 & 2 & 0 \\ 2 & 0 & 1 \end{bmatrix}.$$

Por superposición máxima de casos se emparejan los conjuntos (C_1, C'_2), (C_2, C'_3) y (C_3, C'_1), con las correspondientes etiquetas $(1;2)$, $(2;3)$ y $(3;1)$.

Algunos de los índices tienen otros criterios de emparejamiento, como por ejemplo el máximo de cada fila de la tabla de contingencia, lo cual puede dar lugar a clases pareadas con más de una clase de la segunda partición. En el ejemplo sería $(1;2)$, $(2;2)$ y $(3;1)$.

A continuación se referencian varios índices de esta tipología. El índice Mínimo Error de Clasificación (MCE) [16] es uno de ellos y en este capítulo será introducido con la denominación y notación que encontramos en las referencias bibliográficas. Su definición y propiedades son el principal objeto de estudio de este trabajo y se desarrollan con más detalle en el Capítulo 2, con una notación diferente.

Índice de Meila-Heckerman [17] Se define por:

$$MH(\mathscr{C}, \mathscr{C}') = \frac{1}{n} \sum_{i=1}^{J} \max_{C'_j \in \mathscr{C}} m_{ij}$$

El índice suma los máximos de cada fila (o columna) de la tabla de contingencia. Su asimetría lo hace una medida inapropiada para la comparación de particiones.

Considerando la tabla de contingencia del ejemplo, obtenemos el valor $MH(\mathscr{C}, \mathscr{C}') = 0.75$.

Medida de clasificación correcta y error de clasificación La medida de clasificación correcta fue introducida por Meila en [17]. Es una generalización simétrica del índice anterior. Se define como la proporción de casos correctamente clasificados.

$$MM(\mathscr{C}, \mathscr{C}') = \frac{1}{n} \sum_{i=1}^{min(J,L)} m_{ii'}$$

donde i' es el índice del *cluster* de $\mathscr{C}'$ que es pareado con $C_i \in \mathscr{C}$.

Los grupos se emparejan de manera unívoca según la máxima superposición de casos de la tabla de contingencia, como en el primer caso del ejemplo planteado.

Otro procedimiento para su cálculo consiste en buscar el valor máximo m_{ij} de la tabla de contingencia M y emparejar los *clusters* correspondientes C_i y C'_j, ya que son los que tienen superposición absoluta de casos. Luego suprimir la i-ésima fila y j-ésima columna de la matriz M correspondientes a los grupos pareados y repetir el procedimiento hasta que la matriz tenga tamaño cero. El cálculo del índice finaliza con la suma de las cantidades de los grupos pareados en cada paso y dividiendo por el total de elementos, como puede verse en la fórmula de cálculo de este índice.

Esta medida es el complemento del índice MCE, estudiado con más detalle en el Capítulo 2, que mide el error de clasificación.

$$MCE(\mathscr{C}, \mathscr{C}') = 1 - MM(\mathscr{C}, \mathscr{C}')$$

En [15], Meila define la medida de error de clasificación como:

$$CE(\mathscr{C}, \mathscr{C}') = 1 - \frac{1}{n} \max_{\sigma} \sum_{k=1}^{J} n_{k,\sigma(k)}$$

asumiendo que $J \leq L$, donde $n_{k,k'} = |C_k \cap C'_{k'}|$ y σ es un mapeo inyectivo de $\{1, \ldots, J\}$ en $\{1, \ldots, L\}$.

En el ejemplo que venimos siguiendo a lo largo de la sección, los valores de los índices son $MCE(\mathscr{C}, \mathscr{C}') = CE(\mathscr{C}, \mathscr{C}') = 0.375$ y $MM(\mathscr{C}, \mathscr{C}') = 0.625$.

Medida Van Dongen [3] Es una medida simétrica de disimilaridad entre particiones, basada también en la intersección máxima de *clusters*. Vale cero cuando las particiones son iguales. Su definición es la siguiente:

$$VD(\mathscr{C}, \mathscr{C}') = 2n - \sum_{i=1}^{J} \max_{j} m_{ij} - \sum_{j=1}^{L} \max_{i} m_{ij}$$

Su valor en el ejemplo es $VD(\mathscr{C}, \mathscr{C}') = 3$.

Se debe notar que, para los casos en que $J \neq L$, los índices basados en superposición de conjuntos desestiman a las $|J - L|$ clases de la partición con mayor cardinalidad.

1.2.3. Índices basados en información mutua

Esta aproximación para la comparación de *clusterings* se basa en la teoría de la información y en la noción de entropía como medida de incertidumbre. Aplicado a *clusterings* la entropía puede definirse de la siguiente forma: la probabilidad de que un elemento esté en el *cluster* $C_i \in \mathscr{C}$ es $\mathbb{P}(i) = |C_i| / n$, entonces la entropía asociada a la partición $\mathscr{C}$ con J clases es:

$$H(\mathscr{C}) = -\sum_{i=1}^{J} \mathbb{P}(i) log_2 \mathbb{P}(i).$$

La noción de entropía de una partición puede extenderse a la de *información mutua* entre dos particiones, la cual mide la dependencia entre ellas. Se interpreta como la reducción de incertidumbre de una partición que se produce debido al conocimiento de otra partición del mismo conjunto de datos. Si las particiones son independientes, la primera partición no aportará información de la segunda (y viceversa) por lo que la información mutua es cero. En el extremo opuesto si las particiones $\mathscr{C}$ y $\mathscr{C}'$ son idénticas, la información mutua es uno, es decir, toda la información de la primera partición es compartida por la segunda. Más precisamente, la información mutua entre dos particiones $\mathscr{C}$ y $\mathscr{C}'$ se define como:

$$I(\mathscr{C}, \mathscr{C}') = \sum_{i=1}^{J} \sum_{j=1}^{L} \mathbb{P}(i,j) log_2 \frac{\mathbb{P}(i,j)}{\mathbb{P}(i)\mathbb{P}(j)},$$

donde $\mathbb{P}(i,j)$ es la probabilidad de que un elemento esté en el *cluster* C_i de $\mathscr{C}$ y C'_j de $\mathscr{C}'$.

En nuestro ejempo los valores de entropía e información mutua son los siguientes: $H(\mathscr{C}) = 1.56$, $H(\mathscr{C}') = 1.3$ y $I(\mathscr{C}, \mathscr{C}') = 0.95$.

La información mutua I es una métrica en el espacio de todos los *clusterings*, sin embargo, el hecho de no estar acotada por un valor constante dificulta su interpretación. Como la información mutua entre dos particiones

sí está acotada por el valor de sus entropías $I(\mathscr{C}, \mathscr{C}') \leq \min(H(\mathscr{C}), H(\mathscr{C}'))$, surgen algunas medidas normalizadas que presentamos a continuación.

Información mutua normalizada de Strehl y Ghosh Es una medida de similaridad introducida en [24], citado por [26]. Es la normalización de la información mutua por la media geométrica de las entropías. Se define como

$$NMI_1(\mathscr{C}, \mathscr{C}') = \frac{I(\mathscr{C}, \mathscr{C}')}{\sqrt{H(\mathscr{C})H(\mathscr{C}')}}.$$

El índice toma valores entre cero y uno y el valor calculado en el ejemplo es $NMI_1(\mathscr{C}, \mathscr{C}') = 0.670$.

Información mutua normalizada de Fred y Jain Esta medida es muy similar a la anterior. En este caso los autores en [6] normalizan la información mutua por la media aritmética de las entropías, quedando la siguiente definición:

$$NMI_2(\mathscr{C}, \mathscr{C}') = \frac{2I(\mathscr{C}, \mathscr{C}')}{H(\mathscr{C}) + H(\mathscr{C}')}.$$

El índice toma valores entre cero y uno y en el ejemplo vale $NMI_2(\mathscr{C}, \mathscr{C}') = 0.667$.

Variación de información En [16] la autora propone esta medida basada en la noción de entropía que se define como:

$$VI(\mathscr{C}, \mathscr{C}') = H(\mathscr{C}) + H(\mathscr{C}') - 2I(\mathscr{C}, \mathscr{C}').$$

El índice toma valores en el rango $[2/n, \log n]$ y con los datos del ejemplo $VI(\mathscr{C}, \mathscr{C}') = 0.95$.

1.3. Propiedades de algunos índices de comparación de particiones

Los índices de validación externa deben satisfacer ciertas propiedades para que sean útiles, interpretables y que permitan la comparabilidad de resultados

en diferentes conjuntos de datos. Medidas normalizadas con rangos por ejemplo entre $[0, 1]$ o $[-1, 1]$ son preferibles en este sentido, así como también es ventajoso que cumplan las propiedades de una métrica (ver detalles en [16], [26] y [21]) y que su valor no dependa de ciertos parámetros como la cantidad de grupos, el tamaño relativo de los mismos y la cantidad de observaciones.

En esta sección resumimos estudios sobre las propiedades de los índices de validación externa. Se presentan en primer lugar los antecedente de tipo experimental —que evalúan el desempeño de los índices en diferentes condiciones— y luego los resultados sobre características de los índices que fueron demostrados teóricamente, como las propiedades métricas, con especial interés en los antecedentes que estudian la distribución teórica de las medidas.

1.3.1. Antecedentes empíricos

En [21] se analiza la comparación empírica con varios índices de validación sobre particiones simuladas artificialmente. Los resultados obtenidos muestran que los índices basados en superposición de conjuntos tienen un mejor desempeño que los basados en conteo de pares e información mutua, en aspectos como dependencia al tamaño relativo de los *clusters* (clases no balanceadas), cantidad de *clusters* y en la linealidad de los cambios con la superposición de las particiones.

Milligan y Cooper estudian en [18] el comportamiento de los índices de Rand, Rand ajustado (R_{adj}), Jaccard y Fowlkes Mallows, al variar las condiciones en el número y tamaño de los *clusters* en una muestra pequeña. El índice R_{adj} es el índice que muestra mejor desempeño con una menor dependencia de dichos parámetros.

Steinley realiza en [23] un análisis empírico de las propiedades del índice R_{adj} con otros índices, en particular con el MCE. Demuestra buenas propiedades del R_{adj} en cuanto a invarianza a cambios en el números de *clusters* y tamaño relativo de los grupos, tendiendo a focalizar toda su discriminación en las diferencias de superposición de las particiones. Este trabajo, si bien es de carácter experimental, sugiere profundizar la investigación en técnicas paramétricas, las cuales requieren la derivación de la distribución teórica del

índice R_{adj}.

En los trabajos [1], [27] y [2] se encuentran estudios sobre el comportamiento de índices de validación externa en distintas condiciones experimentales y la asociación de sus valores con el error de clasificación. Se constata la consistencia de resultados en medidas normalizadas. El índice R_{adj} muestra los mejores resultados en contexto de clases no balanceadas.

Por otra parte, se relevaron los trabajos antecedentes que analizan la distribución empírica y el comportamiento ante cambios en la hipótesis de comparación.

Saporta y Youness proponen en [22], [28] y [29] métodos para encontrar la distribución empírica de índices de comparación de particiones. Parten de la hipótesis de que las particiones comparadas son similares. En las dos primeras referencias los datos son simulados a partir de un modelo de clase latente y se comparan las particiones generadas por métodos de *clustering* en base a dos subconjuntos disjuntos de variables. En [29] utilizan el método de proyección: comparan el resultado de un método de *clustering* con la predicción que surge de un método de clasificación supervisada. Los autores estudian los índices Rand, Jaccard, entre otros. Estiman la distribución empírica bajo la hipótesis de particiones similares y los valores críticos de una prueba de hipótesis para la igualdad de particiones, con el percentil 5 % de la distribución. No se obtienen valores críticos universales por el hecho de que la distribución depende de la cantidad de *clusters*. Además estudian la asociación entre los diferentes índices y se verifica que las distribuciones de los índices cambian si las particiones comparadas son independientes, siendo bimodales.

Si bien es común comparar estos índices, son muy pocas las propiedades teóricas desarrolladas hasta el momento. El objeto de la siguiente subsección es repasar algunas de ellas.

1.3.2. Antecedentes teóricos

En el estudio [25] se discuten propiedades importantes en medidas de comparación de *clustering*, tales como ser métrica (satisfacción de las propiedades

de no negatividad, reflexividad, simetría y desigualdad triangular) y la propiedad de normalización (que facilita la interpretación y comparación en diferentes condiciones). Se centra en medidas basadas en información mutua y también considera el índice de Rand normalizado (R_{adj}).

En [16] se demuestran numerosas propiedades del índice de Variación de información (VI), entre ellas:

- Es una métrica en en el espacio de todos los *clustering*.
- n-invarianza: El valor de $VI(\mathscr{C}, \mathscr{C}')$ depende sólo del tamaño relativo de los *clusters*, no dependiendo directamente del número de elementos del conjunto de datos.
- La siguiente cota es alcanzada para todo n: $VI(\mathscr{C}, \mathscr{C}') \leq \log n$.
- Si $|\mathscr{C}| \leq J^*$ y $|\mathscr{C}'| \leq J^*$, con $J^* \leq \sqrt{n}$, entonces $VI(\mathscr{C}, \mathscr{C}') \leq 2 \log J^*$
- Es un criterio sensible para comparar particiones que están alineadas en el enrejado de particiones[1]. Se verifican los axiomas denominados aditividad del refinamiento, aditividad de la unión y convexidad aditiva (ver detalles en [16]).

En [16] la autora demuestra que el índice de Rand, Rand ajustado, Fowlkes-Mallows y Jaccard son asintóticamente n-invariantes cuando n es grande. La métrica de Mirkin y Van Dongen dependen fuertemente de n, si bien son fácilmente escalables para ser n-invariantes. En el mismo trabajo se hace referencia a que la métrica de Mirkin normalizada ($Mirkin(\mathscr{C}, \mathscr{C}')/n^2$) y el índice de Rand Ajustado también satisfacen las propiedades de una métrica.

