Estudio de Métodos de Descomposición para problemas de Cuantificación Ordinal

David Pérez Román
Máster Universitario en Investigación en Inteligencia Artificial
Universidad Internacional Menéndez Pelayo (UIMP)
Madrid, España
100010523@alumnos.uimp.es

***Abstract*—El objetivo de este trabajo es corroborar la hipótesis de que el método de descomposición Frank and Hall (F&H), no es adecuado para problemas de cuantificación ordinal, pese a que en el ámbito de la clasificación sí mejora los modelos existentes. En el caso de la cuantificación, este método asume un reparto homogéneo en la distribución de las clases en los casos en las que estas son agrupadas. Esta asunción raramente es certera provocando que estos métodos tengan peor desempeño que los métodos convencionales. Para corroborar esta hipótesis se plantean dos experimentos en los que se emplearán 5 modelos reconocidos en la literatura y se enfrentarán a sus contrapartes empleando la descomposición F&H. Estos experimentos no solo demuestran que los métodos de descomposición no son adecuados para problemas de cuantificación ordinal, sino que además desmienten la asunción del reparto homogéneo en la clase complementaria.**

***Index Terms*—Machine Learning, Cuantificación, Métodos de Descomposición, Frank and Hall**

I. Introducción

La cuantificación, también conocida como estimación de la prevalencia, es un subcampo del aprendizaje supervisado que tiene como objetivo estimar la distribución de clases en un conjunto de datos sin etiquetar. El enfoque de la cuantificación fue introducido por Forman en 2005 [1] y 2006 [2], donde aborda el problema de la cuantificación, específicamente en el contexto de los problemas de cuantificación binaria, formalizando la tarea de cuantificación, evaluando varios métodos como Classify and Count o Adjusted Count [1] e introduciendo una evaluación experimental extensa de los métodos propuestos en varios conjuntos de datos, demostrando su efectividad para estimar con precisión las proporciones de clases a pesar de las imprecisiones del clasificador empleado.

A diferencia de la clasificación tradicional, que pretende asignar una etiqueta a cada instancia (o multiples etiquetas en el caso de la clasificación multi-etiqueta), la cuantificación se centra en predecir la proporción de cada clase en una muestra determinada. Una idea errónea común en el ámbito del aprendizaje máquina es la suposición de que el problema de la cuantificación es trivial, asumiendo que bastaría con entrenar un clasificador, emplear dicho clasificador para clasificar todas las instancias de una muestra y contar los resultados para determinar la prevalencia de cada clase. Este método, conocido como Clasificar y Contar no funciona ya que, en los problemas de cuantificación, se viola la asunción i.i.d. (independientes e idénticamente distribuidas) en la que se basan los algoritmos de clasificación, donde se asume que la distribución $P(X, Y)$ en los conjuntos de entrenamiento y test es la misma. En los problemas de cuantificación se asume que la distribución $P(X, Y)$ en los conjuntos de entrenamiento y test cambia, por definición al menos $P(y)$ cambia, de otra forma el problema de cuantificación sería trivial. El cambio esperado en la distribución se denomina de diferentes formas en la literatura, siendo las más habituales *prior probability shift* y *label shift*. Formalmente, este tipo de cambio es aquel en el que $P(y)$ cambia, pero $P(x|y)$ no cambia, es decir, la distribución de las clases varía pero la distribución de los ejemplos dada la clase no lo hace. González et al. realizan en [3] un estudio sobre el estado del arte en el campo de la cuantificación en el que se exponen en detalle los diferentes métodos propuestos para abordar dicho problema.

Existen multiples aplicaciones donde la distribución de clases es crucial, algunos ejemplos de aplicaciones en el ámbito de la cuantificación son:

- Análisis de sentimientos [4], [5]; donde el objetivo es determinar el sentimiento general (positivo, negativo, neutral) en un cuerpo de texto, como reseñas de productos, publicaciones en redes sociales o comentarios de clientes. La cuantificación ayuda a estimar la proporción de cada clase de sentimiento, lo que resulta útil para medir la opinión pública y tomar decisiones comerciales.
- Minado de opiniones [6], donde se pretende extraer y analizar información subjetiva de datos de texto. La cuantificación se utiliza para estimar la proporción de opiniones diferentes sobre temas específicos, como opiniones políticas o reseñas de productos.
- Detección de fraude [7], en la que la cuantificación se puede utilizar para estimar la prevalencia de transacciones fraudulentas dentro de un conjunto de datos. Esto ayuda a evaluar el riesgo general y desarrollar estrategias para mitigar el fraude.

Otras aplicaciones comúnmente asociadas a la cuantificación son la investigación de mercados, epidemiología o aplicaciones en el ámbito médico.

Del mismo modo que con la clasificación, en el ámbito de la cuantificación también existen diferentes tipologías de problemas, este trabajo se centra en problemas de cuantificación

ordinales. La cuantificación ordinal es un caso específico dentro del campo del aprendizaje supervisado donde las clases tienen un orden natural. Este orden añade una capa adicional de información, ya que los métodos no sólo deben estimar con precisión las proporciones de las clases sino también tener en cuenta la naturaleza ordinal de las mismas. Algunos ejemplos de este tipo de problemas incluyen escalas de calificación (p. ej., de 1 a 5 estrellas), grados educativos (p. ej., A, B, C) y etapas de las enfermedades (p. ej., temprana, media, tardía). Algunas de las aplicaciones de la cuantificación ordinal son el análisis de encuestas, la evaluación educativa, las aplicaciones médicas y la investigación de mercado, entre otros.

En el paradigma de la clasificación, en concreto en el caso multi-clase, existen numerosos métodos para abordar dicho problema. Sin embargo, se ha demostrado que existen métodos para reducir la complejidad del problema para pasar a multiples problemas binarios, en lugar de un problema multi-clase, mejorando así los resultados obtenidos. Estos métodos son conocidos como métodos de descomposición.

Sin embargo, se ha demostrado que los métodos de descomposición no son adecuados para problemas de cuantificación, obteniendo mejores resultados los métodos convencionales multi-clase que los métodos empleando descomposición. Este trabajo pretende extender esta misma hipótesis al caso de los problemas de cuantificación ordinal, empleando el método de descomposición especifico para este tipología de problemas, conocido como el método de descomposición *Frank and Hall* [8] (F&H) que típicamente se emplea en problemas de clasificación ordinal.

Para corroborar dicha hipótesis, se presentará una experimentación basada en dos experimentos, divididos en dos casos cada uno, es decir, se presentarán cuatro casos. Los dos primeros consisten en datos generados artificialmente y en los dos últimos casos se estudia la hipótesis sobre datos reales, primero empleando datos obtenidos de una competición de cuantificación reciente, LeQua2024 [9], concretamente empelando la tarea T3 que es la de cuantificación ordinal. Segundo, empleando datasets comúnmente utilizados en la literatura.

II. Métodos

En los problemas de cuantificación se parte de un conjunto de entrenamiento $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ en el que $x_i \in \mathcal{X}$ e $y_i \in \mathcal{Y} = \{c_1, c_2, ..., c_k\}$, siendo $\mathcal{X}$ el espacio de entrada e $\mathcal{Y}$ el conjunto de k clases. Mientras que en clasificación un modelo h debe asociar cada ejemplo a una clase: $h : \mathcal{X} \to \{c_1, c_2, ..., c_k\}$, en cuantificación el modelo $\overline{h}$ proporciona una predicción agregada, es decir, no se predice cada ejemplo individualmente, sino la prevalencia $\hat{p}_l$ de cada una de las k clases en todo el conjunto: $\overline{h} : \mathcal{X}^m \to [0, 1]^K$, cumpliendo que $0 \leq \hat{p}_l \leq 1$ y $\sum_{l=1}^{k} \hat{p}_l = 1$.

Para acometer este tipo de problemas se han propuesto diversos métodos en la literatura. De entre los existentes, en las siguientes secciones se expondrán los empleados en los experimentos llevados a cabo. Adicionalmente se expondrá el método Classify and Count (CC), al ser este muy util para presentar Adjusted Count (AC) y el Probabilistic Adjusted Count (PAC), dos métodos que sí se van a emplear en el trabajo. Adicionalmente se presentará un método basado en probabilidades Expectation Maximization (EM) y, finalmente, métodos basados en comparar distribuciones, como HDy (basado en la distancia de Hellinger) y EDy (basado en Energy Distance).

