Actas de las Jenui, vol. 7. 2022.
Páginas: 199-206

Guiando la intervención del profesorado en la evaluación por pares explotando un modelo gráfico probabilístico

Jerónimo Hernández-González
Departament de Matemàtiques i Informàtica
Universitat de Barcelona
08007 Barcelona
jeronimo.hernandez@ub.edu

Pedro Javier Herrera
Dpt. de Ingeniería del Software y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
28040 Madrid
pjherrera@issi.uned.es

Resumen

Entre las metodologías activas de enseñanza-aprendizaje, la evaluación por pares, donde los estudiantes se valoran mutuamente, destaca como técnica de evaluación formativa. Aunque también podría usarse para calificar al propio estudiante, existen dudas sobre la fiabilidad de las calificaciones obtenidas de los pares. Bajo la hipótesis de que se puede modelar la evaluación por pares para guiar eficientemente al profesorado en la elección de qué actividades revisar, el objetivo es obtener una estimación fiable de la calificación de las actividades que el docente no ha revisado. Se usa un modelo gráfico probabilístico para el modelado, y un método de aprendizaje automático con aproximación Bayesiana que se ajusta con las calificaciones de los pares y el profesorado. Se propone un procedimiento que sugiere, uno a uno, qué trabajo debería calificar el docente para reducir la incertidumbre en el modelo. El docente decide cuántos trabajos calificar según su propio criterio de incertidumbre tolerable. Esta propuesta, validada en datos reales, muestra resultados prometedores. Tiene el potencial impacto de ayudar a extender la evaluación por pares como técnica de evaluación y calificación, reduciendo las dudas entre el profesorado acerca de la fiabilidad de las calificaciones obtenidas.

Abstract

Among the active teaching-learning methodologies, formative assessment technique of peer assessment, where students assess each other, stands out. Although this technique could be used to grade students too, there are doubts about the reliability of the ratings given by peers. Under the hypothesis that peer assessment can be modeled to efficiently guide teachers in choosing which activities to correct, the objective is to obtain a reliable estimate of the grade of the tests that the teacher has not reviewed. Probabilistic graphical models are used for modeling, together with a Bayesian machine learning method that adjusts to peer and teacher ratings. A procedure is proposed that suggests, one by one, which work the teacher should grade next to reduce the uncertainty in the model. The teacher decides how many activities to grade based on their own criterion of tolerable uncertainty. This proposal, validated in real data, shows promising results and has the potential impact of helping to extend peer evaluation as an evaluation and grading technique, reducing doubts among teachers about the reliability of the grade estimates.

Palabras clave

Evaluación por pares, modelos gráficos probabilísticos, aprendizaje automático, inferencia estadística, optimización carga de trabajo.

1. Introducción

Las nuevas metodologías de enseñanza-aprendizaje vienen a transformar las técnicas tradicionales de la educación para mejorar el proceso de aprendizaje, por ejemplo, re-concibiendo la evaluación como un proceso formativo. Entre las metodologías de evaluación formativa, el proceso en el que el alumnado participa en la evaluación corrigiendo el trabajo de sus compañeros se conoce como *evaluación por pares* o co-evaluación [14, 17]. Para llevarla a cabo, el alumnado necesita desarrollar destrezas y conocimientos alternativos a los tradicionales en el proceso de aprendizaje. Destaca por su capacidad para promover el aprendizaje, involucrar al alumnado de manera activa, desmitificar la evaluación, aumentar la autoconsideración, etc. [3, 4, 12, 13]

Desde el punto de vista del profesorado, aunque ayudaría a reducir su carga de trabajo [13], se plantea el dilema sobre la conveniencia de usar o no la evaluación por pares para calificar al alumnado. Se podría evaluar tanto a la prueba que realiza, como al *feedback*

que provee como evaluador. Además, la calificación de los pares suele ser imprecisa en diferente medida y en relación con diferentes factores (conocimiento del evaluador, implicación, rúbrica disponible, etc.). En la práctica, sólo se suele evaluar la actividad, y cada actividad es evaluada por varios pares para, mediante agregación, intentar obtener una calificación más precisa. Entre otros cuestionamientos [2], el profesorado esgrime el trabajo extra que le supone organizar este tipo de evaluación y la agregación final de toda la información para rechazar el uso de la evaluación por pares.

En la práctica, lo ideal para compensar posibles sesgos individuales sería pedir a cada alumno que evalúe a *todos* sus pares. Pero este escenario no es realista. Si se quiere validar el proceso, es necesario otro conocimiento, y el más accesible es el del profesorado. Éste puede evaluar y determinar la calidad de cada actividad. Nótese la contradicción: se querría usar la evaluación por pares para calificar (y ahorrar esta tarea al docente), pero para validarlo se recurre al docente.

En este trabajo se modela a la clase en una actividad de evaluación por pares con el objetivo de realizar la agregación de las calificaciones de manera informada. Para ello, usaremos un modelo gráfico probabilístico (PGM, por sus siglas en inglés) [7] para modelar la relación entre el conocimiento de los anotadores y la calidad de su trabajo y las evaluaciones a sus pares. Modelando la competencia del estudiante, la agregación se apoya en mayor medida en la evaluación del alumnado competente. Modelos de diferente complejidad intentan abarcar más o menos fenómenos relacionados con el desempeño del alumnado. Aquí, usaremos dos modelos, con diferentes hipótesis subyacentes, originalmente propuestos por Piech et al. [10] y usados exclusivamente para la agregación de la calificación final.