En [15], además de hacer una caracterización axiomática de las medidas de validación externa como Variación de información, Mirkin, Rand y Van Dongen, se discuten las propiedades del "error de clasificación", que en el presente trabajo denominamos Mínimo error de clasificación (MCE). Allí se expone el cumplimiento de las siguientes propiedades:

- Simetría $MCE(\mathscr{C}, \mathscr{C}') = MCE(\mathscr{C}', \mathscr{C})$
- Escala. Se denota $\mathscr{C}_J^U$ el *clustering* uniforme, es decir, la partición con J clases iguales y $\hat{1}$ el *clustering* que contiene los n elementos del conjunto en una clase. Si $\mathscr{C}_J^U$ existe, entonces $MCE(\hat{1}, \mathscr{C}_J^U) = 1 - \frac{1}{J}$

[1] También conocido como diagrama de Hess del enrejado. La descripción puede verse con más detalle en el trabajo referenciado.

- Convexidad aditiva. $\mathscr{C}'$ es un refinamiento de $\mathscr{C}$ si para cada $C'_j \in \mathscr{C}'$ hay un único *cluster* $C_j \in \mathscr{C}$ tal que $C'_j \subseteq C_j$. Dado $\mathscr{C} = \{C_1, ..., C_J\}$ un *clustering* y $\mathscr{C}'$ un refinamiento de $\mathscr{C}$. Se denota C'_j a la partición inducida por $\mathscr{C}'$ en el conjunto de elementos C_j y $\hat{1}_{n_j}$ al *clustering* del conjunto de datos C_j que contiene los n_j puntos en un grupo. Entonces $MCE(\mathscr{C}', \mathscr{C}) = \sum_{j=1}^{J} \frac{n_j}{n} MCE(\hat{1}_{n_j}, C'_j)$.

Si bien no son los aspectos a profundizar en este trabajo, cabe mencionar que —de los tres axiomas de aditividad sobre el enrejado de particiones verificados para el índice Variación de información (aditividad del refinamiento, aditividad de la unión y convexidad aditiva)— Meila demuestra que el índice error de clasificación sólo cumple con la convexidad aditiva. En [19] también se demuestran numerosas propiedades del MCE en comparación de subespacios de *clustering*.

Con respecto a la distribución teórica de los índices, en [10] (citado por [28]) se deriva el siguiente resultado asintótico para la distribución del índice de Rand ($R(\mathscr{C}, \mathscr{C}')$): si el índice de Rand compara dos particiones independientes, con J clases equiprobables, la distribución asintótica es Normal con esperanza $\mathbb{E}(R(\mathscr{C}, \mathscr{C}')) = 1 - \frac{2}{J} + \frac{2}{J^2}$ y varianza $\mathbb{V}(R(\mathscr{C}, \mathscr{C}')) = \frac{1}{n^2}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{J} + \frac{2}{J^2}\right)\left(\frac{2}{J} - \frac{2}{J^2}\right)$. Este resultado no es válido para J pequeño, especialmente $J = 2$, y sólo es aproximadamente válido para n grande.

En la Tabla 1.1 se presenta un resumen de los índices de comparación seleccionados, con el rango de valores posibles, señalando cuáles de ellos tienen demostradas teóricamente características que tienen que ver con propiedades de una métrica y de su función de distribución.

Índice de comparación de particiones	**Rango**	**Métrica**	**Distribución**
Rand	$[0, 1]$		✓
Rand ajustado	$[0, 1]$	✓	
Jaccard	$[0, 1]$		
Fowlkes Mallows	$[0, 1]$		
Mirkin	$[0, n(n-1)]$	✓	
Diferencia de particiones	$[0, n(n-1)/2]$		
Meila-Heckerman	$[0, 1]$		
Medida de clasificación correcta	$[0, 1]$	✓	
Mínimo error de clasificación	$\left[0, \frac{J-1}{J}\right]$	✓	
Van Dongen	$[0, 2n)$	✓	
Información mutua normalizada de Strehl y Ghosh	$[0, 1]$		
Información mutua normalizada de Fred y Jain	$[0, 1]$		
Variación de información	$[2/n, \log n]$	✓	

Tabla 1.1: Índices de comparación de particiones donde se indica si fueron probadas propiedades de métrica y distribución en la literatura.

Una observación a destacar que podemos hacer a partir de la revisión de antecedentes es que el error de clasificación, que coincide con el índice Mínimo error de clasificación (MCE) que estamos estudiando, es tomado como referencia en varias oportunidades para evaluar el desempeño de otros índices de comparación de particiones, por lo cual en dichos trabajos es considerado a *priori* como una medida intuitiva de validación externa. Vimos en los antecedentes bibliográficos que los índices Variación de información y el MCE fueron muy estudiados en el cumplimiento de sus propiedades como métricas. El índice de Rand Ajustado, la medida de Mirkin y Van Dongen también tienen menciones que indican que cumplen propiedades de una métrica. Además se identificaron trabajos que evidencian un mejor desempeño de las medidas basadas en superposición de conjuntos respecto a otras y buenos desempeños del índice de Rand normalizado en cuanto a robustez ante cambios en las condiciones experimentales. Sin embargo, sólo se encontraron resultados teóricos sobre la función de distribución para el índice de Rand y en ciertas condiciones. Queda en evidencia que son escasos los desarrollos para derivar la función de distribución de los índices de comparación de particiones, que permitan evaluar los resultados de *clustering* a través de técnicas basadas en la distribución paramétrica de los índices que sean generalizables, levantando supuestos y restricciones. Este es el punto donde el presente trabajo pretende avanzar con el índice MCE.

1.4. Un primer ejemplo simulado de distintos métodos de *clustering* y evaluación de los índices

En esta sección generamos una simulación de datos simple, donde se aplicarán varios métodos de *clustering*, para evaluar el desempeño de algunos índices presentados en la sección anterior. Tal como en [11], simulamos 75 observaciones bivariadas provenientes de una variable aleatoria normal bivariada estándar. A partir de éstas se generan $J = 3$ grupos de 25 observaciones, variando la media de cada grupo (sumando a las variables simuladas una constante diferente para cada uno de los tres grupos). La Figura 1.1 muestra el gráfico de la simulación. Se guarda el vector con las etiquetas del verdadero grupo para cada observación, que luego será comparado con los resultados obtenidos por los métodos de *clustering*.

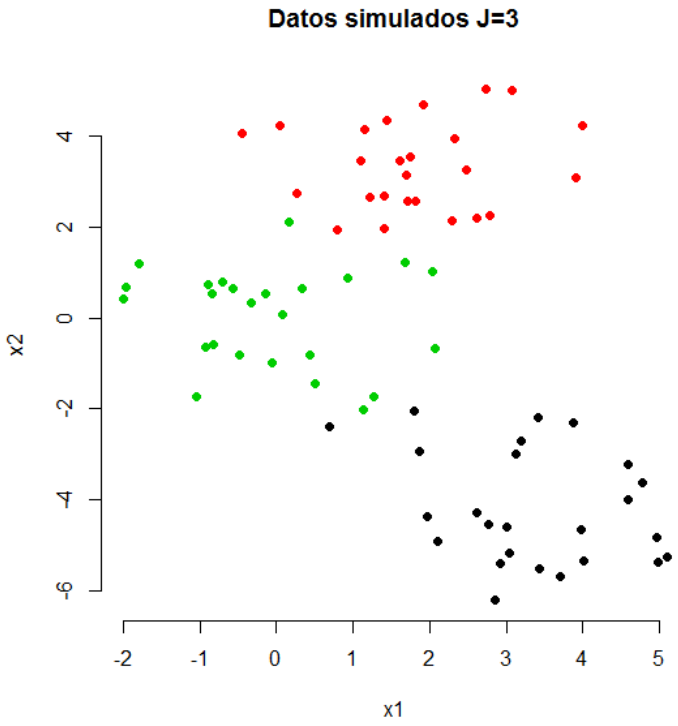

Figura 1.1: Simulación de observaciones con $J = 3$.

El siguiente gráfico muestra la clasificación de las observaciones en $J = 3$

grupos, según distintos métodos: *k-means* [11] con $k = J$, *clustering* jerárquico (utilizandos los criterios de unión *complete*, *average* y *single*) [11], *Partitioning Around Medoids* (PAM) [12] y *clustering* basados en modelos [5]. Los métodos de *clustering* se grafican con distintos colores, mientras que los grupos originales se identifican con la forma del punto. Una primera observación gráfica muestra que el *clustering* jerárquico con el método *single* no tuvo un buen desempeño (Figura 1.2).

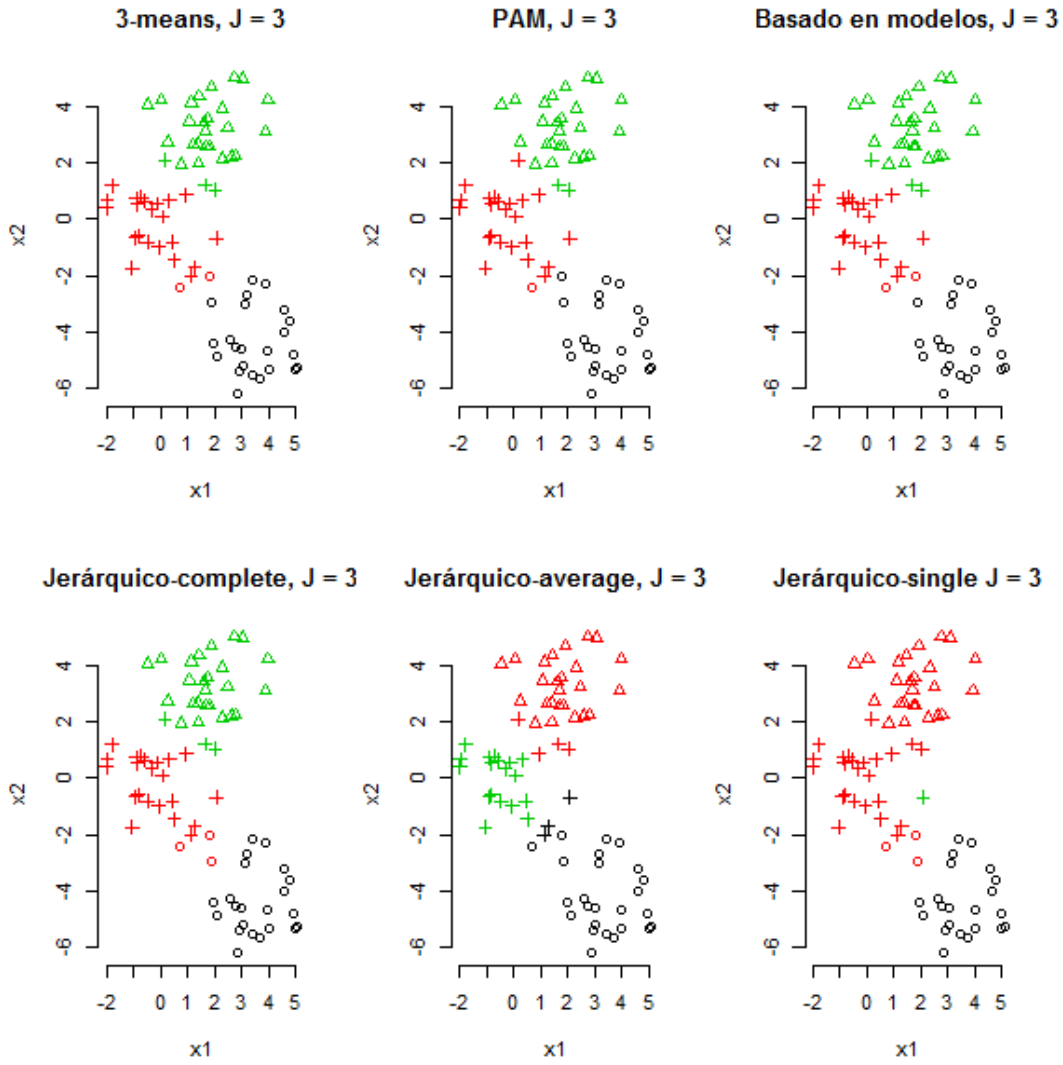

Figura 1.2: Clasificación por distintos métodos de *clustering* de un conjunto de 75 observaciones simuladas como en [11].

Comparación de los métodos a partir de los índices

Se implementa una función en **R** que devuelve las cantidades a, b, c y d, resultantes del conteo de pares para dos particiones dadas $\mathscr{C}$ y $\mathscr{C}'$. En la Tabla 1.2 se presentan dichas cantidades para la comparación de cada resultado de *clustering* con el vector de grupos original. El total de pares evaluados es $n(n-1)/2 = 75 \times 74/2 = 2775$.

	a	b	c	d
3−*means*	788	1756	119	112
PAM	830	1802	73	70
Basado en modelos	788	1756	119	112
Jerárquico (*complete*)	768	1734	141	132
Jerárquico (*average*)	762	1700	175	138
Jerárquico (*single*)	768	1734	141	132

Tabla 1.2: Conteo de pares entre las particiones obtenidas por *clustering* y el vector de grupos verdadero.

En las Tablas 1.3, 1.4 y 1.5 se presentan los índices basados en conteo de pares, las medidas basadas en la superposición de conjuntos o tabla de contingencia y los índices de información mutua, respectivamente, calculados para comparar el resultado del *clustering* con la verdadera etiqueta.

	Rand	Rand Ajustado	Jaccard	Fowlkes-Mallows	Mirkin	Diferencia de particiones
3-*means*	0.92	0.81	0.77	0.87	462	1756
PAM	0.95	0.88	0.85	0.92	286	1802
Basado en modelos	0.92	0.81	0.77	0.87	462	1756
Jerárquico (*complete*)	0.90	0.78	0.74	0.85	546	1734
Jerárquico (*average*)	0.89	0.75	0.71	0.83	626	1700
Jerárquico (*single*)	0.70	0.42	0.49	0.68	1674	1128

Tabla 1.3: Resultados de los índices basados en conteo de pares. Comparación entre las particiones obtenidas por *clustering* y el vector de grupos verdadero.