A. Classify and Count (CC)

Classify and Count o Clasificar y Contar (CC) es uno de los métodos de cuantificación más simples e intuitivos. Implica utilizar un modelo de clasificación estándar $h : \mathcal{X} \to \{c_1, c_2, ..., c_k\}$ para clasificar primero cada instancia en un conjunto de datos y luego simplemente contar el número de instancias que se predice que pertenecen a cada clase. Seguidamente, los recuentos obtenidos se normalizan para estimar la prevalencia (proporción) de cada clase en el conjunto de datos. El paso de conteo y normalización se pueden unir obteniendo:

$$\hat{p}_l = \frac{1}{m} \sum_{x_j \in T} [\![h(x_j) = c_l]\!], \tag{1}$$

donde $[\![q]\!]$ devuelve 1 cuando el predicado es cierto y 0 en el caso contrario.

Como se ha expuesto previamente, CC no funciona bien para problemas de cuantificación. CC tiende a subestimar o sobrestimar las prevalencias cuando el clasificador empleado no es perfecto, que es el caso obviamente habitual en cualquier problema real.

Forman, para derivar el método Adjusted Count que se verá a continuación, explicó que la prevalencia que predice CC para cada clase se puede expresar como:

$$\hat{p}_l = \sum_{j=1}^{k} P(h(x) = c_i | y = c_j) P_T(c_j), \tag{2}$$

donde $P(h(x) = c_i | y = c_j)$ es la probabilidad de que el clasificador h de como predicción la clase c_i cuando el ejemplo realmente pertenece a la clase c_j y $P_T(c_j)$ representa la prevalencia real de la clase c_j en el conjunto T.

B. Adjusted Count (AC)

El método Adjusted Count (AC) [10], es una mejora con respecto al método Classify and Count (CC). AC tiene como objetivo corregir el sesgo inherente en CC ajustando las estimaciones de prevalencia de clase en función de la matriz de confusión del clasificador.

Del mismo modo que con CC, un clasificador estándar se entrena con datos de entrenamiento etiquetados y luego se utiliza para predecir las etiquetas de clase para cada instancia en el conjunto de test o en un conjunto de datos sin etiquetar. Para ajustar los errores de clasificación, se emplea la matriz de confusión del clasificador. Usando la matriz de confusión, se pueden ajustar los recuentos brutos del método CC para estimar la prevalencia real de cada clase.

Partiendo de (2), $\hat{p}_l$ se puede estimar a partir del conjunto de entrenamiento mediante validación cruzada. $P_T(h(x) = c_a)$

se obtiene simplemente aplicando h sobre T. Al escribir esta misma expresión para todas las clases se obtiene un sistema de k ecuaciones con k incógnitas, los valores de $\hat{p}_l$. Por ejemplo, para $k = 3$ y denotando $P(h(x) = c_a|y = c_l)$ como $P(c_a|c_l)$:

$$\begin{pmatrix} P(c_1|c_1) & P(c_1|c_2) & P(c_1|c_3) \\ P(c_2|c_1) & P(c_2|c_2) & P(c_2|c_3) \\ P(c_3|c_1) & P(c_3|c_2) & P(c_3|c_3) \end{pmatrix} * \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \hat{p}_3 \end{pmatrix} = \begin{pmatrix} P_T(h(x)=c_1) \\ P_T(h(x)=c_2) \\ P_T(h(x)=c_3) \end{pmatrix}. \quad (3)$$

AC mejora las estimaciones obtenidas mediante CC corrigiendo los errores del clasificador utilizando la matriz de confusión. Este ajuste a menudo da como resultado estimaciones más precisas de prevalencia, especialmente cuando el desempeño del clasificador es imperfecto. Sin embargo, la precisión de AC depende de la confiabilidad de la matriz de confusión, que idealmente debería estimarse de la forma mas precisa posible, por ejemplo, empleando Validación Cruzada.

C. Probabilistic Adjusted Count (PAC)

Probabilistic Adjusted Count (PAC) originalmente nombrado *Scaled Probability Average* (SPA) en [11], y renombrado como *Probabilistic Adjusted Count* (PAC) en [3], es una extensión del método PCC, siendo este ultimo una version de CC que utiliza las predicciones probabilísticas (probabilidades de pertenencia a una clase) producidas por el clasificador en lugar de solo predicciones estrictas.

En el método PCC, la prevalencia de cada clase es simplemente el promedio de las probabilidades de esa clase en los ejemplos de T:

$$\hat{p}_l = \frac{1}{m} \sum_{x_j \in T} h(x_j, c_l). \quad (4)$$

Del mismo modo que hace el método AC respecto a CC, PAC resuelve un sistema conceptualmente equivalente a (3) pero basado en el cálculo de probabilidades posteriores promediadas sobre D (nuevamente empleando Validación Cruzada) y T:

$$\begin{pmatrix} \overline{h(x_i,c_1)}_{\substack{x_i\in D\\ y_i=c_1}} & \overline{h(x_i,c_1)}_{\substack{x_i\in D\\ y_i=c_2}} & \overline{h(x_i,c_1)}_{\substack{x_i\in D\\ y_i=c_3}} \\ \overline{h(x_i,c_2)}_{\substack{x_i\in D\\ y_i=c_1}} & \overline{h(x_i,c_2)}_{\substack{x_i\in D\\ y_i=c_2}} & \overline{h(x_i,c_2)}_{\substack{x_i\in D\\ y_i=c_3}} \\ \overline{h(x_i,c_3)}_{\substack{x_i\in D\\ y_i=c_1}} & \overline{h(x_i,c_3)}_{\substack{x_i\in D\\ y_i=c_2}} & \overline{h(x_i,c_3)}_{\substack{x_i\in D\\ y_i=c_3}} \end{pmatrix} * \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \hat{p}_3 \end{pmatrix} = \begin{pmatrix} \overline{h(x_j,c_1)}_{x_j\in T} \\ \overline{h(x_j,c_2)}_{x_j\in T} \\ \overline{h(x_j,c_3)}_{x_j\in T} \end{pmatrix}. \quad (5)$$

Este enfoque puede proporcionar una cuantificación más precisa, especialmente cuando el clasificador está bien calibrado y las probabilidades de pertenencia a una clase son confiables. Sin embargo, del mismo modo que con el método AC, es necesario estimar correctamente la matriz de confusión.

D. Expectation Maximization (EM)

Expectation Maximization (EM) es un método iterativo muy conocido para encontrar estimaciones de máxima verosimilitud de parámetros en modelos probabilísticos, especialmente cuando el modelo depende de variables latentes no observadas [12]. En el contexto de la cuantificación, el algoritmo EM se puede adaptar para estimar las verdaderas prevalencias de clase en un conjunto de pruebas basado en las predicciones observadas [13].

Primero, se parte de una suposición inicial para las prevalencias de clase $p^{(0)}(y_l)$, habitualmente se toma misma probabilidad a priori que la del conjunto de entrenamiento $\hat{p}_t(y_l)$. Seguidamente se procede con el paso de estimación, donde se reestima la probabilidad a posteriori de cada ejemplo para cada clase $P(y_i = c_l|x_i)$. La idea de este paso es que no hace falta reetrenar el clasificador para una distribución similar al conjunto de test, se van reajustando esas probabilidades en las iteraciones del algoritmo. Finalmente se procede con el paso de maximización, donde se actualizan las estimaciones de las prevalencias de clase en función de los valores esperados calculados en el paso de estimación.

Se repiten los pasos de estimación y maximización hasta que el algoritmo converge (cuando no se aprecia mucho cambio en las prevalencias). Finalmente, devuelve como predicción las prevalencias que se obtienen en el ultimo paso de maximización que se ha ejecutado.

Definiendo $h_l(x_i)$ como la salida del clasificador correspondiente a la clase y_l para la observación x_i del conjunto de datos sobre los que predecir las prevalencias.

$$\hat{p}_t(y_l|x_i) = h_i(x_i), \quad (6)$$

Se definen las ecuaciones para los tres pasos del algoritmo *Inicialización*, *Estimación* y *Maximización*:

$$\hat{p}^{(0)}(y_l) = \hat{p}_t(y_l), \quad (7)$$

$$\hat{p}^{(s)}(y_l|x_i) = \frac{\frac{\hat{p}^{(s)}(y_l)}{\hat{p}_t(y_l)} \hat{p}_t(y_l|x_i)}{\sum_{j=1}^{l} \frac{\hat{p}^{(s)}(y_j)}{\hat{p}_t(y_j)} \hat{p}_t(y_l|x_i)}, \quad (8)$$

$$\hat{p}^{(s+1)}(y_l) = \frac{1}{m} \sum_{i=1}^{m} \hat{p}^{(s)}(y_l|x_i). \quad (9)$$

E. Métodos basados en Hellinger Distance (HDy)

El método HDy (Método basado Distancia de Hellinger) está diseñado para abordar tareas de cuantificación, particularmente cuando las distribuciones de clases en los conjuntos de entrenamiento y test difieren. Como los métodos anteriores, asume que $P(x|y)$ es constante y aprovecha la distancia de Hellinger, una medida de divergencia entre dos distribuciones de probabilidad, para ajustar las estimaciones de prevalencia de las clases [14].