	Meila Heckerman	Clasificación correcta	Van Dongen	MCE
3-*means*	0.93	0.93	10	0.07
PAM	0.96	0.96	6	0.04
Basado en modelos	0.93	0.93	10	0.07
Jerárquico (*complete*)	0.92	0.92	12	0.08
Jerárquico (*average*)	0.91	0.91	14	0.09
Jerárquico (*single*)	0.95	0.64	31	0.36

Tabla 1.4: Resultados de los índices basados en superposición de conjuntos. Comparación entre las particiones obtenidas por *clustering* y el vector de grupos verdadero.

	Información mutua normalizada de Strehl y Ghosh	Información mutua normalizada de Fred y Jain	Variación de información
3-*means*	0.80	0.80	0.63
PAM	0.86	0.86	0.43
Basado en modelos	0.80	0.80	0.63
Jerárquico (*complete*)	0.77	0.77	0.71
Jerárquico (*average*)	0.75	0.75	0.78
Jerárquico (*single*)	0.57	0.56	1.13

Tabla 1.5: Resultados de los índices basados información mutua. Comparación entre las particiones obtenidas por *clustering* y el vector de grupos verdadero.

En primer lugar, se observa que la mayoría de las medidas producen el mismo ordenamiento en la evaluación de las particiones, por lo cual llevarían a una interpretación similar de los resultados. El índice de Meila-Heckerman es el único discordante que computa un mejor puntaje para el *clustering* jerárquico por el método *single* (0.95), debido a que no empareja unívocamente las clases.

Además podemos ver que:

- las medidas normalizadas permiten hacer la interpretación de los resultados de manera más intuitiva, por ejemplo índice de Rand que vale 0.92 para el *clustering* obtenido por el método PAM, a diferencia los índices Mirkin, Diferencia de particiones y Van Dongen que no están normalizados.
- verificamos que se distingue bien a los índices de similaridad (Rand, Rand ajustado, Jaccard, Fowlkes-Mallows, diferencia de particiones, Meila-Hckerman, clasificación correcta) de los de disimilaridad (Mirkin, Van Dongen y MCE), por ejemplo la medida de clasificación correcta es 0.96, al comparar la partición del método PAM con la verdadera partición de los datos, y el MCE es cercano a cero (0.04) para la misma comparación.
- la mayoría de los métodos tienen un buen desempeño, clasificando a una proporción alta de observaciones correctamente, exceptuando al método jerárquico *single*.
- la clasificación del método PAM es la mejor según todas las medidas calculadas y la obtenida por el método jerárquico *single* es la peor evaluada.
- las particiones generadas por el método 3-*means* y el método basado en modelos obtienen iguales valores en todos los índices de comparación.
- en cuanto a los métodos jerárquicos, según todos los índices, la mejor partición se obtiene por el método *complete* y le sigue el método *average*.

Capítulo 2

El índice mínimo error de clasificación (MCE)

En este capítulo se expone la definición y análisis del índice mínimo error de clasificación (MCE), introducido por Meila en [17]. El objetivo del análisis es establecer la función de distribución del índice, a partir de la cual se intentará diseñar técnicas de utilidad para evaluar si dos particiones son estadísticamente similares. La evaluación de una partición obtenida mediante *clustering* puede hacerse con respecto a las verdaderas etiquetas de los datos, es decir, como en el ejemplo anterior con datos etiquetados.

Este capítulo se estructura de la siguiente manera. En la primera sección se formaliza la definición del índice y se presenta un ejemplo ilustrativo. En la segunda y tercera sección se demuestran propiedades de la distribución y se deriva la expresión analítica de la función de distribución para dos particiones con dos *clusters*. En la cuarta sección se muestran avances para dos particiones con tres *clusters*. En la quinta sección se exponen resultados sobre datos simulados. Finalmente se comparan con los resultados de otros índices a partir de los datos simulados y en una aplicación a datos reales.

2.1. Definición del índice MCE

El índice de validación Mínimo error de clasificación (MCE) mide la "distancia" entre dos particiones según el error de clasificación de una partición con respecto a la otra. Como las etiquetas de las particiones son arbitrarias, la medida toma el mínimo error permutando de todas las maneras posibles las etiquetas de las observaciones.

Más precisamente, dado un conjunto de datos $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^p$ y $\mathscr{C}$ y $\mathscr{C}' \in \mathcal{P}(\mathscr{L})$ dos particiones de $\mathscr{L}$, denotamos las clases obtenidas en la primera partición $\mathscr{C}$ por $\{C_1, \ldots, C_J\}$ y las clases obtenidas en la segunda partición por $\{C'_1, \ldots, C'_L\}$. Las etiquetas de cada observación en la primera partición se denotan $\{y_1, \ldots, y_n\}$ y las de la segunda partición $\{\hat{y}_1, \ldots, \hat{y}_n\}$, es decir $y_i \in \{1, \ldots, J\}$ y $\hat{y}_i \in \{1, \ldots, L\}\, \forall i = 1, \ldots, n$.

Si S_J es el conjunto de permutaciones del vector de etiquetas $\{1, \ldots, J\}$, el error de clasificación (CE) asociado a una permutación $\sigma \in S_J$ se define como:

$$\tau_\sigma = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}}$$

Entonces el MCE de la segunda partición respecto de la primera es:

$$\tau = MCE(\mathscr{C}, \mathscr{C}') = \begin{cases} \text{mín}_{\sigma \in S_J}\, \tau_\sigma & \text{si } J \leq L \\ \text{mín}_{\sigma \in S_J}\, \frac{1}{n} \sum \mathbf{1}_{\{\sigma(y_i) \neq \widehat{y}_i\}} & \text{si no} \end{cases}$$

De aquí en adelante supondremos que las dos particiones comparadas tienen la misma cantidad de clases, es decir $J = L$.

Ejemplo:

Supongamos que tenemos $n = 6$ observaciones repartidas en $J = 3$ grupos, según las particiones $\mathscr{C}, \mathscr{C}'$ y $\mathscr{C}''$ (podemos pensar que $\mathscr{C}, \mathscr{C}'$ y $\mathscr{C}''$ son resultados de tres métodos de *clustering* distintos). Denotando a cada partición como un vector de etiquetas y con la notación del Capítulo 1 tenemos que:

$\mathscr{C} = \{y_1 = 1, y_2 = 1, y_3 = 2, y_4 = 2, y_5 = 2, y_6 = 3\} =$
$\{C_1 = \{x_1, x_2\}, C_2 = \{x_3, x_4, x_5\}, C_3 = \{x_6\}\}$

$\mathscr{C}' = \{\hat{y}_1 = 3, \hat{y}_2 = 3, \hat{y}_3 = 1, \hat{y}_4 = 1, \hat{y}_5 = 1, \hat{y}_6 = 2\} =$
$\{C_1 = \{x_3, x_4, x_5\}, C_2 = \{x_6\}, C_3 = \{x_1, x_2\}\}$

$\mathscr{C}'' = \{\hat{y}'_1 = 3, \hat{y}'_2 = 1, \hat{y}'_3 = 1, \hat{y}'_4 = 1, \hat{y}'_5 = 2, \hat{y}'_6 = 2\} =$
$\{C_1 = \{x_2, x_3, x_4\}, C_2 = \{x_5, x_6\}, C_3 = \{x_1\}\}$

Recordemos que el conjunto de permutaciones S_J de las etiquetas $\{1, 2, 3\}$ tiene $3! = 6$ elementos que se muestran en la Tabla 2.1:

$\sigma_1 = id$	$\sigma_2 = (23)$	$\sigma_3 = (12)$	$\sigma_4 = (123)$	$\sigma_5 = (132)$	$\sigma_6 = (13)$
$1 \longrightarrow 1$	$1 \longrightarrow 1$	$1 \longrightarrow 2$	$1 \longrightarrow 2$	$1 \longrightarrow 3$	$1 \longrightarrow 3$
$2 \longrightarrow 2$	$2 \longrightarrow 3$	$2 \longrightarrow 1$	$2 \longrightarrow 3$	$2 \longrightarrow 1$	$2 \longrightarrow 2$
$3 \longrightarrow 3$	$3 \longrightarrow 2$	$3 \longrightarrow 3$	$3 \longrightarrow 1$	$3 \longrightarrow 2$	$3 \longrightarrow 1$

Tabla 2.1: Permutaciones de $\{1, 2, 3\}$

Calculamos el error de clasificación (CE) $\tau_\sigma = \frac{1}{6} \sum \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}}$ para cada una de las seis permutaciones y luego el índice MCE.

Al comparar $\mathscr{C}$ con $\mathscr{C}'$ tenemos que $\tau_{\sigma_1} = \frac{1}{6} \sum_{i=1}^{6} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}} = 1$; $\tau_{\sigma_2} = \frac{1}{2}$; $\tau_{\sigma_3} = \frac{5}{6}$; $\tau_{\sigma_4} = \frac{4}{6}$; $\tau_{\sigma_5} = 1$; $\tau_{\sigma_6} = 0$

Por lo tanto $MCE(\mathscr{C}, \mathscr{C}') = \text{mín}\{\tau_{\sigma_1}, \tau_{\sigma_2}, \tau_{\sigma_3}, \tau_{\sigma_4}, \tau_{\sigma_5}, \tau_{\sigma_6}\} = 0$, por lo que $\mathscr{C}$ y $\mathscr{C}'$ son la misma partición del conjunto de datos.

Al comparar $\mathscr{C}$ con $\mathscr{C}''$, $\tau_{\sigma_1} = \frac{4}{6}$; $\tau_{\sigma_2} = \frac{4}{6}$; $\tau_{\sigma_3} = \frac{4}{6}$; $\tau_{\sigma_4} = \frac{4}{6}$; $\tau_{\sigma_5} = \frac{6}{6}$; $\tau_{\sigma_6} = \frac{2}{6}$

Por lo tanto $MCE(\mathscr{C}, \mathscr{C}'') = \frac{1}{3}$, por lo cual las particiones coinciden en $\frac{2}{3}$ de sus elementos.

2.2. Distribución del error de clasificación (CE)

Para derivar la distribución del MCE necesitamos conocer la distribución de los errores de clasificación τ_σ. Para y_i y $\hat{y}_i$ dos realizaciones de variables aleatorias independientes discretas con igual distribución y con etiquetas

$\{1, \ldots, J\}$, definimos:

- $p_j = \mathbb{P}[y_i = j]$, $\hat{p_j} = \mathbb{P}[\hat{y}_i = j]$, y
- $\theta = \mathbb{P}[y_i \neq \sigma(\hat{y}_i)] = 1 - \mathbb{P}[y_i = \sigma(\hat{y}_i)] = 1 - \sum_{j=1}^{J} \mathbb{P}[y_i = j, \sigma(\hat{y}_i) = j] = 1 - \sum_{j=1}^{J} p_j \hat{p}_j$

Suponiendo que y_i y $\hat{y}_i$ provienen de una variable aleatoria que toma valores uniformes en $\{1, \ldots, J\}$, entonces $p_j = \hat{p}_j = 1/J$ y $\theta = 1 - \frac{1}{J}$.

Vamos a trabajar con la variable $n\tau_\sigma$, ya que al ser suma de n variables aleatorias independientes Bernoulli de parámetro $\theta = 1 - \frac{1}{J}$, conocemos que $n\tau_\sigma$ es Binomial de parámetros n y $\theta = 1 - \frac{1}{J}$:

$$n\tau_\sigma \sim \mathcal{B}(n, \theta)$$

Por lo tanto $\quad \mathbb{E}(\tau_\sigma) = 1 - \frac{1}{J} \quad$ y $\quad \mathbb{V}(\tau_\sigma) = \frac{1}{n}\frac{1}{J}\left(1 - \frac{1}{J}\right)$

y es sabido que para n grande su distribución se aproxima a la distribución gaussiana:

$$n\tau_\sigma \sim \mathcal{N}\left(n\theta, n\theta(1-\theta)\right)$$

Observación 2.2.1 *1. Notar que $n\tau_{\sigma_1}, n\tau_{\sigma_2}, \ldots, n\tau_{\sigma_{J!}}$ no son independientes, pues $\sum_{j=1}^{J!} n\tau_{\sigma_j} = n(J-1)!(J-1)$.*

Esto se debe a que:

$$n\sum_{j=1}^{J!} \tau_{\sigma_j} = n\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{\{y_i \neq \sigma_1(\hat{y}_i)\}} + \cdots + \mathbf{1}_{\{y_i \neq \sigma_{J!}(\hat{y}_i)\}}\right) =$$
$$\sum_{i=1}^{n}\left[\#\left\{\sigma \in S_J : \sigma(\hat{y}_i) \neq y_i\right\}\right]$$

Como para y_i y $\hat{y}_i$ fijos $\#\{\sigma \in S_J : \sigma(a) \neq b\} =$
$\#S_J - \#\{\sigma \in S_J : \sigma(a) = b\} = J! - (J-1)! = (J-1)!(J-1)$

se deduce que $n\sum_{j=1}^{J!} \tau_{\sigma_j} = n(J-1)!(J-1)$.