Dado el conjunto de entrenamiento D, y un conjunto de test T, se estima la función de distribución de entrenamiento y test respectivamente. Se modifica la distribución de D, variando las prevalencias de las clases, hasta que se aproxime lo máximo posible a la distribución de T. Esta aproximación consiste en ponderar los casos del conjunto D por sus prevalencias estimadas, dando lugar a una distribución mixta denotada por D':

$$D' = D^{c_1} \cdot \hat{p}_1 + D^{c_2} \cdot \hat{p}_2 + ... + D^{c_k} \cdot \hat{p}_k, \quad (10)$$

donde D^{c_l} es la distribución de los ejemplos de la clase c_l. D^{c_l} variará de forma constante según cambie la prevalencia $\hat{p}_l$. El objetivo es minimizar la distancia Δ entre D' y T.

$$\arg\min_{\hat{p}_k} \Delta(D', T) = \arg\min_{\hat{p}_k} \Delta\left(\sum_{l=1}^{k} D^{c_l}\hat{p}_l, T\right). \quad (11)$$

Gonzalez et al. presentan dos métodos para resolver este problema en [14]: a través de los atributos (método HDx) o bien estimando las distribuciones D^{c_l} y T a través de las predicciones dadas por el clasificador (método HDy).

Para emplear el método HDy, se entrena un clasificador probabilístico en el conjunto de entrenamiento y se obtienen predicciones probabilísticas para cada instancia del conjunto de entrenamiento y de test. Sea $p(y = c_l|x_i)$ la probabilidad prevista de que la instancia x_i pertenezca a la clase c_l. Se divide el rango de probabilidades predichas en b contenedores (bins). Cada bin representa un rango de probabilidades predichas. Para cada clase, se crea un histograma de las probabilidades previstas tanto para el conjunto de entrenamiento como para el de test.

Por ejemplo, en el caso de HDy, la distancia de Hellinger entre la distribución de salida del clasificador para datos de test T y la distribución de salida del clasificador para datos de entrenamiento D se puede estimar como:

$$\min_{\hat{p}_1,\ldots,\hat{p}_k} \frac{1}{k}\sum_{l=1}^{k}\sqrt{\sum_{r=1}^{b}\left(\sqrt{\frac{|T_{r,l}|}{m}} - \sqrt{\sum_{l=1}^{k}\frac{|D_{r,l}^{c_l}|}{n^{c_l}}\hat{p}_l}\right)^2}, \quad (12)$$

donde $D_{r,l}^{c_l}$ es el numero de ejemplos de la clase c_l que caen en el bin r, n^{c_l} es el numero de ejemplos de la clase c_l y $T_{r,l}$ es el numero de ejemplos de test que caen en el bin r clase l.

Se emplean las distancias de Hellinger calculadas para ajustar las estimaciones de prevalencia brutas. La idea es encontrar las prevalencias de D' que minimicen la distancia de Hellinger entre las distribuciones D' y T.

F. Métodos basados en Energy Distance (EDy)

El método EDy (Método basado en Energy Distance) está diseñado para abordar tareas de cuantificación, particularmente cuando las distribuciones de clases en los conjuntos de entrenamiento y prueba difieren, similar al caso de HDy. Aprovecha la Energy Distance, una medida de distancia estadística entre dos distribuciones, para ajustar las estimaciones brutas de prevalencia de clase obtenidas de un clasificador [15].

El método EDy sigue la misma lógica y pasos que el método HDy, pero en este caso se minimiza la Energy Distance (ED) entre las predicciones de los conjuntos D' y T. Edy minimiza la siguiente expresión:

$$\begin{aligned}\min\ & 2\cdot\mathbb{E}_{x_i\backsim D', x_j\backsim T}\,||h(x_j) - h(x_i)|| \\ & -\mathbb{E}_{x_i, x_i'\backsim D'}\,||h(x_i) - h(x_i')||.\end{aligned} \quad (13)$$

El uso de la Energy Distance permite una medida sólida de la divergencia entre las distribuciones de entrenamiento y test, consiguiendo así una cuantificación más precisa.

III. Métodos de descomposición

El concepto de métodos de descomposición, como *uno contra todos* (OvA) y *uno contra uno* (OvO), surgió originalmente en el contexto de problemas de clasificación multi-clase. Estos métodos son técnicas bien conocidas para extender clasificadores binarios para abordar la clasificación multi-clase [16].

En el caso ordinal, el método de descomposición empleado es el método *Frank and Hall* [8], introducido por Eibe Frank y Mark Hall, conocido por su eficacia y simplicidad. El método es una extensión y refinamiento de las técnicas de descomposición existentes como uno contra todos (OvA) y uno contra uno (OvO), con un enfoque en reducir la complejidad computacional y mejorar el rendimiento de los clasificadores ordinales.

A. Descomposición Frank & Hall (F&H)

El método de Frank y Hall tiene como objetivo combinar las ventajas de OvA y OvO y al mismo tiempo mitigar sus inconvenientes. La idea clave es utilizar una estructura jerárquica para reducir la cantidad de clasificadores binarios y hacer que el proceso de aprendizaje sea más eficiente teniendo en cuenta la estructura ordenada de las clases.

El método se basa en tres pasos principales. Primero, binarización del problema ordinal: En lugar de entrenar un clasificador para todo el problema, se divide el problema y se crean clasificadores binarios. Se parte de un problema ordinal con k clases, la idea es dividir el problema con k clases en $k-1$ problemas binarios, agrupando las clases respetando el orden de las mismas, de este modo se respeta la ordinalidad del problema. Segundo, entrenamiento de clasificadores binarios: Para cada uno de los $k-1$ problemas binarios generados en el primer paso, se entrena un clasificador binario. La tarea del clasificador es predecir el valor la clase correspondiente (0 o 1) en función de las características de entrada. El entrenamiento comienza derivando nuevos conjuntos de datos del conjunto de datos original, uno para cada uno de los $k-1$ nuevos atributos de clase binaria. Tercero, combinación de clasificadores: Para predecir el valor de clase de una instancia desconocida es necesario estimar las probabilidades de las k clases ordinales originales usando los modelos $k-1$. La estimación de la probabilidad del primer y último valor de clase ordinal depende de un único clasificador, mientras que la probabilidad de las clases intermedias dependerá de dos clasificadores.

Se denota P_i como la probabilidad $P_r(Target > c_i)$ para $i = 1, 2, ..., k-1$ clases. Usando estas probabilidades, se puede calcular la probabilidad para cada clase c_i de la siguiente manera:

- Clase c_1: La probabilidad de estar en la primera clase viene dada por $P_r(Target = c_1) = 1 - P_1$
- Clase c_k: La probabilidad de estar en la última clase viene dada por: $P_r(Target = c_k) = P_{k-1}$
- Clases intermedias $(c_2, c_3, ..., c_{k-1})$: Para estas clases, la probabilidad depende de dos clasificadores binarios: $P_r(Target = c_i) = P_i - 1\cdot(1 - P_i)$ para $i = 2, 3, ..., k-1$

Un ejemplo común de aplicación del método Frank and Hall es el siguiente:

Imagínese un problema de clasificación ordinal donde el objetivo es predecir las calificaciones de satisfacción del cliente en una escala del 1 al 4. El método de Frank y Hall se puede adaptar de la siguiente manera:

- **Binarización del problema ordinal**: Se binariza el problema, transformando el problema de 4 clases en 3 problemas binarios (1 vs 2-3-4, 1-2 vs 3-4 y 1-2-3 vs 4).
- **Entrenamiento de clasificadores binarios**: Se entrenan 3 clasificadores binarios, uno para cada problema binario generado en el primer paso.
- **Predicción de nuevos casos**: Se emplean los clasificadores entrenados en el paso anterior para predecir la clase de una nueva instancia desconocida. Aplicando las formulas expuestas en la explicación anterior, se puede obtener la probabilidad de cada clase por separado.

Este método se puede extender a problemas de cuantificación, el paso de binarización del problema ordinal es el mismo, pero en este caso se entrena un cuantificador por cada modelo de la descomposición de Frank y Hall (F&H), en lugar de un clasificador, es decir, se entrenan 4 cuantificadores: (1 vs 2-3-4, 1-2 vs 3-4 y 1-2-3 vs 4) y se combinan sus pronósticos. La clase positiva corresponde al grupo izquierdo de cada cuantificador (1, 1-2, etc.) En este tipo de estrategia de descomposición es importante garantizar que las prevalencias p consecutivas agregadas no disminuyan. Esta situación puede ocurrir ya que los cuantificadores binarios se entrenan de forma independiente. Se supone el siguiente caso:

cuantificador 1 vs 2-3-4	$p(1) = 0,3$
cuantificador 1-2 vs 3-4	$p(1,2) = 0,2$
cuantificador 1-2-3 vs 4	$p(1,2,3) = 0,6$

Esto es inconsistente. Destercke y Yang [17] proponen emplear el método que se basa en calcular las prevalencias acumuladas superior (ajustando de izquierda a derecha) e inferior (de derecha a izquierda). Estos conjuntos de valores aumentan monótonamente (de izquierda a derecha) y disminuyen monótonamente (de derecha a izquierda), respectivamente. El valor promedio se asigna a cada grupo y la prevalencia para cada clase se calcula como:

$$p(y_k) = p(y_1, ..., y_k) - p(y_1, ..., y_{k-1}). \tag{14}$$

Tomando las siguientes prevalencias como ejemplo:

(1)	(1-2)	(1-2-3)
0,3	0,3	0,6
0,2	0,2	0,6
0,25	0,25	0,6

Donde la primera fila corresponde con las prevalencias acumuladas superiores (ajustando de izquierda a derecha), la segunda con las prevalencias acumuladas más bajas (ajustando de derecha a izquierda) y la tercera con las prevalencias promedio. Se pueden obtener las prevalencias de cada clase aplicando la metodología descrita anteriormente, de modo que:

$$p(1) = 0,25$$
$$p(2) = p(1,2) - p(1) = 0,25 - 0,25 = 0$$
$$p(3) = p(1,2,3) - p(1,2) = 0,6 - 0,25 = 0,35$$

La última clase se calcula como 1 menos la suma de prevalencias del resto de clases:

$$p(4) = 1 - p(1,2,3) = 1 - 0,6 = 0,4$$

Al aplicar la descomposición F&H en problemas de cuantificación se supone uniformidad entre las clases, es decir, se supone que $P(x|y)$ se mantiene constante entre las clases. Si se sigue el ejemplo anterior, al aprender un cuantificador binario cualquiera en el caso (1 vs 2, 3, 4) y obteniendo una prevalencia de 0.7 para la clase 1, implicaría un reparto homogéneo del resto de la probabilidad entre las clases restantes, ya que $P(x|y = \{2, 3, 4\})$ se asume que permanece constante. Por lo tanto, implicaría un reparto 0.1 por clase restante (2, 3 y 4), una asunción muy fuerte que en muchos casos no se cumple. Partiendo de esta suposición, se puede asumir que, a medida que las muestras tengan prevalencias muy diferentes a una distribución homogénea, el error tenderá a crecer en los cuantificadores binarios, afectando al error del método de descomposición. En su defecto, las variaciones entre las prevalencias de la clase de interés y la clase complementaria no varían uniformemente, provocando mayores errores en los métodos de descomposición.

En este trabajo se plantea la hipótesis de que no es cierto que $P(x|y)$ permanezca constante en las clases agregadas y por tanto la distribución de prevalencias entre dichas clases no es homogénea sino independiente. El error de los cuantificadores binarios que forman la descomposición Frank and Hall sería mayor cuánto más diferentes sean las prevalencias de las clases que forman la clase complementaria, ya que eso implica que es más difícil que $P(x|y)$ se mantenga constante en esa clase agregada. Dado que los algoritmos de cuantificación hacen esa asunción, se espera que la precisión de sus estimaciones baje al aplicar métodos de descomposición como Frank and Hall.

IV. Experimentación

Para demostrar la hipótesis propuesta al inicio del trabajo, que es que el método de descomposición Frank and Hall no es adecuado para problemas de cuantificación ordinales, se proponen dos experimentos en los que se pretende demostrar esta hipótesis de forma empírica. Se ha seguido la misma metodología y se han empleado las mismas herramientas en todos los experimentos, de este modo será posible comparar los resultados obtenidos. Todos los modelos descritos anteriormente se implementan en Python, empleando la librería QuantificationLib [18]. En todos los casos, se parte de un clasificador subyacente a los modelos, es importante utilizar siempre el mismo clasificador sin que el proceso tenga que reejecutar el entrenamiento del modelo porque uno de los requisitos para tener en cuenta en esta comparativa es que todo método que requiera de un clasificador debe de ser exactamente el mismo. El clasificador empleado es un FrankAndHallMonotoneClassifier [19] con un clasificador Logistic Regression (LR) subyacente. Se realiza un Grid Search para escoger el mejor LR que pasar al clasificador F&H y asi

obtener resultados mas fiables. FrankAndHallMonotoneClassifier es una versión probabilística de F&H que aplica la idea de Destercke y Yang, explicada en la sección III, para garantizar probabilidades consistentes.

En el primer experimento se emplean datasets generados artificialmente, en este caso se opta por un tamaño de conjunto de entrenamiento variable desde 50 a 2000 que se evaluará sobre 300 conjuntos de tests de 2000 ejemplos cada uno. Es decir, se dispondrá de 300 conjuntos de test de tamaño 2000 los cuales serán testeados en conjuntos de entrenamiento con tamaño variable. Para que este proceso aporte resultados robustos se repetirá 10 veces, de esta forma el valor reflejado será la media de 300 muestras multiplicado por 10 repeticiones, es decir es la media de 3000 resultados. Se plantean dos casos en los que el primero tiene las clases bien diferenciadas, sin mucho solapamiento en valores fronterizos, mientras que en el segundo existe más solapamiento en valores fronterizos, haciendo el test más difícil.

En el segundo experimento se emplean datasets obtenidos de datos reales, en el primer caso se emplea el dataset obtenido de la competición LeQua2024 [9], en concreto se emplea el dataset correspondiente a la tarea T3 de la competición, que es la correspondiente al caso ordinal. Para el segundo caso se emplean 5 datasets comúnmente utilizados en la literatura para formalizar un banco de pruebas con el que obtener resultados en varios datasets con datos reales.

A. *Métricas de error*

Para determinar el desempeño de los métodos descritos en la sección II, y poder compararlos entre ellos, es necesario emplear métricas que reflejen cómo se comporta el modelo. El objetivo principal es medir con qué precisión las prevalencias estimadas coinciden con las prevalencias reales. Existen varias métricas empleadas en la literatura para evaluar el rendimiento de los métodos de cuantificación.

Algunas de las métricas comúnmente empleadas en problemas de cuantificación son el Error Absoluto (AE), Error Absoluto Medio (MAE), Error Cuadrático (SE), Error Cuadrático Medio (MSE), Divergencia de Kullback-Leibler (KLD), etc. En problemas convencionales de cuantificación, MAE y KLD se suelen utilizar extensamente. Sin embargo, es posible que las métricas de cuantificación tradicionales como MAE o KLD no capturen completamente la importancia de la naturaleza ordinal de las clases. Métricas como el Earth Mover's Distance (EMD) son más adecuadas ya que tienen en cuenta el orden de las clases.