2. *Por otro lado, podemos calcular la varianza de cada $n\tau_\sigma$ de la siguiente manera, ya que usaremos procedimientos similares en cuentas ulteriores:*

$$\mathbb{V}(n\tau_\sigma) = \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}}\right)^2 - \left[\mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}}\right)\right]^2 =$$

$$\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}}\mathbf{1}_{\{y_j \neq \sigma(\hat{y}_j)\}}\right) - n^2\mathbb{P}\left(y_i \neq \sigma(\hat{y}_i)\right)^2 =$$

$$(n^2-n)\mathbb{P}\big(y_i \neq \sigma(\hat{y}_i)\big)\mathbb{P}\big(y_j \neq \sigma(\hat{y}_j)\big) + n\mathbb{P}\big(y_i \neq \sigma(\hat{y}_i)\big)$$

$$-n^2\mathbb{P}\big(y_i \neq \sigma(\hat{y}_i)\big)^2 = (n^2-n)\left(1-\tfrac{1}{J}\right)^2 + n\big(1-\tfrac{1}{J}\big) - n^2\big(1-\tfrac{1}{J}\big)^2 =$$

$$n\big(1-\tfrac{1}{J}\big)\left[1-\big(1-\tfrac{1}{J}\big)\right] = n\big(1-\tfrac{1}{J}\big)\tfrac{1}{J}.$$

2.3. El índice MCE y el caso de J=2 grupos

2.3.1. Suma, esperanza y varianza de CE

Si $J = 2$, los valores de $n\tau$ son naturales en el rango $\left[0, \frac{n}{2}\right]$ y las permutaciones involucradas son $\sigma_1 = Id$ y $\sigma_2 = (12)$. A partir de la observación de la sección anterior se tiene que:

- $n\tau_{\sigma_1} + n\tau_{\sigma_2} = n, \quad \text{con} \quad n\tau_{\sigma_1}, n\tau_{\sigma_2} \sim \mathcal{B}(n, 1/2).$
- $\mathbb{E}(n\tau_\sigma) = \frac{n}{2} \qquad \text{y} \qquad \mathbb{V}(n\tau_\sigma) = \frac{n}{4}.$

2.3.2. Covarianzas de CE

$$\mathbb{COV}(n\tau_{\sigma_1}, n\tau_{\sigma_2}) = \mathbb{E}(n\tau_{\sigma_1}, n\tau_{\sigma_2}) - \mathbb{E}(n\tau_{\sigma_1})\mathbb{E}(n\tau_{\sigma_2}) =$$

$$\mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma_1(\hat{y}_i)\}} \sum_{j=1}^{n} \mathbf{1}_{\{y_j \neq \sigma_2(\hat{y}_j)\}}\right) - \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma_1(\hat{y}_i)\}}\right)\mathbb{E}\left(\sum_{j=1}^{n} \mathbf{1}_{\{y_j \neq \sigma_2(\hat{y}_j)\}}\right) =$$

$$\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{1}_{\{y_i \neq \hat{y}_i\}}\mathbf{1}_{\{y_j = \hat{y}_j\}}\right) - \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \hat{y}_i\}}\right)\mathbb{E}\left(\sum_{j=1}^{n} \mathbf{1}_{\{y_j = \hat{y}_j\}}\right) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left(\mathbf{1}_{\{y_i \neq \hat{y}_i, y_j = \hat{y}_j\}}\right) - n^2\mathbb{P}(y_i \neq \hat{y}_i)\mathbb{P}(y_j = \hat{y}_j) =$$

$(n^2-n)\mathbb{P}(y_i \neq \hat{y}_i, y_j = \hat{y}_j) - n^2\mathbb{P}(y_i \neq \hat{y}_i)\mathbb{P}(y_j = \hat{y}_j)$,

porque si $i = j : \{y_i \neq \hat{y}_i, y_j = \hat{y}_j\} = \emptyset$.

$\mathbb{COV}(n\tau_{\sigma_1}, n\tau_{\sigma_2}) = (n^2-n)\mathbb{P}(y_i \neq \hat{y}_i)\mathbb{P}(y_j = \hat{y}_j) - n^2\mathbb{P}(y_i \neq \hat{y}_i)\mathbb{P}(y_j = \hat{y}_j) =$

$(n^2-n)\left(1-\frac{1}{J}\right)\frac{1}{J} - n^2\left(1-\frac{1}{J}\right)\frac{1}{J} = -n\left(1-\frac{1}{J}\right)\frac{1}{J} = \frac{-n}{4}$.

Por lo que la matriz de correlaciones entre $n\tau_{\sigma_1}$ y $n\tau_{\sigma_2}$ es $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$,

lo cual es acorde a que $\tau_{\sigma_1} + \tau_{\sigma_2} = 1$, ya que tenemos relación de linealidad perfecta.

2.3.3. Probabilidad conjunta y función de distribución del MCE

De la expresión $n\tau_{\sigma_1} + n\tau_{\sigma_2} = n$, si $n\tau_{\sigma_1} = x$ y $n\tau_{\sigma_2} = y$ se tiene que:

$$\mathbb{P}(n\tau_{\sigma_1} = x, n\tau_{\sigma_2} = y) = \begin{cases} 0 & \text{si } x+y \neq n \\ \mathbb{P}(n\tau_{\sigma_1} = x | n\tau_{\sigma_2} = n-x) & \\ \mathbb{P}(n\tau_{\sigma_2} = n-x) = & \\ \mathbb{P}(n\tau_{\sigma_2} = n-x) = \binom{n}{x}\left(\frac{1}{2}\right)^n & \text{si } x+y = n \end{cases}$$

A partir de lo cual se obtiene la siguiente tabla de probabilidad conjunta para $n\tau_{\sigma_1}$ y $n\tau_{\sigma_2}$:

$n\tau_{\sigma_1}/n\tau_{\sigma_2}$	0	1	2	$\cdots$	$n-2$	$n-1$	n
0	0	0	0	0	0	0	$\binom{n}{0}\left(\frac{1}{2}\right)^n$
1	0	0	0	0	0	$\binom{n}{1}\left(\frac{1}{2}\right)^n$	0
2	0	0	0	0	$\binom{n}{2}\left(\frac{1}{2}\right)^n$	0	0
$\vdots$	0	0	0	$\cdots$	0	0	0
$n-2$	0	0	$\binom{n}{n-2}\left(\frac{1}{2}\right)^n$	0	0	0	0
$n-1$	0	$\binom{n}{n-1}\left(\frac{1}{2}\right)^n$	0	0	0	0	0
n	$\binom{n}{n}\left(\frac{1}{2}\right)^n$	0	0	0	0	0	0

.

Veamos cuál es la distribución de $n\tau = \text{mín}\{n\tau_{\sigma_1}, n\tau_{\sigma_2}\}$.

Tenemos que $\mathbb{P}(n\tau = z) =$

$\mathbb{P}(n\tau_{\sigma_1} = z, n\tau_{\sigma_2} = n - z) + \mathbb{P}(n\tau_{\sigma_2} = z, n\tau_{\sigma_1} = n - z) =$

$2\mathbb{P}(n\tau_{\sigma_1} = z, n\tau_{\sigma_2} = n - z) = 2\binom{n}{z}\left(\frac{1}{2}\right)^n$, para $z \neq \frac{n}{2}$.

Si n es par, $\mathbb{P}\left(n\tau = \frac{n}{2}\right) = \mathbb{P}\left(n\tau\sigma_1 = \frac{n}{2}, n\tau\sigma_2 = \frac{n}{2}\right) = \binom{n}{n/2}\left(\frac{1}{2}\right)^n$

$$\Rightarrow \mathbb{P}(n\tau = z) = \begin{cases} \binom{n}{n/2}\left(\frac{1}{2}\right)^n & \text{si } z = \frac{n}{2} \text{ y } n \text{ es par} \\ 2\binom{n}{z}\left(\frac{1}{2}\right)^n & \text{en otro caso} \end{cases}$$

Entonces, la función de distribución de $n\tau$ es:

$F_{n\tau}(z) = \mathbb{P}(n\tau \leq z) = \mathbb{P}(\text{mín}\{n\tau_{\sigma_1}, n\tau_{\sigma_1}\} \leq z) =$

$1 - \mathbb{P}(\text{mín}\{n\tau_{\sigma_1}, n\tau_{\sigma_1}\} \geq z + 1) =$

$1 - \mathbb{P}(n\tau_{\sigma_1} \geq z + 1, n\tau_{\sigma_2} \geq z + 1)$ y queda definida por:

$$\Rightarrow F_{n\tau}(z) = \mathbb{P}(n\tau \leq z) = \begin{cases} \sum_{i=0}^{z} 2\binom{n}{i}\left(\frac{1}{2}\right)^n = 2\Phi(z) & \text{si } z+1 \leq \frac{n}{2} \\ 1 & \text{si } z > \frac{n}{2} \end{cases}$$

con $\Phi = F_{\mathcal{B}(n,1/2)}$ la función de distribución binomial de parámetros n y $\frac{1}{2}$.

Observación 2.3.1 *Es fácil ver que* $F_{n\tau}(z) = F_{\tau}(z/n)$

La Figura 2.1 muestra la función de distribución $F_{n\tau}$ en el caso de simular 1000 observaciones del índice $n\tau$ con $n = 100$ observaciones y $J = 2$ grupos.

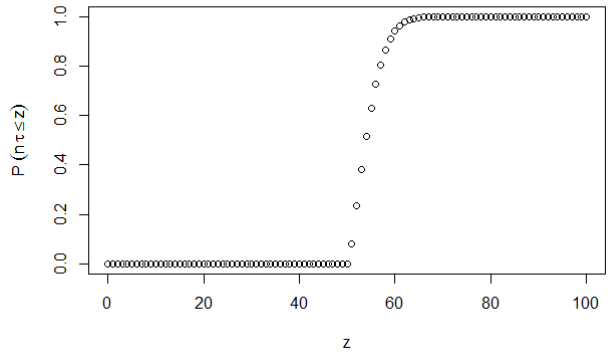

Figura 2.1: Distribución $F_{n\tau}$ de $n\tau$, con $n = 100$ observaciones y $J = 2$ grupos.

2.3.4. Esperanza del MCE

El cálculo de la esperanza de $n\tau$ difiere según si n es par o impar.

- Si n es par

$$\mathbb{E}(n\tau) = \sum_{k=0}^{n/2-1} 2\left(\frac{1}{2}\right)^n k\binom{n}{k} + \frac{n}{2}\left(\frac{1}{2}\right)^n \binom{n}{n/2}$$

El primer término es:

$$2\left(\tfrac{1}{2}\right)^n \sum_{k=1}^{n/2-1} k\binom{n}{k} = \left(\tfrac{1}{2}\right)^{n-1} \sum_{k=1}^{n/2-1} kn\frac{\binom{n-1}{k-1}}{k} = \left(\tfrac{1}{2}\right)^{n-1} n \sum_{k=1}^{n/2-1} \binom{n-1}{k-1}$$

$$= \left(\tfrac{1}{2}\right)^{n-1} n \sum_{k=0}^{n/2-2} \binom{n-1}{k} = \left(\tfrac{1}{2}\right)^{n-1} n \left(\frac{2^{n-1}}{2} - \binom{n-1}{n/2-1}\right),$$

porque: $\sum_{k=0}^{n-1}\binom{n-1}{k} = \sum_{k=0}^{n/2-1}\binom{n-1}{k} + \sum_{k=n/2}^{n-1}\binom{n-1}{k} = 2^{n-1}$ y

$\sum_{k=0}^{n/2-1}\binom{n-1}{k} = \sum_{k=n/2}^{n-1}\binom{n-1}{k} = \frac{2^{n-1}}{2}$.

Entonces,

$$\mathbb{E}(n\tau) = \left(\tfrac{1}{2}\right)^{n-1} n\left(\frac{2^{n-1}}{2} - \binom{n-1}{n/2-1}\right) + \frac{n}{2}\left(\tfrac{1}{2}\right)^{n}\binom{n}{n/2} =$$

$$\frac{n}{2} - \left(\tfrac{1}{2}\right)^{n-1} n\binom{n-1}{n/2-1} + \frac{n}{2}\left(\tfrac{1}{2}\right)^{n}\binom{n}{n/2} =$$

$$\frac{n}{2} - \left(\tfrac{1}{2}\right)^{n-1}\binom{n}{n/2}\frac{n}{2} + \left(\tfrac{1}{2}\right)^{n}\binom{n}{n/2}\frac{n}{2} = \frac{n}{2} - \left(\tfrac{1}{2}\right)^{n}\binom{n}{n/2}\frac{n}{2}.$$

- Cuando n es impar

$$\mathbb{E}(n\tau) = \sum_{k=0}^{\frac{n-1}{2}} k\mathbb{P}(n\tau = k) = \sum_{k=0}^{\frac{n-1}{2}} 2\left(\tfrac{1}{2}\right)^{n} k\binom{n}{k} =$$

$$\left(\tfrac{1}{2}\right)^{n-1}\sum_{k=0}^{\frac{n-1}{2}} k\binom{n}{k} = \left(\tfrac{1}{2}\right)^{n-1}\sum_{k=0}^{\frac{n-1}{2}} n\binom{n-1}{k-1} = \left(\tfrac{1}{2}\right)^{n-1} n\sum_{k=0}^{\frac{n-1}{2}-1}\binom{n-1}{k} =$$

$$\left(\tfrac{1}{2}\right)^{n-1} n\left[2^{n-2} - \tfrac{1}{2}\binom{n-1}{\frac{n-1}{2}}\right],$$

porque $\sum_{k=0}^{n-1}\binom{n-1}{k} = \sum_{k=0}^{\frac{n-1}{2}-1}\binom{n-1}{k} + \binom{n-1}{\frac{n-1}{2}} + \sum_{k=\frac{n-1}{2}+1}^{n-1}\binom{n-1}{k} =$

$2\sum_{k=0}^{\frac{n-1}{2}-1}\binom{n-1}{k} + \binom{n-1}{\frac{n-1}{2}} = 2^{n-1}$

Entonces: $\sum_{k=0}^{\frac{n-1}{2}-1}\binom{n-1}{k} = 2^{n-2} - \frac{1}{2}\binom{n-1}{\frac{n-1}{2}}$.