Castaño et al. [19], proponen el uso de EMD en problemas de cuantificación ordinales. Esta métrica mide la cantidad mínima de trabajo necesaria para transformar una distribución en otra, considerando la distancia entre distribuciones. Considerando que la cuantificación ordinal requiere comparar dos distribuciones de probabilidad, el conjunto de prevalencias verdaderas, p, y la predicha, $\hat{p}$, EMD se puede calcular eficientemente como:

$$EMD(p,\hat{p}) = \sum_{l=1}^{k-1} \left| \sum_{a=1}^{l} p_a - \sum_{a=1}^{l} \hat{p}_a \right| \qquad (15)$$

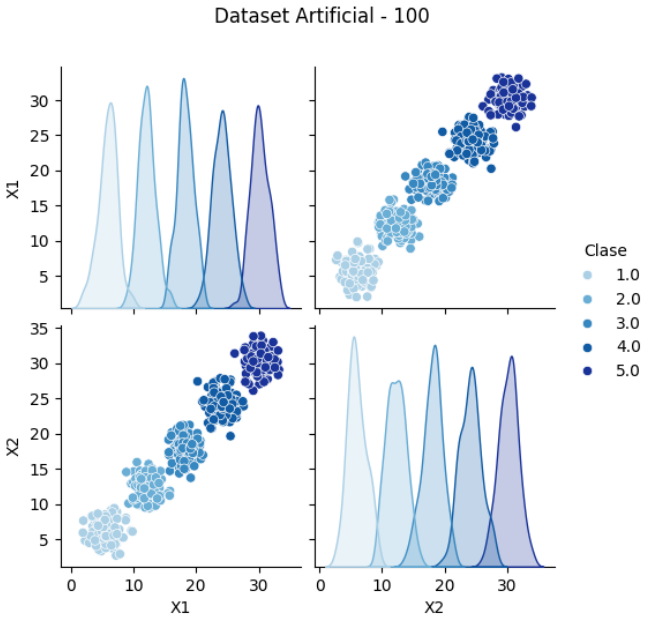

Fig. 1. Caso Artificial 1: 5 clases bien diferenciadas con poco solapamiento en valores fronterizos.

EMD calcula la masa de probabilidad que debe ser desplazada para convertir una distribución en otra y oscila entre 0 y $k-1$ en esta configuración. Cuanto menor sea el valor de EMD, más se parecerán las dos distribuciones comparadas, indicando así que las prevalencias predichas $\hat{p}$ se acercan a las prevalencias reales p. Como se ha expuesto al inicio de la sección, EMD es una alternativa al error absoluto medio (MAE), calculado como $MAE(p,\hat{p}) = \frac{1}{k}\sum_{l=1}^{k}|p_l - \hat{p}_l|$. Pero, como se expone en [19], EMD captura mucho mejor la similitud cuando las clases tienen una relación de ordinalidad.

B. *Caso Artificial 1*

En el primer caso artificial se generan dataset con clases bien diferenciadas, es decir, sin superposición (o con muy poca) en valores fronterizos. El experimento se lanzará para distintos tamaños de D: [50, 100, 500, 1000, 1500, 2000]. De este modo también se comprobará que los modelos sigan la consistencia de Fisher que, en el contexto de la cuantificación, garantiza que el error de cuantificación tiende a 0 a medida que aumenta el número de ejemplos del conjunto de entrenamiento D y el número de ejemplos del conjunto de test T.

En la Figura 1 se puede ver la distribución de los datos empleados en el caso artificial 1. Este es el caso más favorable al tener las clases bien diferenciadas. Empleando EMD como la métrica de error, se obtienen los gráficos de la Figura 2, en la que se puede apreciar los dos sucesos presentados al inicio de la sección. Por un lado se aprecia como los métodos que emplean Frank and Hall obtienen peores resultados, y además se comprueba la consistencia de Fisher, viendo como el error tiende a 0 cuanto mayor es el número de ejemplos del conjunto de entrenamiento D.

Por otro lado, se observa que métodos como AC, EM y HDy muestran resultados similares en sus versiones F&H y en las convencionales. Esto es debido a la naturaleza del caso presentado, el cual plantea el problema de cuantificación sobre un conjunto con clases claramente diferenciables. Esto favorece a los métodos F&H, puesto que el sesgo introducido por el clasificador empleando este método es menos apreciable gracias a la estructura de los datos, haciendo que se reduzca la subestimación/sobreestimación del los clasificadores binarios subyacentes. Los métodos como PAC y EDy, que dependen en gran medida de resultados probabilísticos o medidas de distancia, podrían ser más sensibles a los cambios introducidos por la descomposición. Los ajustes probabilísticos son particularmente propensos a que se amplifiquen pequeños errores en el paso final de cuantificación. En el caso del método HDy, la asignación del error hacia el grupo complementario es de mayor calidad. Al disponer de grupos claramente diferenciables, los histogramas son separables aumentando así la calidad de la descomposición.

Cabe destacar que, como se ha reiterado en varias ocasiones, este es el caso mas favorable, por lo que incluso los métodos que emplean Frank and Hall devuelven resultados consistentes, que incluso mejoran en algunos casos a los otros modelos, como es el caso del HDy. De hecho, todos los modelos muestran errores muy pequeños, del orden de 0.01. Por lo tanto, las conclusiones del caso expuesto no son generalizables ya que las clases son claramente diferenciables, favoreciendo a lo métodos F&H al mitigar el sesgo introducido por estos métodos.

C. Caso Artificial 2

En el segundo caso artificial, se mantienen las mismas bases del experimento, esta vez cambiando la separación entre clases para que haya solapamiento en valores fronterizos, tal y como se puede ver en la Figura 3. La principal diferencia con el caso 1 es que estos muestran cinco clases distinguibles, pero no triviales. Es decir, existe cierto solapamiento difuminando la frontera entre las mismas simulando una situación más pareja a datos reales. Esto es importante ya que en el supuesto de experimentar con una nube de puntos en la que no se distinga ninguna de las clases, implica que el estimador no tendría posibilidad alguna de clasificar y por tanto la cuantificación arrastra un sesgo derivado del estimador muy elevado. En caso contrario, en el supuesto de enfrentarse a un conjunto de datos extremadamente evidente con grupos muy bien diferenciados, existirá una ventaja clara en algunos métodos frente a otros.

Tras lanzar el experimento, del mismo modo que con el caso artificial 1, se emplea la métrica EMD y se generan gráficos para observar la evolución del error durante las deferentes muestras empleadas. En la Figura 4 se pueden observar las comparaciones entre los modelos. A diferencia del caso anterior, en el que las clases estaban bien diferenciadas, en este caso se aprecia una diferencia mayor entre los modelos convencionales y los modelos aplicando Frank and Hall. Este comportamiento es el esperado, los modelos Frank and Hall sufren en estos casos debido al sesgo introducido por la

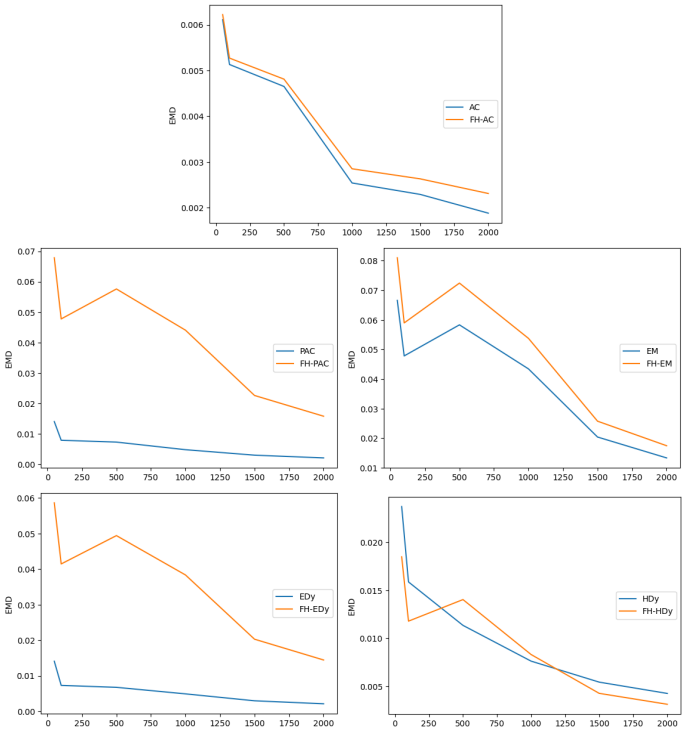

Fig. 2. Cuantificación sobre el conjunto de datos del caso artificial 1, para los métodos de AC, PAC, EM, HDy y EDy con la finalidad de compararlas con sus versiones F&H.

naturaleza del método de descomposición, los límites entre clases se vuelven difusos, lo que dificulta que los clasificadores binarios separen las clases con precisión. Esto conduce a errores compuestos al combinar los resultados de los clasificadores binarios en predicciones de múltiples clases. Al asumir que $P(x|y)$ es constante, se espera un reparto homogéneo de las prevalencias en la clase complementaria, pero esta asunción no es correcta puesto que este caso raramente se puede dar. Este hecho perjudica a los cuantificadores, sobretodo en las clases intermedias de la escala ordinal, es decir, las que no están en los extremos. Esto es debido a que los errores y sesgos de los clasificadores binarios individuales pueden propagarse y amplificarse cuando se combinan. Cada clasificador binario puede tener dificultades con la superposición de límites de clases, y sus errores combinados pueden dar como resultado mayores imprecisiones en la cuantificación final.