Por lo tanto,

$$\mathbb{E}(n\tau) = \frac{n}{2} - \left(\frac{1}{2}\right)^{n} n\binom{n-1}{\frac{n-1}{2}}$$

En resumen:

$$\Rightarrow \mathbb{E}(n\tau) = \begin{cases} \frac{n}{2} - (\frac{1}{2})^{n}\binom{n}{n/2}\frac{n}{2} & \text{si } n \text{ es par} \\ \frac{n}{2} - (\frac{1}{2})^{n}\binom{n-1}{\frac{n-1}{2}} n & \text{si } n \text{ es impar} \end{cases}$$

2.4. El MCE en el caso de J=3 grupos

Calculamos la covarianza entre $n\tau_{\sigma_k}$ y $n\tau_{\sigma_l}$, con $k, l \in \{1, \ldots, 6\}$

$$\mathbb{COV}(n\tau_{\sigma_k}, n\tau_{\sigma_l}) = \mathbb{E}(n\tau_{\sigma_k}, n\tau_{\sigma_l}) - \mathbb{E}(n\tau_{\sigma_k})\mathbb{E}(n\tau_{\sigma_l}) =$$

$$\mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma_k(\hat{y}_i)\}} \sum_{j=1}^{n} \mathbf{1}_{\{y_j \neq \sigma_l(\hat{y}_j)\}}\right) - \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma_k(\hat{y}_i)\}}\right) \mathbb{E}\left(\sum_{j=1}^{n} \mathbf{1}_{\{y_j \neq \sigma_l(\hat{y}_j)\}}\right) =$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{P}\big(y_i \neq \sigma_k(\hat{y}_i), y_j \neq \sigma_l(\hat{y}_j)\big) - n^2\big(1 - 1/J\big)^2.$$

Se tiene que:

$$\mathbb{P}\big(y_i \neq \sigma_k(\hat{y}_i), y_j \neq \sigma_l(\hat{y}_j)\big) = 1 - \mathbb{P}([y_i = \sigma_k(\hat{y}_i)] \cup [y_j = \sigma_l(\hat{y}_j)]) =$$

$$1 - \mathbb{P}\big(y_i = \sigma_k(\hat{y}_i)\big) - \mathbb{P}\big(y_j = \sigma_l(\hat{y}_j)\big) + \mathbb{P}\big(y_i = \sigma_k(\hat{y}_i), y_j = \sigma_l(\hat{y}_j)\big)$$

Para $i \neq j$, por independencia:

$$\mathbb{P}\left(y_i = \sigma_k(\hat{y}_i), y_j = \sigma_l(\hat{y}_j)\right) = \mathbb{P}\left(y_i = \sigma_k(\hat{y}_i)\right) \mathbb{P}\left(y_j = \sigma_l(\hat{y}_j)\right) = 1/J^2$$

Para $i = j$:

$$\mathbb{P}\left(y_i = \sigma_k(\hat{y}_i), y_i = \sigma_l(\hat{y}_i)\right) = \sum_{j=1}^{J} \mathbb{P}(y_i = j, \sigma_k(\hat{y}_i) = j, \sigma_l(\hat{y}_i) = j) =$$

$$\sum_{j=1}^{J} \mathbb{P}\left(y_i = j\right) \underbrace{\mathbb{P}\left(\sigma_k(\hat{y}_i) = j, \sigma_l(\hat{y}_i) = j\right)}_{(*)}$$

El término señalado con $(*)$ puede tomar dos valores distintos según las permutaciones σ_k y σ_l que se consideren. Las permutaciones de $\{1, 2, 3\}$ son $\sigma_1 = id, \sigma_2 = (23), \sigma_3 = (12), \sigma_4 = (123), \sigma_5 = (132)$ y $\sigma_6 = (13)$. Se debe tener en cuenta las coincidencias entre las permutaciones, en cuanto a la correspondencia de etiquetas, por ejemplo σ_2 y σ_4 coinciden en hacer corresponder la etiqueta 1 con la 2. El total de coincidencias identificadas en las permutaciones son las siguientes:

$$\sigma_3, \sigma_4 : 1 \to 2 \quad \sigma_5, \sigma_6 : 1 \to 3 \quad \sigma_1, \sigma_2 : 1 \to 1$$

$$\sigma_3, \sigma_5 : 2 \to 1 \quad \sigma_1, \sigma_6 : 2 \to 2 \quad \sigma_2, \sigma_4 : 2 \to 3$$

$$\sigma_1, \sigma_3 : 3 \to 3 \quad \sigma_2, \sigma_5 : 3 \to 2 \quad \sigma_4, \sigma_6 : 3 \to 1$$

Entonces,

$$\mathbb{COV}(n\tau_{\sigma_k}, n\tau_{\sigma_l}) = (n^2-n)\left(1 - \frac{1}{J} - \frac{1}{J} + \frac{1}{J^2}\right) + n\left(1 - \frac{1}{J} - \frac{1}{J} + (*)\right) - n^2\left(1 - \frac{1}{J}\right)^2$$

donde $(*) = \mathbb{P}(y_i = \sigma_k(\hat{y}_i), y_i = \sigma_l(\hat{y}_i)) = \sum_{j=1}^{3} \mathbb{P}(y_i = j)[\mathbb{P}(\sigma_k(\hat{y}_i) = j, \sigma_l(\hat{y}_i) = j)] =$

$$\sum_{j=1}^{3} \mathbb{P}(y_i = j)\left[\sum_{j'=1}^{3} \mathbb{P}(\hat{y}_i = j', \sigma_k(j') = j, \sigma_l(j') = j)\right].$$

El término entre corchetes es igual a $1/J^2$ cuando las permutaciones coinciden en la correspondencia de etiquetas y vale 0 si no coinciden (porque el conjunto $\{\hat{y}_i = j', \sigma_k(j') = j, \sigma_l(j') = j\} = \emptyset$). Por ejemplo, se tiene que:

$$\mathbb{COV}(n\tau_{\sigma_2}, n\tau_{\sigma_4}) = 0 \quad \text{y} \quad \mathbb{COV}(n\tau_{\sigma_2}, n\tau_{\sigma_3}) = -n/J^2.$$

La Tabla 2.2 muestra la estimación de la matriz de correlaciones para $J = 3$ con $n = 500$ observaciones.

	$n\tau_{\sigma_1}$	$n\tau_{\sigma_2}$	$n\tau_{\sigma_3}$	$n\tau_{\sigma_4}$	$n\tau_{\sigma_5}$	$n\tau_{\sigma_6}$
$n\tau_{\sigma_1}$	1.00	-0.01	-0.06	-0.50	-0.50	0.06
$n\tau_{\sigma_2}$	-0.01	1.00	-0.47	-0.01	0.02	-0.53
$n\tau_{\sigma_3}$	-0.06	-0.47	1.00	0.04	0.02	-0.50
$n\tau_{\sigma_4}$	-0.50	-0.01	0.04	1.00	-0.50	-0.03
$n\tau_{\sigma_5}$	-0.50	0.02	0.02	-0.50	1.00	-0.03
$n\tau_{\sigma_6}$	0.06	-0.53	-0.50	-0.03	-0.03	1.00

Tabla 2.2: Matriz de correlaciones para particiones de $n = 500$ observaciones y $J = 3$ grupos, sobre 1000 repeticiones de cada τ_{σ_j}.

Notar que:

$$\frac{\frac{-n}{J^2}}{n\left(1-\frac{1}{J}\right)\frac{1}{J}} = \frac{\frac{-500}{9}}{500 \times 2/9} = -0,5; \qquad n\left(1 - \frac{1}{J}\right)\frac{1}{J} = 500 \times 2/9 = 111,111$$

lo cual muestra la coherencia con el resultado encontrado anteriormente. En la Figura 2.2 ilustramos a través de un *scatterplot* estas correlaciones.

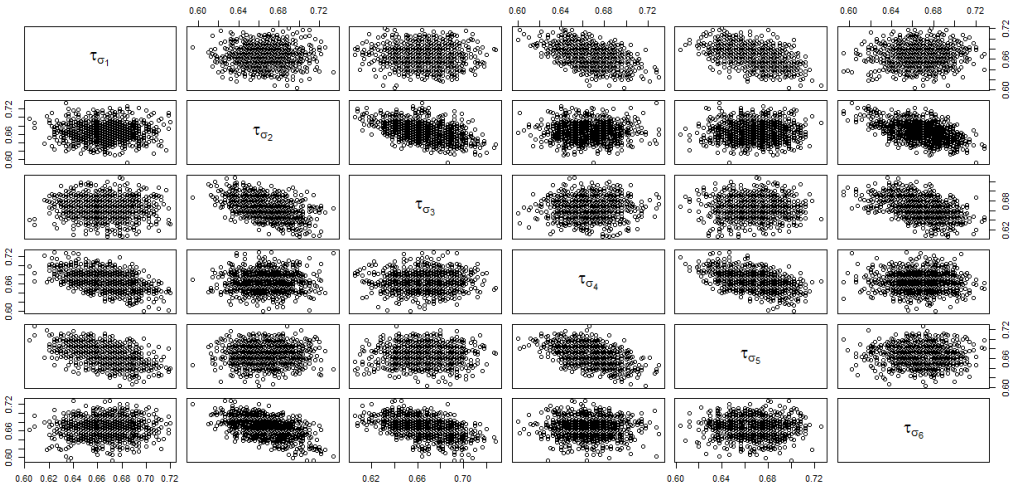

Figura 2.2: Correlaciones entre $n\tau_{\sigma_1}, n\tau_{\sigma_2}, n\tau_{\sigma_3}, n\tau_{\sigma_4}, n\tau_{\sigma_5}, n\tau_{\sigma_6}$, para particiones de $n = 500$ observaciones y $J = 3$ grupos, sobre 1000 repeticiones de cada $n\tau_{\sigma_j}$.

Para determinar la distribución del mínimo debemos hallar la distribución conjunta de $(n\tau_{\sigma_1}, n\tau_{\sigma_2}, n\tau_{\sigma_3}, n\tau_{\sigma_4}, n\tau_{\sigma_5}, n\tau_{\sigma_6})$ sujeto a que:

- $n\tau_{\sigma_1} + n\tau_{\sigma_2} + n\tau_{\sigma_3} + n\tau_{\sigma_4} + n\tau_{\sigma_5} + n\tau_{\sigma_6} = 4n$ y a que
- $\mathbb{COR}(n\tau_{\sigma_2}, n\tau_{\sigma_6}) = \mathbb{COR}(n\tau_{\sigma_3}, n\tau_{\sigma_2}) =$
  $\mathbb{COR}(n\tau_{\sigma_4}, n\tau_{\sigma_1}) = \mathbb{COR}(n\tau_{\sigma_5}, n\tau_{\sigma_1}) =$
  $\mathbb{COR}(n\tau_{\sigma_5}, n\tau_{\sigma_4}) = \mathbb{COR}(n\tau_{\sigma_6}, n\tau_{\sigma_3}) = -0.5$
  y $\mathbb{COR}(n\tau_{\sigma_i}, n\tau_{\sigma_j}) = 0$ en otros casos.

Esta probabilidad conjunta es objeto de estudio actualmente.

2.5. Simulaciones

En esta sección se estudia la distribución empírica del MCE entre dos particiones $\mathscr{C}$ y $\mathscr{C}'$. Se consideran y comparan los resultados bajo distintas condiciones experimentales: diferente cantidad de grupos (J), diferente cantidad de observaciones (n), particiones independientes o con distintos grados de dependencia y balance o desbalance en el tamaño relativo de los grupos.

Para cada uno de los cuatro escenarios definidos según condiciones de independecia/dependencia de las particiones y balance/desbalance de los grupos se describe el proceso de simulación, una presentación gráfica de la distribución del índice y las medias obtenidas para cada valor de J y n.

2.5.1. Particiones independientes con clases balanceadas

Para el caso de particiones independientes con J clases balanceadas se simulan directamente los vectores de particiones $\mathscr{C}$ y $\mathscr{C}'$, que contienen las etiquetas de cada observación. Estas podrían ser el resultado de un análisis de *clustering*.

El procedimiento de simulación utilizado fue el siguiente: se genera un vector fijo de etiquetas equidistribuidas $\mathscr{C}_{aux}$, en cada paso se simula $\mathscr{C}$ como un reordenamiento aleatorio de $\mathscr{C}_{aux}$ y $\mathscr{C}'$ como un reordenamiento aleatorio del $\mathscr{C}$. De esta manera los pares $(\mathscr{C}, \mathscr{C}')$ son independientes entre sí.

Realizamos simulaciones para J cantidad de clases y n número de observaciones, con $J \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ y $n \in \{50, 100, 200, 300, 400, 500, 1000\}$. Para cada valor de J y n se generan $N = 1000$ parejas de particiones $(\mathscr{C}, \mathscr{C}')$ independientes y se calcula el índice $\tau = MCE(\mathscr{C}, \mathscr{C}')$. Así se obtienen 1000 valores del índice para cada valor de J y n.

La Figura 2.3 muestra el gráfico de la densidad del $MCE(\mathscr{C}, \mathscr{C}')$ estimada a partir de $N = 1000$ valores del índice, para distintos valores de J y con $n = 1000$ observaciones. Vemos que la forma de la densidad es más simétrica cuando aumenta la cantidad de grupos J.

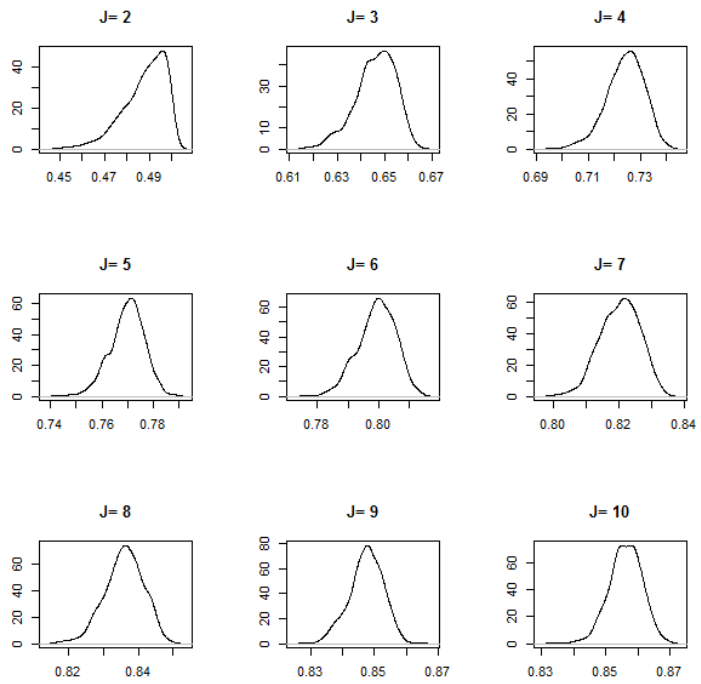

Figura 2.3: Estimación de la densidad del $MCE(\mathscr{C}, \mathscr{C}')$ con $N = 1000$ valores del índice, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones independientes de $n = 1000$ observaciones con J clases balanceadas.

En la Tabla 2.3 se presentan las medias del $MCE(\mathscr{C}, \mathscr{C}')$ para cada valor de n y J sobre las $N = 1000$ repeticiones. Se observa que estos valores aumentan con la cantidad de clases J y con el tamaño de muestra n.

	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	0.44	0.57	0.63	0.66	0.67	0.68	0.68	0.68	0.68
n=100	0.46	0.60	0.67	0.70	0.72	0.74	0.74	0.75	0.75
n=200	0.47	0.62	0.69	0.73	0.76	0.77	0.78	0.79	0.80
n=300	0.48	0.63	0.70	0.74	0.77	0.79	0.80	0.81	0.82
n=400	0.48	0.63	0.71	0.75	0.78	0.80	0.81	0.82	0.83
n=500	0.48	0.64	0.71	0.76	0.79	0.81	0.82	0.83	0.84
n=1000	0.49	0.65	0.72	0.77	0.80	0.82	0.84	0.85	0.86

Tabla 2.3: Medias del $MCE(\mathscr{C}, \mathscr{C}')$ con $N = 1000$ valores del índice, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones independientes de n observaciones con J clases balanceadas.

2.5.2. Particiones independientes con clases desbalanceadas

Para simular particiones independientes con J clases desbalanceadas simplemente se generan (de manera independiente) pares de vectores $(\mathscr{C}, \mathscr{C}')$ a partir de la distribución binomial $\mathcal{B}(J-1, 0.6)$, donde J se corresponde con la cantidad de clases de cada partición. Como los números generados aleatoriamente a partir de la distribución binomial incluyen al cero (dando como resultado valores en el rango $[0, J-1]$), para obtener el vector de etiquetas $\{1, 2 \ldots, J\}$ sumamos a las simulaciones el valor 1. La Figura 2.4 y la Tabla 2.4 muestran que los resultados son similares a los observados en el caso balanceado aunque con magnitudes menores en el valor del índice.

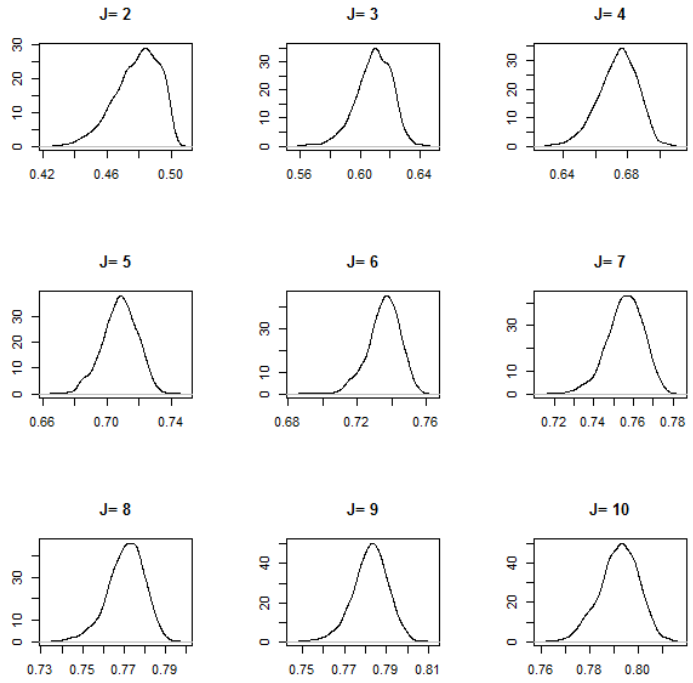

Figura 2.4: Estimación de la densidad del $MCE(\mathscr{C}, \mathscr{C}')$ con $N = 1000$ valores del índice, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones independientes de $n = 1000$ observaciones con J clases desbalanceadas.

	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	0.44	0.56	0.61	0.64	0.65	0.66	0.67	0.68	0.68
n=100	0.46	0.58	0.64	0.67	0.69	0.70	0.71	0.72	0.73
n=200	0.47	0.60	0.66	0.69	0.71	0.73	0.74	0.75	0.76
n=300	0.47	0.60	0.66	0.70	0.72	0.74	0.75	0.76	0.77
n=400	0.47	0.60	0.67	0.70	0.72	0.74	0.76	0.77	0.78
n=500	0.48	0.60	0.67	0.70	0.73	0.75	0.76	0.77	0.78
n=1000	0.48	0.61	0.67	0.71	0.74	0.76	0.77	0.78	0.79

Tabla 2.4: Medias del $MCE(\mathscr{C}, \mathscr{C}')$ con $N = 1000$ valores del índice, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones independientes de n observaciones con J clases desbalanceadas.

2.5.3. Particiones dependientes con clases balanceadas

El siguiente escenario analizado es la comparación de particiones dependientes con clases balanceadas. Para esto se simulan vectores de observaciones con diferentes grados de dependencia entre sí.

El procedimiento de simulación parte de vectores de etiquetas iguales $\mathscr{C} = \mathscr{C}'$, con J grupos de igual tamaño. Luego se modifica aleatoriamente una proporción de las etiquetas etiquetas de $\mathscr{C}'$. La modificación varía en proporciones $p \in \{0.1, 0.4, 0.6, 0.9\}$, por ejemplo cuando $p = 0.40$ y $n = 100$ la segunda partición podrá diferir de la primera en las etiquetas de 40 observaciones. Se realizan $N = 1000$ repeticiones de las muestras dependientes para cada valor de p, J y n. En la Figura 2.5 y en la Tabla 2.5 se presentan los resultados de la distribución para $p = 0.4$, que se corresponden con un grado de dependencia moderada de las particiones. Observando los resultados constatamos, como es esperar, que el error de clasificación es menor al obtenido en la comparación de particiones independientes. Los valores del índice MCE aumentan en promedio cuando aumenta p, ya que es el parámetro que determina la diferencia entre las particiones comparadas. Estos resultados no se exponen en este trabajo por motivos de espacio.

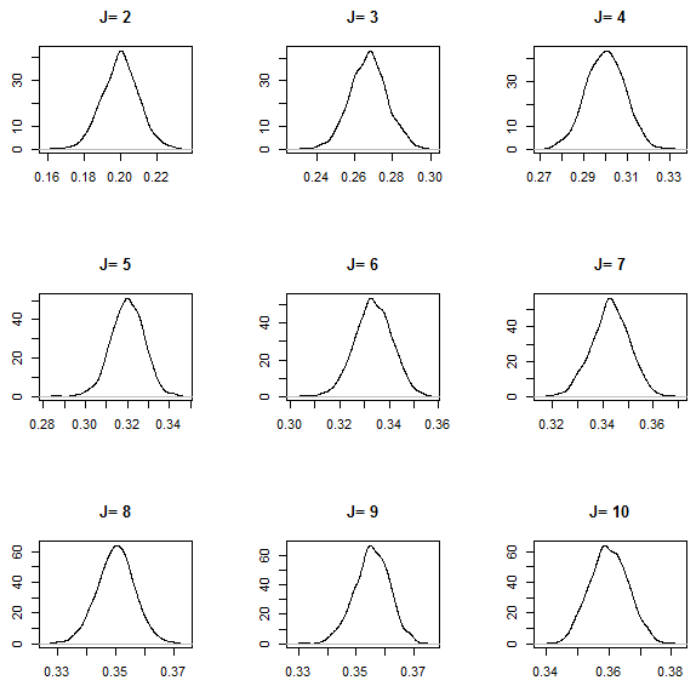

Figura 2.5: Estimación de la densidad del $MCE(\mathscr{C},\mathscr{C}')$ con $N = 1000$ valores del índice, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones dependientes ($p = 0.4$) de $n = 1000$ observaciones con J clases balanceadas.

	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	0.202	0.263	0.299	0.320	0.334	0.343	0.351	0.355	0.360
n=100	0.201	0.267	0.299	0.320	0.333	0.343	0.351	0.353	0.361
n=200	0.200	0.268	0.301	0.320	0.334	0.343	0.349	0.357	0.360
n=300	0.201	0.266	0.301	0.320	0.333	0.343	0.350	0.355	0.360
n=400	0.200	0.267	0.300	0.320	0.333	0.343	0.350	0.355	0.360
n=500	0.200	0.266	0.300	0.320	0.333	0.343	0.350	0.356	0.360
n=1000	0.200	0.267	0.300	0.320	0.334	0.343	0.350	0.355	0.360

Tabla 2.5: Medias del $MCE(\mathscr{C},\mathscr{C}')$ con $N = 1000$ valores del índice, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones dependientes ($p = 0.4$) de n observaciones con J clases balanceadas.

2.5.4. Particiones dependientes con clases desbalanceadas

Para la generación de particiones dependientes con clases de diferente tamaño se combinan los mecanismos de simulación de los escenarios anteriores. El procedimiento parte de dos vectores de etiquetas iguales ($\mathscr{C} = \mathscr{C}'$), generados a partir de la distribución binomial para garantizar clases de diferente

tamaño (como fue descrito en la página 36). Luego se modifica aleatoriamente una proporción de las etiquetas de $\mathscr{C}'$. La modificación varía en proporciones $p = \{0.1, 0.4, 0.6, 0.9\}$ (como en la página 37). Se realizan $N = 1000$ repeticiones de las muestras dependientes para cada valor de p, J y n.

Al igual que en el ejemplo anterior, se muestra el gráfico de densidad (Figura 2.6) en el caso de $n = 1000$ observaciones y la media del índice (Tabla 2.6) calculada para cada combinación de J y n, ambos para un grado de dependencia determinado en la simulación por la proporción $p = 0.4$. Se observa que los valores medios son similares a los resultantes en particiones con clases balanceadas.

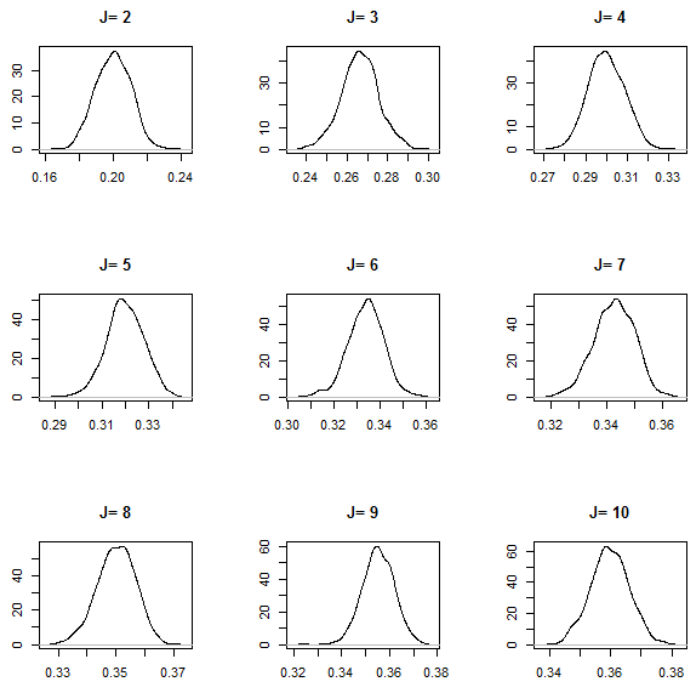

Figura 2.6: Estimación de la densidad del $MCE(\mathscr{C}, \mathscr{C}')$ con $N = 1000$ valores del índice, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones dependientes ($p = 0.4$) de $n = 1000$ observaciones con J clases desbalanceadas.

	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	0.200	0.264	0.299	0.320	0.330	0.341	0.349	0.353	0.357
n=100	0.200	0.268	0.300	0.321	0.332	0.343	0.349	0.355	0.360
n=200	0.200	0.266	0.301	0.320	0.334	0.344	0.350	0.355	0.360
n=300	0.200	0.267	0.300	0.321	0.333	0.343	0.350	0.356	0.360
n=400	0.201	0.266	0.300	0.320	0.333	0.343	0.350	0.356	0.360
n=500	0.201	0.266	0.300	0.319	0.333	0.343	0.350	0.356	0.360
n=1000	0.200	0.266	0.300	0.320	0.334	0.343	0.350	0.355	0.360

Tabla 2.6: Medias del $MCE(\mathscr{C}, \mathscr{C}')$ con $N = 1000$ valores del índice, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones dependientes ($p = 0.4$) de n observaciones con J clases desbalanceadas.

2.6. Comparación del MCE con otros índices en distintas situaciones

Para la analizar comparativamente los resultados del MCE seleccionamos los índices de Rand y Jaccard, conocidos y popularmente utilizados en comparación de *clustering*. Los resultados se examinan en las distintas condiciones experimentales simuladas en la sección anterior y en un aplicación a un conjunto de datos reales.

2.6.1. Simulaciones

Calculamos los valores de los índices MCE, Rand y Jaccard sobre los datos simulados en la Sección 2.5 y la relación entre los índices es evaluada mediante el coeficiente de correlación de Pearson, calculado con $N = 1000$ valores de los índices.

En las tablas 2.7 a 2.8 se presentan los coeficientes obtenidos para las particiones simuladas con $n = 1000$ observaciones, según la cantidad de clases (J), en los escenarios de particiones con independencia o con dependencia (con $p = 0.4$) y distinguiendo según condiciones de desbalance o balance en los grupos.