Este fenómeno en las clases intermedias (2, 3 y 4 en este caso) se puede observar si se enfrentan las prevalencias obtenidas por los cuantificadores contra las prevalencias reales. Como se ha expuesto anteriormente, la asunción de que $P(x|y)$ es constante puede resultar problemática cuando se aplican métodos como F&H, sobretodo en las clases intermedias. Para la primera y la última clase (1 y 5 en este caso), los cuantificadores binarios comparan estas clases en los extremos con el agregado de todas las demás clases, por ejemplo: 1 vs. (2, 3, 4 y 5) y (1, 2, 3 y 4) vs. 5. Dado que estas clases están en los extremos, la separación entre estas clases y el resto es más clara, lo que conduce a un mejor rendimiento y predicciones

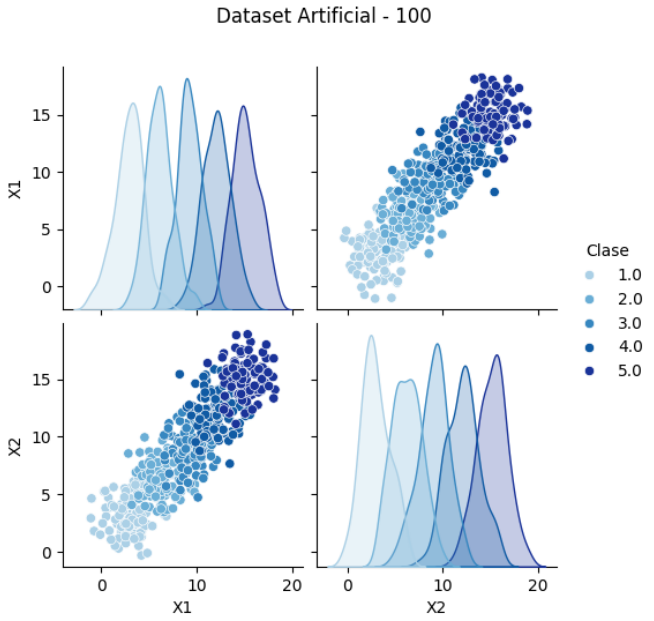

Fig. 3. Caso Artificial 2: 5 clases difusas con solapamientos en valores fronterizos.

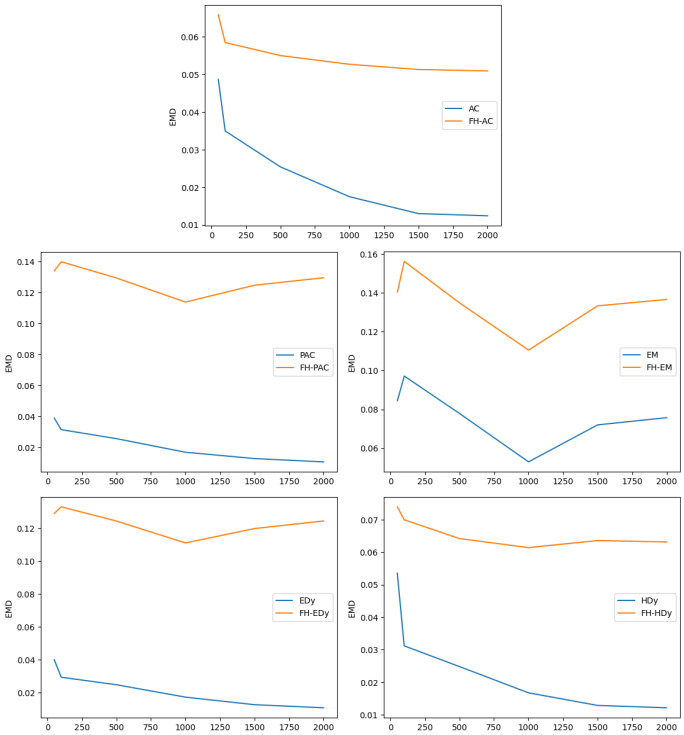

Fig. 4. Cuantificación sobre el conjunto de datos del caso artificial 2, para los métodos de AC, PAC, EM, HDy y EDy con la finalidad de compararlas con sus versiones F&H.

más cercanas al ideal.

De manera opuesta, las clases intermedias se comparan en combinaciones más complejas, como por ejemplo: (1 y 2) vs. (3, 4 y 5) y (1, 2 y 3) vs. (4 y 5). Estas clases intermedias a menudo tienen distribuciones de características superpuestas con sus clases vecinas, lo que dificulta que los clasificadores binarios las distingan con precisión. La suposición de que $P(x|y)$ es consistente entre descomposiciones no se cumple aquí porque las distribuciones de características se superponen de manera más significativa. Esta superposición introduce sesgos que afectan a las prevalencias previstas.

Partiendo de estas consideraciones, en las Figuras 5 y 6 se puede apreciar la diferencia entre los modelos Frank and Hall y los convencionales. En la Figura 5 se aprecia como los modelos convencionales predicen prevalencias cercanas al ideal (indicado en los gráficos con una línea discontinua), $y = x$, que se daría si las predicciones coincidiesen con las prevalencias reales. De manera opuesta, en la Figura 6, se puede observar como los modelos que emplean Frank and Hall tienden a subestimar las clases intermedias, alejándose del ideal. Este comportamiento está presente en todos los casos, pero es especialmente apreciable en el caso del método PAC, donde el método convencional muestra una gran cercanía al ideal, mientras que F&H tiene una clara desviación en las clases 2, 3 y 4, donde claramente el cuantificador las está subestimando. Similar a los resultados observados mediante EMD, se puede observar como tanto PAC como EDy tienen comportamientos similares.

Por este motivo, en este caso se observan diferencias mayores en el EMD de los modelos comparados, donde en el caso 1 en el modelo HDy Frank and Hall llegaba a igualar, e incluso mejorar, al modelo convencional. En el caso 2, hay una clara diferencia entre ambos, obteniendo mejores resultados el modelo convencional. Asimismo, mientras que los modelos convencionales siguen cumpliendo con la Consistencia de Fisher, algunos de los modelos F&H muestran comportamientos poco consistentes, principalmente debido a que las clases intermedias tienen distribuciones de características superpuestas con las de sus vecinas, lo que dificulta la clasificación binaria, asimismo los errores de los cuantificadores binarios individuales se acumulan, lo que resulta en mayores desviaciones en las prevalencias previstas para las clases intermedias.

D. Caso Real 1

En el primer caso artificial se mostraba un conjunto formado por cinco clases las cuales eran claramente diferenciables, en el segundo se planteaban clases con solapamiento en valores fronterizos, donde se ha observado el sesgo del método F&H para las clases intermedias y, en el primer caso real, se ha empleado un dataset obtenido de una competición de cuantificación reciente, LeQua2024 [9].

Learning to Quantify (LeQua) es una comunidad orientada exclusivamente en problemas de cuantificación en diversos campos como la recuperación de información, la minería de datos, aprendizaje automático y estadística, entre muchas otras. Uno de los objetivos de LeQua es crear un laboratorio específico para la cuantificación, así como generar sinergias entre distintas disciplinas para volcar conocimiento y compartir de técnicas. LeQua pone a disposición un nuevo laboratorio para la evaluación de métodos de cuantificación tanto en un

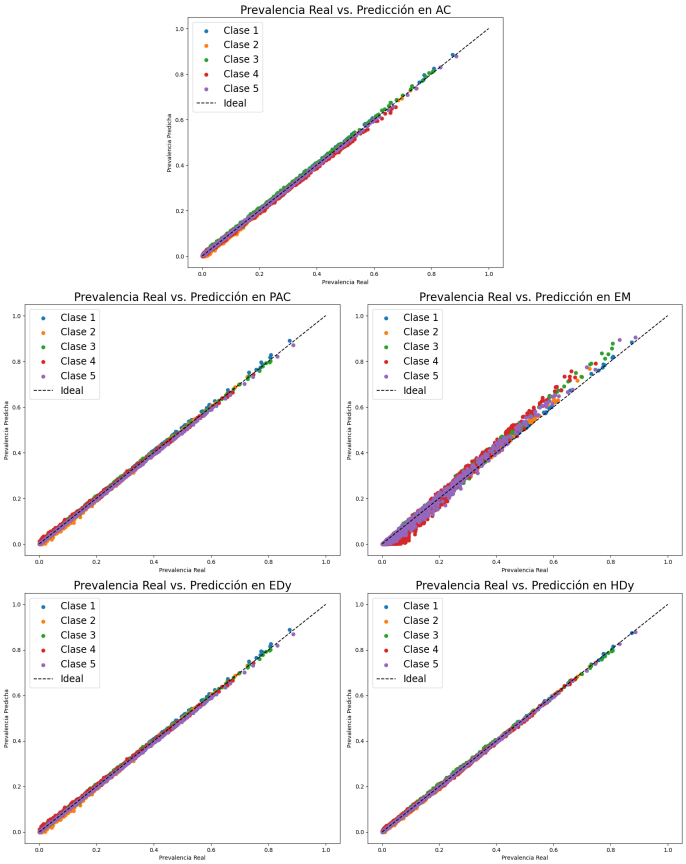

Fig. 5. Comparación entre las prevalencias reales y las predicciones de los métodos de AC, PAC, EM, HDy y EDy empleando las versiones convencionales en el conjunto de datos artificiales del caso 2