		J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10	promedio
Dependencia	Grupos desbalanceados	-1	-0,93	-0,89	-0,83	-0,81	-0,79	-0,73	-0,74	-0,70	-0,83
	Grupos balanceados	-1	-1	-1	-1	-0,99	-0,99	-0,99	-0,99	-0,98	-0,99
Independencia	Grupos desbalanceados	-0,94	-0,28	-0,43	-0,21	-0,21	-0,12	-0,14	-0,13	-0,05	-0,28
	Grupos balanceados	-0,93	-0,86	-0,83	-0,81	-0,8	-0,79	-0,78	-0,78	-0,77	-0,82

Tabla 2.7: Correlaciones entre $MCE(\mathscr{C}, \mathscr{C}')$ y Rand$(\mathscr{C}, \mathscr{C}')$ sobre $N = 1000$ valores de los índices, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones con $J = 2, \ldots, 10$ grupos y $n = 1000$ observaciones, con dependencia/independencia y desbalance/balance en los grupos.

		J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10	promedio
Dependencia	Grupos desbalanceados	-0,99	-0,95	-0,91	-0,87	-0,85	-0,82	-0,78	-0,77	-0,75	-0,86
	Grupos balanceados	-1	-1	-1	-1	-1	-1	-1	-1	-1	-0,99
Independencia	Grupos desbalanceados	-0,58	-0,45	-0,6	-0,52	-0,56	-0,59	-0,57	-0,63	-0,54	-0,56
	Grupos balanceados	-0,93	-0,86	-0,83	-0,81	-0,8	-0,79	-0,78	-0,78	-0,77	-0,82

Tabla 2.8: Correlaciones entre $MCE(\mathscr{C}, \mathscr{C}')$ y Jaccard$(\mathscr{C}, \mathscr{C}')$ sobre $N = 1000$ valores de los índices, donde $\mathscr{C}$ y $\mathscr{C}'$ son particiones con $J = 2, \ldots, 10$ grupos y $n = 1000$ observaciones, con dependencia/independencia y desbalance/balance en los grupos.

En primer lugar se evidencia que la asociación entre los resultados del MCE y los indices de Rand y Jaccard varía según las condiciones experimentales. Cuando las particiones comparadas tienen clases balanceadas, la correlación es negativa (lo cual es de esperar porque MCE mide la "distancia" o disimilaridad entre particiones, mientras que Rand y Jaccard son índices de similaridad) y fuerte. Si se trata de particiones con dependencia la asociación es casi perfecta (cercana a menos uno), mientras que con particiones independientes la asociación es fuerte pero menor (entre -0.77 y -0.93). También se observa que la correlación disminuye cuando aumenta el número de clases J.

En condiciones de grupos desbalanceados los resultados de los índices difieren. Podemos observar que al comparar particiones dependientes las correlaciones —si bien disminuyen respecto al caso de grupos de igual tamaño— tienen coeficiente de entre -0.7 y -0.99. Sin embargo, cuando las particiones son independientes y además desbalanceadas la asociación entre las medidas es sensiblemente menor y, exceptuando el caso con $J = 2$ grupos, es mucho más débil con el índice de Rand que con el índice de Jaccard.

Si observamos los valores medios que se obtienen de cada índice en las simulaciones del caso de dos particiones independientes con grupos desbalanceados (Tabla 2.9), podemos describir la diferencia en el comportamiento de

los índices en este caso particular. En el ejemplo de particiones con $n = 1000$ observaciones, el MCE muestra un error de clasificación alto, lo que es de esperar al comparar particiones que son independientes. El mismo comportamiento puede observarse en los valores medios del índice de similaridad de Jaccard, pero en sentido opuesto al MCE que mide disimilaridad. Sin embargo, el índice de Rand muestra un comportamiento contraintuitivo, con valores del índice altos en promedio. Por este motivo consideramos que el índice de Rand no muestra un buen desempeño para identificar las diferencias de dos particiones independientes, cuando éstas tienen grupos desbalanceados.

	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
MCE	0.48	0.61	0.67	0.71	0.74	0.76	0.77	0.78	0.79
Rand	0.50	0.55	0.62	0.68	0.72	0.75	0.78	0.80	0.81
Jaccard	0.33	0.20	0.14	0.11	0.09	0.08	0.07	0.06	0.06

Tabla 2.9: Medias de los índices $MCE(\mathscr{C}, \mathscr{C}')$, Rand$(\mathscr{C}, \mathscr{C}')$ y Jaccard$(\mathscr{C}, \mathscr{C}')$ con $N = 1000$ valores de los índices, donde $(\mathscr{C}, \mathscr{C}')$ son particiones independientes de $n = 1000$ observaciones con J clases desbalanceadas.

2.6.2. Una aplicación a datos reales (*MNIST*)

La base de datos *MNIST* contiene 70000 imágenes de dígitos escritos a mano. Los datos describen las imágenes en escala de grises de tamaño 28×28 pixeles, como los que muestra la Figura 2.7. Cada una de las $28 \times 28 = 784$ columnas del conjunto de datos se corresponde con el rango del pixel (de 0 a 783) y toman valores entre 0 y 255 , que representan la intensidad del gris. La base contiene una columna adicional con la etiqueta que indica el dígito del 0 a 9 que representa la imagen.

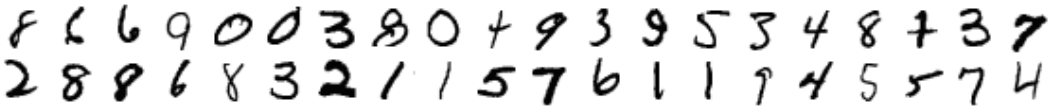

Figura 2.7: Ejemplos de imágenes de 28×28 pixeles de *MNIST*.

Se aplican algoritmos de clasificación no supervisada: 10-*means* y CLARA[1] [12], como en el Capítulo 1, con $J = 10$ clases. Luego se comparan los resultados con la verdadera etiqueta de las imágenes.

[1] Adaptación del método PAM a bases de datos con una gran cantidad de observaciones.

La Tabla 2.10 muestra la matriz de contingencia que surge de comparar el *clustering* obtenido por el método 10-*means* y la verdadera clasificación de los dígitos $\{0, 1, \ldots, 9\}$, donde C_i es el *cluster* emparejado con el grupo del dígito i, con la permutación de etiquetas que minimiza el error de clasificación. Como puede observarse, calculamos la tasa de error de cada dígito a partir de dicha tabla de contingencia. Observamos que ésta oscila alrededor de 0.3 para la mayoría de los dígitos. Los errores relativos más altos están en el dígito 5 —con un error de 0.7— que como puede observarse es clasificado frecuentemente como 3 u 8, y en el dígito 9, que el 90 % de las veces fue asignado al grupo de los dígitos 4 y 7.

dígito	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
C_0	5046	0	56	22	9	60	76	21	37	50
C_1	2	4293	427	457	181	166	201	375	337	267
C_2	9	10	4862	210	29	7	57	53	53	20
C_3	291	8	319	4581	0	2118	38	5	1174	86
C_4	42	6	215	188	3735	425	67	2088	208	3456
C_5	1253	7	238	500	261	1851	1940	12	346	30
C_6	173	7	152	33	170	70	4434	4	54	16
C_7	7	10	79	46	2193	212	4	4405	188	2852
C_8	76	9	205	1051	17	1141	16	20	4115	89
C_9	4	3527	437	53	229	263	43	310	313	92
error	0,27	0,45	0,30	0,36	0,45	0,71	0,36	0,40	0,40	0,99

Tabla 2.10: Tabla de contingencia entre el vector con la verdadera etiqueta de los dígitos y la partición obtenida por el método 10-*means*, donde C_i representa el *cluster* emparejado con el dígito i, según el mínimo error de clasificación.

Los valores de los índices de Rand, Jaccard y MCE en la comparación de los *clustering* y la verdadera etiqueta de los dígitos se presentan en la Tabla 2.11. El error según el índice MCE es menor para la partición obtenida por 10-*means*. Los índices Rand y Jaccard también muestran que la clasificación de 10-*means* es mejor que la resultante del método CLARA.

	MCE	Rand	Jaccard
k-means	0.47	0.88	0.28
CLARA	0.55	0.85	0.19

Tabla 2.11: Índices MCE, Rand y Jaccard en la comparación de las particiones obtenidas por *clustering* con el vector de dígitos de *MNIST*.

Las diferencias entre las magnitudes de los índices al comparar las *clustering* con la verdadera etiqueta pueden interpretarse a partir del análisis de la tabla de contingencia 2.10. Podemos ver que según el índice de Rand las particiones comparadas son bastante similares. Si repasamos la definición, vemos que este índice contabiliza a los pares de observaciones que son clasificados en el mismo grupo (a) y en distintos grupos (b) por ambas particiones, logrando captar por ejemplo agrupamientos de los dígitos 5 y 9 que en el MCE, por su emparejamiento unívoco de los grupos, se computan como errores. El índice de Jaccard desestima la cantidad b en su definición, motivo que explica la diferencia con el índice de Jaccard, si bien ambos miden similaridad.

En este sentido, los índices muestran diferencias que pueden ser provechosas según el contexto y los objetivos de la comparación. Estas diferencias deben ser consideradas al momento de la interpretación de los resultados, para poder evaluar con el índice que más se adecue al propósito del trabajo pudiendo incluso utilizarlos de manera complementaria.

Capítulo 3

Test de hipótesis para particiones independientes a partir del MCE

Uno de los propósitos principales de conocer la distribución teórica del índice es el diseño de pruebas que permitan determinar si dos particiones son estadísticamente equivalentes. Las propiedades que fueron demostradas hasta el momento se derivaron bajo algunos supuestos y nos permiten elaborar un test de hipótesis para particiones independientes, con $J = 2$ grupos igualmente distribuidos. A continuación presentamos el test diseñado basado en el índice MCE y una evaluación del desempeño sobre datos simulados.

En el caso de dos grupos demostramos que si las particiones comparadas $\mathscr{C}$ y $\mathscr{C}'$ son independientes entonces $n\tau_\sigma \sim \mathcal{B}(n, 1 - 1/J)$ y vimos cuál es la distribución de $n\tau$.

Ahora proponemos un test de hipótesis no paramétrico para contrastar si dos particiones son independientes entre sí. Dadas dos particiones $\mathscr{C}$ y $\mathscr{C}'$, el planteo del test es:

(H_0): Las particiones $\mathscr{C}$ y $\mathscr{C}'$ son independientes
(H_1): Las particiones $\mathscr{C}$ y $\mathscr{C}'$ no son independientes

El estadístico de prueba es $n\tau$ cuya distribución bajo H_0 es conocida para

el caso $J = 2$. A partir de que $\mathbb{P}(n\tau \geq q_\alpha) \geq 1 - \alpha$ se propone la región crítica definida por el intervalo $[0, q_\alpha)$, con q_α el cuantil α de la distribución del estadístico. Si $\alpha = 0.05$:

$$F_{n\tau}(q_{0.05}) = \mathbb{P}(n\tau \leq q_{0.05}) \leq 0.05$$

En la Tabla 3.1 se presentan los cuantiles 0.05 de $n\tau$ calculados sobre las 1000 simulaciones de particiones independientes con clases balanceadas. El caso $J = 2$ se puede comparar con el cuantil teórico 0.05 que está en la última columna. La Tabla 3.2 contiene la misma información para el índice τ (MCE), es decir, con el error expresado en proporción de casos.

	Experimental									Teórico
	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10	J=2
n=50	18	25	28	30	31	31	31	31	31	18
n=100	40	55	62	66	68	70	71	71	71	40
n=200	86	116	132	140	146	149	152	154	155	86
n=300	132	181	204	217	225	231	235	238	240	133
n=400	180	244	274	293	304	312	318	322	326	180
n=500	228	308	347	371	385	395	402	408	412	228
n=1000	470	630	712	759	789	811	827	838	847	469

Tabla 3.1: Cuantil 0.05 de la distribución empírica de $n\tau$ ($nMCE$) sobre 1000 simulaciones con $J \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ grupos y $n \in \{50, 100, 200, 300, 400, 500, 1000\}$ observaciones y cuantil 0.05 de la distribución teórica del $n\tau$ para $J = 2$ grupos y $n \in \{50, 100, 200, 300, 400, 500, 1000\}$ (última columna).

	Experimental									Teórico
	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10	J=2
n=50	0.36	0.50	0.56	0.60	0.62	0.62	0.62	0.62	0.62	0.36
n=100	0.40	0.55	0.62	0.66	0.68	0.70	0.71	0.71	0.71	0.40
n=200	0.43	0.58	0.66	0.70	0.73	0.74	0.76	0.77	0.78	0.43
n=300	0.44	0.60	0.68	0.72	0.75	0.77	0.78	0.79	0.80	0.44
n=400	0.45	0.61	0.68	0.73	0.76	0.78	0.80	0.80	0.81	0.45
n=500	0.46	0.62	0.69	0.74	0.77	0.79	0.80	0.82	0.82	0.46
n=1000	0.47	0.63	0.71	0.76	0.79	0.81	0.83	0.84	0.85	0.47

Tabla 3.2: Cuantil 0.05 de la distribución empírica del MCE sobre 1000 simulaciones con $J \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ grupos y $n \in \{50, 100, 200, 300, 400, 500, 1000\}$ observaciones y cuantil 0.05 de la distribución teórica del $n\tau$ para $J = 2$ grupos y $n \in \{50, 100, 200, 300, 400, 500, 1000\}$ (última columna).

3.1. Desempeño experimental del test de hipótesis

Para evaluar el desempeño del test planteado se estudian las tasas de error obtenidas en simulaciones generadas con diferentes condiciones experimentales. Los puntos críticos de la prueba son los estimados en la sección anterior, presentados en la Tabla 3.2. En el caso $J = 2$ se toman los resultados teóricos y en el resto de los casos los estimados de forma empírica.

Se consideran el error tipo I ($= \mathbb{P}_{H_0}(H_1)$) y el error tipo II ($= \mathbb{P}_{H_1}(H_0)$). Estos se calculan como la proporción de casos donde rechazo H_0 suponiendo H_0 cierta y la proporción de casos en los cuales no rechazo H_0 suponiendo H_1 cierta, respectivamente. Para evaluar el error tipo I se generan otras 1000 muestras de particiones $(\mathscr{C}, \mathscr{C}')$ independientes, con n observaciones y J clases balanceadas. En la Tabla 3.3 se presentan los errores tipo I obtenidos al comparar las particiones independientes simuladas según J y n. Este computa alrededor de 5 % en todos los casos, como era de esperar.

	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	3.3	3.0	2.1	4.0	4.3	2.8	1.3	2.4	2.9
n=100	2.3	3.4	3.3	3.5	2.7	4.4	4.3	1.8	4.7
n=200	2.9	2.5	4.6	3.8	4.3	3.6	4.2	4.7	3.3
n=300	5.2	4.3	4.4	6.3	2.7	5.9	3.8	4.4	3.6
n=400	4.6	5.3	5.3	3.9	2.5	4.4	4.1	3.2	5.3
n=500	6.6	2.8	4.2	4.6	5.0	4.4	3.2	4.0	3.9
n=1000	3.8	4.6	4.1	4.5	3.8	4.3	4.5	4.3	3.3

Tabla 3.3: Porcentaje de error tipo I obtenido con el test en $N = 1000$ comparaciones de particiones independientes con clases balanceadas.

Por otra parte, para evaluar el error tipo II, se utilizan las particiones generadas bajo H_1, es decir, con dependencia. Como fue descrito en la Sección 2.5, se simulan particiones con clases balanceadas y diferentes grados de dependencia entre sí. La particiones comparadas varían en proporciones $p = \{0.1, 0.4, 0.6, 0.9\}$. Se utilizan la $N = 1000$ repeticiones de las muestras dependientes generadas para cada valor de p, J y n.

El error tipo II es 0 cuando $p = 0.1$ y $p = 0.4$ (alta dependencia de las particiones). Las Tablas 3.4 y 3.5 refieren a los errores cuando las particiones

son modificadas en una proporción 0.6 y 0.9 de los casos, respectivamente. Puede observarse que, en el primer caso ($p = 0.6$), el error tipo II es siempre 0, salvo en el caso con $n = 50$ y $J \leq 3$, y con $n = 100$ y $J = 2$, donde el error es positivo pero pequeño. Cuando $p = 0.9$ (escenario más cercano a independencia) la proporción de veces que se acepta la hipótesis nula de independencia es considerable. El error de tipo II en este ejemplo disminuye cuando crece n y J.

	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	19.4	3.2	1.5	0.0	0.0	0.0	0.0	0.0	0.0
n=100	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
n=200	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
n=300	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
n=400	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
n=500	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
n=1000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tabla 3.4: Porcentaje de error tipo II obtenido en $N = 1000$ comparaciones de particiones dependientes ($p = 0.6$) con clases balanceadas.

	J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	94.2	94.7	95.4	93.5	93.7	96.2	96.9	97.7	97.8
n=100	87.5	88.0	88.7	89.6	88.8	87.7	88.5	91.2	86.2
n=200	73.7	67.8	66.6	64.2	68.9	63.3	71.0	69.2	70.5
n=300	60.0	53.2	46.1	36.2	40.0	36.9	39.4	37.0	41.1
n=400	46.4	33.6	25.3	21.7	21.5	15.2	18.8	22.8	14.1
n=500	36.9	19.1	14.3	9.8	6.6	8.1	7.1	4.1	5.0
n=1000	10.9	2.0	0.6	0.1	0.0	0.0	0.0	0.0	0.0

Tabla 3.5: Porcentaje de error tipo II obtenido en $N = 1000$ comparaciones de particiones dependientes ($p = 0.9$) con clases balanceadas.

La evaluación experimental muestra un buen desempeño del test de independencia de particiones con clases de igual tamaño. Los errores de tipo I están en el entorno esperado para un test de nivel $\alpha = 0.05$. En la evaluación del error de tipo II el test demostró ser potente cuando la dependencia entre las particiones que se comparan es relativamente alta y que pierde precisión cuando las particiones comparadas tienen condiciones que se aproximan a la hipótesis nula de independencia.

3.2. Comparación con un test para particiones independientes a partir del índice de Rand

Por último nos interesa reflexionar sobre los resultados del test basado en el MCE presentado en este capítulo, en relación con el índice de Rand, que como vimos en el Capítulo 1, también tiene resultados teóricos sobre su distribución demostrados en [10] y citado por [28].

Considerando que la distribución teórica derivada para Rand (ver página 16) es aproximadamente válida en la comparación de particiones independientes con clases equidistribuidas (para J y n grande), podemos plantear un test de manera análoga al test anterior pero utilizando como estadístico de prueba al índice de Rand. En este caso la región crítica queda definida por el intervalo $(q_{1-\alpha}, 1]$, con $q_{1-\alpha}$ el cuantil $1-\alpha$ de la distribución del estadístico, ya que si $\alpha = 0.05$:

$$F_{n\tau}(q_{0.95}) = \mathbb{P}(n\tau \leq q_{0,95}) \leq 1 - 0.05$$

por ser Rand un índice de similaridad a diferencia del MCE.

En la Tabla 3.6 podemos ver que los cuantiles al 95 % de la distribución teórica y empírica de Rand son similares. Sin embargo si consideramos las $N = 1000$ simulaciones de particiones independientes con grupos balanceados (que generamos para evaluar el desempeño del test con el MCE) y calculamos los errores tipo I utilizando el índice Rand observamos lo siguiente: el error es pequeño si consideramos los puntos definidos por la distribución empírica, pero es muy elevado si utilizamos como punto de corte los cuantiles que surgen de la distribución teórica del estadístico de Rand. Esto último se observa también en los casos con particiones con $J > 3$, donde los resultados teóricos sobre la distribución son aproximadamente válidos.

		J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
Teórico	$q_{0.95}$	0.501	0.556	0.626	0.681	0.723	0.756	0.782	0.803	0.821
	Error tipo I en %	11.8	22.3	40.1	63.0	82.6	95.0	99.2	100.0	100.0
Experimental	$q_{0.95}$	0.502	0.557	0.627	0.682	0.724	0.757	0.783	0.804	0.822
	Error tipo I en %	2.1	4.9	4.4	4.8	4.9	6.5	5.3	5.2	4.9

Tabla 3.6: Cuantil 0.95 de la distribución teórica y empírica sobre 1000 simulaciones del índice de *Rand* y porcentaje de error tipo I del test de Rand, calculado con los respectivos $q_{0.95}$ como puntos de corte, para $N = 1000$ comparaciones de particiones independientes.

Estas diferencias en la magnitud de los errores tipo I se explican por la pequeña variabilidad del índice de Rand al comparar particiones independientes con clases balanceadas (en la simulaciones la desviación típica del índice de Rand es ≈ 0.0003). Con estos resultados podemos ver que el test basado en la distribución teórica del índice de Rand no tiene un buen desempeño a diferencia del test basado en el MCE.

Más allá de estos resultados reconocemos las limitaciones de la discusión en torno a la hipótesis nula de independencia de particiones en el contexto de evaluación externa de resultados de *clustering*, ya que dos particiones pueden ser muy diferentes (siendo un *clustering* una muy mala clasificación respecto a la verdadera etiqueta de los datos) y no ser independientes. Por eso enfatizamos la contribución que hace este trabajo al escaso conocimiento existente sobre las propiedades teóricas de los índices, que requiere continuidad para levantar supuestos, con el objetivo de obtener un test que evalúe la igualdad de dos particiones.

Capítulo 4

Consideraciones finales

Uno de los objetivos de este trabajo fue la elaboración de un estado del arte sobre los índices empleados para comparar particiones y sus propiedades. En la revisión bibliográfica se constató que existen numerosos índices utilizados para la validación externa de resultados de *clustering*. Sin embargo, los estudios sobre sus propiedades son escasos y principalmente de tipo experimental. Los desarrollos teóricos que permitan avanzar hacia pruebas de aplicación más genérica son casi inexistentes en la literatura.

En este trabajo nos centramos en las propiedades del índice Mínimo error de clasificación, introducido por Meila en [17]. Se estudió la función de distribución y se obtuvo la expresión analítica para un caso particular: particiones independientes con dos clases balanceadas. A partir del conocimiento de la distribución teórica del índice se planteó una prueba de hipótesis para particiones independientes y se verificó que la región crítica del estadístico es cercana a la estimada con datos simulados artificialmente. Además se demostraron otras propiedades de interés, para el caso de tres clases y para el caso general.

El índice fue evaluado empíricamente a través de datos simulados y con relación a otros índices de validación. Los resultados muestran que los cambios en los parámetros experimentales como cantidad de observaciones, cantidad de clases y clases desbalanceadas o balanceadas condicionan la distribución de los índices, reforzando la idea de que no es conveniente generalizar pruebas de significación basadas en resultados empíricos. Los hallazgos presentados constituyen una contribución teórica y son insumos para profundizar el análisis de

la distribución del índice que permita ir levantando los supuestos y restricciones para diseñar metodologías generalizables.

Las líneas de investigación inmediatas que pueden dar continuidad a este estudio son derivar la distribución teórica para el caso de tres clases (o más), levantar el supuesto de clases uniformemente distribuidas para considerar el caso de clases no balanceadas y también el estudio de comparación de particiones con diferente cantidad de clases.

Por último, a partir de los supuestos que permitieron conocer la distribución teórica del índice se pudo diseñar un estadístico para probar independencia de particiones. En este punto se plantea una nueva interrogante sobre la distribución de un estadístico de prueba bajo la hipótesis de particiones similares, que permita la definición y el estudio teórico de un test para la igualdad de particiones que sería de gran utilidad práctica.

Referencias bibliográficas

[1] Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., and Dougherty, E. R. Model-based evaluation of clustering validation measures. *Pattern recognition 40*, 3 (2007), 807–824.

[2] de Souto, M. C., Coelho, A. L., Faceli, K., Sakata, T. C., Bonadia, V., and Costa, I. G. A comparison of external clustering evaluation indices in the context of imbalanced data sets. In *Neural Networks (SBRN), 2012 Brazilian Symposium on* (2012), IEEE, pp. 49–54.

[3] Dongen, S. Performance criteria for graph clustering and markov cluster experiments. *Technical Report INS-R0012* (2000).

[4] Fowlkes, E. B., and Mallows, C. L. A method for comparing two hierarchical clusterings. *Journal of the American statistical association 78*, 383 (1983), 553–569.

[5] Fraley, C., Raftery, A., Murphy, T., and Scrucca, L. Mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report No. 597* (2012).

[6] Fred, A. L. N., and Jain, A. K. Robust data clustering. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on* (2003), vol. 2, IEEE, pp. 128–136.

[7] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. On clustering validation techniques. *Journal of intelligent information systems 17*, 2-3 (2001), 107–145.

[8] Hubert, L., and Arabie, P. Comparing partitions. *Journal of classification 2*, 1 (1985), 193–218.

[9] Hultsch, L. Untersuchung zur besiedlung einer sprengfläche im pockautal durch die tiergruppen heteroptera (wanzen) und auchenorrhyncha(zikaden). *Studienarbeit TU Bergakademie Freiberg, Studiengang Geoökologie* (2004).

[10] Idrissi, A. N. *Contribution à l'unification de critères d'association pour variables qualitatives*. PhD thesis, Thèse de doctorat de l'Université de Paris 6, 2000.

[11] James, G., Witten, D., Hastie, T., and Tibshirani, R. *An introduction to statistical learning*, vol. 112. Springer, 2013.

[12] Kaufman, L., and Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

[13] Li, T., Ogihara, M., and Ma, S. On combining multiple clusterings. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (2004), ACM, pp. 294–303.

[14] Meila, M. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*. Springer, 2003, pp. 173–187.

[15] Meila, M. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd international conference on Machine learning* (2005), ACM, pp. 577–584.

[16] Meila, M. Comparing clusterings-an information based distance. *Journal of multivariate analysis 98*, 5 (2007), 873–895.

[17] Meila, M., and Heckerman, D. An experimental comparison of model-based clustering methods. *Machine learning 42*, 1-2 (2001), 9–29.

[18] Milligan, G. W., and Cooper, M. C. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research 21*, 4 (1986), 441–458.

[19] Patrikainen, A., and Meila, M. Comparing subspace clusterings. *IEEE Transactions on knowledge and data engineering 18*, 7 (2006), 902–916.

[20] Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66* (12 1971).

[21] Rezaei, M., and Fränti, P. Set matching measures for external cluster validity. *IEEE Transactions on Knowledge and Data Engineering 28*, 8 (2016), 2173–2186.

[22] Saporta, G., and Youness, G. Comparing two partitions: Some proposals and experiments. In *Compstat* (2002), Springer, pp. 243–248.

[23] Steinley, D. Properties of the hubert-arable adjusted rand index. *Psychological methods 9*, 3 (2004), 386.

[24] Strehl, A., and Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research 3*, Dec (2002), 583–617.

[25] Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research 11*, Oct (2010), 2837–2854.

[26] Wagner, S., and Wagner, D. *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

[27] Wu, J., Chen, J., Xiong, H., and Xie, M. External validation measures for k-means clustering: A data distribution perspective. *Expert Systems with Applications 36*, 3 (2009), 6050–6061.

[28] Youness, G., and Saporta, G. Some measures of agreement between close partitions. *Student 51* (2004), 1–12.

[29] Youness, G., and Saporta, G. Comparing partitions of two sets of units based on the same variables. *Advances in data analysis and classification 4*, 1 (2010), 53–64.

Librerías de R utilizadas

- *cluster*, *mclust*
  Utilizadas para la aplicación de algoritmos de *clustering*.
- *cubt*
  Se utilizó para el cálculo del MCE.
- *clue*, *clues*, *mclust*, *mcclust*, *partitions*, *MixSim*, *clusteval*, *clusterSim*, *partitionComparison*, *NMI*
  Utilizadas en el cálculo de otros índices de comparación.