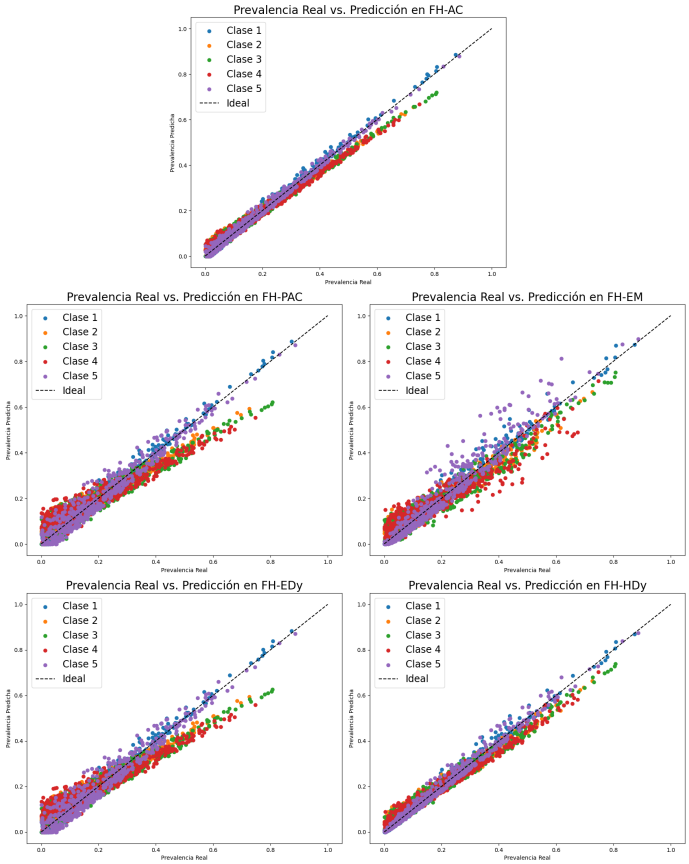

Fig. 6. Comparación entre las prevalencias reales y las predicciones de los métodos de AC, PAC, EM, HDy y EDy empleando las versiones F&H en el conjunto de datos artificiales del caso 2

problemas binarios como multi-clase, algunos de estos últimos de tipo ordinal.

El objetivo de LeQua 2024 (la segunda competición de Learning to Quantify) es permitir la evaluación comparativa de métodos para "aprender a cuantificar" en conjuntos de datos textuales, es decir, métodos para entrenar predictores de las frecuencias relativas de las clases de interés en conjuntos de documentos textuales sin etiquetar.

LeQua 2024 ofrece cuatro tareas (T1 a T4). Para cada una de estas tareas, se proporcionan datasets ya convertidos en forma vectorial. Como el foco de este trabajo son los cuantificadores ordinales, se emplean los datos de la tarea T3 que es la que corresponde para el caso ordinal. Esta tarea se ocupa de evaluar cuantificadores ordinales, es decir, cuantificadores que operan sobre un conjunto de $n > 2$ clases totalmente ordenadas; los datos utilizados se ven afectados por prior probability shift (también conocido como label shift).

Los datasets consisten en reseñas de productos del sitio web de Amazon, ya convertidas en formato vectorial. En la Tarea T3, la etiqueta de clase de cada reseña es la polaridad original basada en el sentimiento, expresada como una calificación de estrellas (de "1 estrella" a "5 estrellas"). Se le pide al predictor que prediga cuántas de las reseñas en la muestra de test tienen una determinada etiqueta de clase, y que lo haga para todas las clases del conjunto. Se dispone de 100 archivos destinados a entrenamiento que constan de 257 columnas y 200 ejemplos cada uno. La primera columna, contiene una identificación numérica de base cero para la etiqueta. Las 256 columnas restantes, etiquetadas del 0 al 255, son las 256 dimensiones de los vectores que representan el contenido de los documentos. Asimismo, se dispone de 1000 archivos de test los cuales disponen de 256 columnas que corresponden con dimensiones de los vectores que representan el contenido de los documentos y 200 ejemplos cada uno. Finalmente, se dispone de un archivo con las prevalencias reales para todos los conjuntos de test.

Tras lanzar el experimento se obtienen los resultados de la Tabla I. Como se puede observar, los modelos Frank and Hall tienen más dificultades para predecir las prevalencias correctas. Como ya se ha visto en el caso Artificial 2 (que intentaba simular datos reales), los métodos convencionales consiguen valores de EMD más bajos (mejores). Un factor importante a tener en cuenta es el sesgo en los propios datos, el dataset proviene de las reseñas de los usuarios en Amazon, por lo que es esperable que reseñas de 1 o 5 estrellas sean más comunes que las demás. Este aspecto juega un papel importante, tal y como se ha visto en el caso artificial 2, los métodos F&H tienden a introducir sesgos en las clases

TABLE I
COMPARACIÓN DE EMD DE LOS RESULTADOS OBTENIDOS EN LA TAREA T3 DE LEQUA

Método	EMD
AC	**0.10677**
FH-AC	0.22217
PAC	**0.11346**
FH-PAC	0.13417
EM	**0.11134**
FH-EM	0.22728
EDy	**0.11166**
FH-EDy	0.14403
HDy	**0.1056**
FH-HDy	0.16041

TABLE II
COMPARACIÓN DE EMD DE LOS RESULTADOS OBTENIDOS EN DIFERENTES DATASETS

Method	ESL	LEV	SWD	bostonhousing	abalone
AC	0.48352	0.51604	0.38505	0.21369	0.75811
FH-AC	0.31354	0.54485	0.45720	0.18502	1.30454
PAC	0.22652	0.72200	0.32999	0.12021	0.91858
FH-PAC	0.44142	0.48022	0.33904	0.22573	1.15396
EM	2.00087	2.00094	1.49978	0.11468	4.45345
FH-EM	1.00736	1.00542	0.74554	0.20867	2.25395
EDy	0.13116	0.25265	0.21357	0.11094	0.56207
FH-EDy	0.36638	0.43977	0.30862	0.21630	1.09975
HDy	0.27232	0.58865	0.28911	0.14281	0.90444
FH-HDy	0.30971	0.63312	0.41181	0.20483	0.94579

intermedias, por lo que el desbalanceo en las clases agrava este comportamiento empeorando el resultado obtenido por cuantificadores que empleen este método.

E. Caso Real 2

Tras realizar los experimentos empleando datos artificiales, y con datos reales provenientes de la competición LeQua se decide experimentar empleando datasets conocidos comunmente empleados en tareas de Machine Learning.

Se experimentará con 5 datasets: *Employee Selection* (ESL), *Lecture Evaluation* (LEV), *Social Workers Decisions* (SWD), *Boston Housing* y *Abalone*. Estos datasets se han obtenido de diferentes fuentes, ESL [20], LEV [21], SWD [22], Boston housing y Abalone [23].

La metodología aplicada será la misma que con los experimentos artificiales: se emplea un Grid Search para buscar hiperparametros del estimador subyacente a los modelos, se implementan todos los modelos con el mismo estimador subyacente y, finalmente, se compararán los resultados empleando la métrica EMD.

El dataset ESL contiene 488 perfiles de solicitantes de determinados puestos de trabajo industriales. Los psicólogos expertos de una empresa de contratación determinaron los valores de los atributos de entrada basándose en los resultados de pruebas psicométricas y entrevistas con los candidatos. El resultado es una puntuación global correspondiente al grado de aptitud del candidato para este tipo de trabajo. El dataset LEV contiene datos de una encuesta de evaluación de cursos universitarios, donde los estudiantes califican diferentes aspectos de un curso. Las calificaciones suelen ser ordinales y reflejan la naturaleza ordenada de las respuestas (p. ej., deficiente, regular, buena, muy buena, excelente). El dataset SWD contiene evaluaciones del mundo real de trabajadores sociales calificados sobre el riesgo que corren los niños si se quedan con sus familias en casa. Cuenta con 1000 ejemplos y 10 atributos. El dataset Boston Housing contiene información sobre diversos atributos de las casas en Boston. Cuenta con 506 ejemplos y 13 atributos. El dataset abalone se recogen datos para predecir la edad del abulón (una familia de moluscos gasterópodos) a partir de medidas físicas. La edad del abulón se determina cortando la concha a través del cono, teñiéndola y contando el número de anillos (clase a predecir) a través de un microscopio. Cuenta con con 4177 ejemplos y 8 atributos. Como la clase a predecir es el numero de anillos, existe un amplio abanico de clases disponibles. Como se trata de un numero muy elevado de clases, se decide reducirlo normalizando el conjunto a 10 clases.

Se puede observar en la Tabla II un resumen de los resultados obtenidos para el conjunto de datasets empleados usando EMD como métrica de error. Generalmente, la hipótesis planteada en el trabajo se sostiene: los modelos que emplean F&H obtienen peores resultados que sus contrapartes convencionales. No obstante, se pueden realizar algunas observaciones.

Se observa que, de los métodos comparados, los dos que consistentemente siempre obtienen mejores resultados que sus contrapartes Frank and Hall, son HDy y EDy. Esto es debido a que son metodos mas robustos y que, al realizar las predicciones mediante la similitud entre las distribuciones, los hace inherentemente más adecuados para manejar superposiciones porque evalúan la alineación general de las distribuciones de clases en lugar de depender de decisiones binarias individuales.

Por otro lado, EM obtiene los resultados mas pobres, esto se debe a que el método EM necesita que las probabilidades devueltas por el clasificador sean muy precisas. Si se da este caso funciona muy bien, que en el experimento se puede ver en con los resultados del dataset *Boston Housing*, donde consigue resultados parejos con el resto de modelos y mejora a su versión Frank and Hall. Por el contrario, si las probabilidades devueltas por el clasificador subyacente son poco precisas, el algoritmo devuelve resultados pobres, tal y como se puede apreciar en el resto de datasets.

V. CONCLUSIONES

Este trabajo extiende la hipótesis de que los métodos de descomposición no son adecuados para problemas de cuantificación a problemas de tipo ordinal. Principalmente se busca refutar la afirmación de que $P(x|y)$ permanece constante y que no hay un reparto homogéneo en las clases complementarias. Para ello, se introduce la cuantificación y su variante ordinal, los métodos de descomposición y el método especifico para problemas ordinales: Frank and Hall. Seguidamente, se presentan los métodos AC, PAC, EM, EDy y HDy con los que se

van a corroborar la hipótesis inicial, indicando características y su funcionamiento. Adicionalmente se expone el método CC al ser este una base para algunos de los métodos empleados.

Con el fin de contrastar la hipótesis se realizan dos experimentos, uno con datos artificiales y otro con datos reales, cada uno de ellos dividido en 2 casos, 4 en total. En el primero de los casos no se observa una diferencia significativa entre los métodos convencionales y F&H, no obstante se observa la consistencia de Fisher en un caso ideal.

Como el caso 1 representa un caso ideal que no se da en casos reales, se propone un segundo caso artificial con solapamiento de valores fronterizos en las clases. En este caso si se puede observar con claridad la diferencia entre los métodos convencionales y F&H. Debido a la naturaleza del método de descomposición, se sospecha que en dos de las clases predichas, las prevalencias obtenidas de los cuantificadores deberían acercarse mas al ideal que el resto (la primera y ultima clase). Esta comprobación refuerza la hipótesis planteada al inicio del trabajo (no hay reparto homogéneo en las clases complementarias) puesto que las dos únicas clases que se aproximan al valor real son las que se encuentran aisladas en las fases de la descomposición.

Finalmente se experimenta con datos reales, primero se emplean datos procedentes de LeQua2024, una competición de cuantificación reciente. Los resultados son similares a los obtenidos en el caso artificial dos, extendiendo la hipótesis a casos reales. Por ultimo, se emplean dataset comúnmente utilizados en la literatura como banco de pruebas final. Los resultados reflejan como los métodos basados en comparar distribuciones funcionan mejor que el resto de modelos comparados.

Se extiende la hipótesis de que $P(x|y)$ no permanece constante en todos los casos y por tanto las variaciones entre prevalencias no es homogénea, al caso ordinal. Se comprueba de este modo que el método de descomposición Frank and Hall no mejora a los métodos convencionales en este tipo de problemas.

REFERENCES

[1] G. Forman, "Counting positives accurately despite inaccurate classification," in *Proceedings of ECML'05*, 2005, pp. 564–575.

[2] ——, "Quantifying trends accurately despite classifier error and class imbalance," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. Association for Computing Machinery, 2006, pp. 157–166. [Online]. Available: https://doi.org/10.1145/1150402.1150423

[3] P. González, A. Castaño, N. V. Chawla, and J. J. D. Coz, "A review on quantification learning," *ACM Comput. Surv.*, vol. 50, no. 5, sep 2017. [Online]. Available: https://doi.org/10.1145/3117807

[4] A. Esuli and F. Sebastiani, "Sentiment quantification," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 72–75, 2010.

[5] A. Esuli, A. Moreo, F. Sebastiani, and E. Cambria, "Cross-lingual sentiment quantification," *IEEE Intelligent Systems*, vol. 35, no. 3, pp. 106–114, may 2020. [Online]. Available: https://doi.org/10.1109/MIS.2020.2979203

[6] W. Gao and F. Sebastiani, "From classification to quantification in tweet sentiment analysis," *Social Network Analysis and Mining*, vol. 6, no. 1, pp. 1–22, 2016.

[7] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," *Data Mining and Knowledge Discovery*, vol. 18, no. 1, pp. 30–55, Feb 2009. [Online]. Available: https://doi.org/10.1007/s10618-008-0116-z

[8] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Machine Learning: ECML 2001*, P. De Raedt, Lucand Flach, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 145–156.

[9] A. Esuli, A. Moreo Fernández, and F. Sebastiani, "Learning to quantify: Lequa 2024 datasets," Feb. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10654475

[10] G. Forman, "Quantifying counts and costs via classification," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, pp. 164–206, 2008.

[11] A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana, "Quantification via probability estimators," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 737–742.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://www.jstor.org/stable/2984875

[13] M. Saerens, P. Latinne, and C. Decaestecker, "Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure," *Neural Computation*, vol. 14, no. 1, pp. 21–41, 01 2002. [Online]. Available: https://doi.org/10.1162/089976602753284446

[14] V. González-Castro, R. Alaiz-Rodríguez, and E. Alegre, "Class distribution estimation based on the hellinger distance," *Information Sciences*, vol. 218, pp. 146–164, 2013.

[15] A. Castaño, L. Morán-Fernández, J. Alonso, V. Bolón-Canedo, A. Alonso-Betanzos, and J. J. del Coz, "An analysis of quantification methods based on matching distributions," *Pattern Recognition*, p. 35, 2021.

[16] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: a unifying approach for margin classifiers," *J. Mach. Learn. Res.*, vol. 1, pp. 113–141, sep 2001. [Online]. Available: https://doi.org/10.1162/15324430152733133

[17] S. Destercke and G. Yang, "Cautious ordinal classification by binary decomposition," in *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2014, pp. 323–337.

[18] A. Castaño, J. Alonso, P. González, P. Pérez, and J. J. del Coz, "Quantificationlib: A python library for quantification and prevalence estimation," *SoftwareX*, vol. 26, p. 101728, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352711024000992

[19] A. Castaño, P. González, J. A. González, and J. J. del Coz, "Matching distributions algorithms based on the earth mover's distance for ordinal quantification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022.

[20] U. of Waikato, "Weka datasets," 2021. [Online]. Available: https://waikato.github.io/weka-wiki/datasets

[21] M. Kelly, R. Longjohn, and K. Nottingham, "The uci machine learning repository," 2019. [Online]. Available: http://archive.ics.uci.edu/ml

[22] OpenML, "Openml datasets," 2021. [Online]. Available: https://www.openml.org/

[23] C. Wei, "Ordinal regression datasets," 2021. [Online]. Available: http://www.gatsby.ucl.ac.uk/ chuwei/ordinalregression.html