Más allá de la agregación estática de las calificaciones, nuestra propuesta implica un proceso dinámico en que el docente revisa y califica un subconjunto de actividades para calibrar y validar la estimación del modelo. Los PGMs permiten lidiar con la incertidumbre e incluso estimarla. Esto es clave en la propuesta, ya que permitiría al docente revisar tantas actividades como considere hasta que el nivel de incertidumbre baje de cierto umbral de tolerancia, el cual fijará según su propio criterio.

El proceso de aprendizaje automático se nutre inicialmente de las calificaciones asignadas por los pares. Nuestra propuesta, basada en el aprendizaje automático activo [9], consiste en pedir al docente la calificación, paso a paso, de una actividad. La introducción y propagación de esta calificación, completamente veraz, en el modelo permitirá recalibrar la competencia de cada estudiante y, en consecuencia, las agregaciones de las notas finales. En este estudio se contraponen dos técnicas para sugerir al docente qué examen revisar en cada momento: una completamente aleatoria, y otra basada en la varianza de las calificaciones de los pares.

Este trabajo presenta una validación inicial del método con datos reales de dos casos de estudio en la etapa educativa universitaria, con resultados positivos. Aunque la metodología es suficientemente general para ser aplicada en cualquier clase, la generalización de estos resultados a cualquier contexto universitario, o incluso a otras etapas educativas donde se emplee la evaluación por pares, necesitará de un estudio de validación exhaustivo.

Desde el punto de vista del profesor que quiere usar las evaluaciones por pares de sus alumnos para calificar los exámenes, este método aporta confianza y rigor a la estimación, y deja en manos del docente la decisión última de cuántas actividades corregir. En otras palabras, permite establecer un compromiso entre la confianza del profesor en las calificaciones estimadas y la carga de trabajo que asume. El aprovechamiento será mayor a medida que aumenta el tamaño de la clase, o el número de instrumentos de evaluación empleados. Además, tiene el potencial de fomentar el aprovechamiento de la evaluación por pares para calificar, incluso por docentes precavidos que se resisten a considerar las calificaciones de los pares para este fin. Así, podría suponer un pequeño respaldo para la implantación de esta metodología.

Este documento continúa con la descripción del método y los materiales (datos). Posteriormente, se presenta el análisis experimental. Se acaba con la discusión, conclusiones e ideas de trabajo futuro.

2. Materiales y métodos

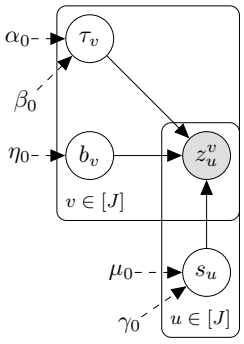
En esta sección se presentan los diferentes componentes del método que guía al profesorado en la supervisión de una prueba de evaluación por pares. Se modela la clase y el alumnado, junto con su desempeño en la prueba, mediante un modelo gráfico probabilístico. Se usa una técnica de selección para elegir qué actividad (de qué alumno) se sugiere al docente corregir. Dada la calificación asignada por el profesor a cierta actividad, se utiliza inferencia Bayesiana estándar para actualizar el modelo[1].

El material utilizado para validar la propuesta está compuesto por los dos conjuntos de datos que se presentan más abajo.

2.1. Modelos

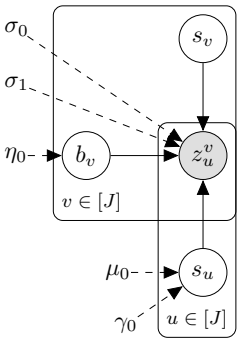
Un modelo gráfico probabilístico (PGM, por sus siglas en inglés) es una herramienta matemática que co-

[1] Implementado con el *software* estadístico STAN [16]



para $v \in \{1, \ldots, J\}$:
$\tau_v \sim \mathcal{G}(\alpha_0, \beta_0)$
$b_v \sim \mathcal{N}(0, 1/\eta_0)$
para $u \in \{1, \ldots, J\}$:
$s_u \sim \mathcal{N}(\mu_0, 1/\gamma_0)$
para $v \in \{1, \ldots, J\}$:
$z_u^v \sim \mathcal{N}(s_u + b_v, 1/\tau_v)$

Figura 1: PG_1 y su proceso generativo asociado. Cada estudiante, como evaluador, tiene fiabilidad τ_v y sesgo b_v individualmente. Cada estudiante, como examinado, alcanza una calificación real s_u (desconocida), y ésta también determina la calificación que le asigna cada evaluador v.



para $v \in \{1, \ldots, J\}$:
$b_v \sim \mathcal{N}(0, 1/\eta_0)$
para $u \in \{1, \ldots, J\}$:
$s_u \sim \mathcal{N}(\mu_0, 1/\gamma_0)$
para $v \in \{1, \ldots, J\}$:
$z_u^v \sim \mathcal{N}(s_u + b_v, 1/(\sigma_1 \cdot s_v + \sigma_0))$

Figura 2: PG_3 y su proceso generativo asociado. Cada estudiante, como evaluador, tiene un sesgo b_v individual. Cada estudiante, como examinado, alcanza una calificación real s_u (desconocida), que puede entenderse también como grado de conocimiento. La calificación que le asigna un evaluador v a un compañero u está determinada por el grado de conocimiento de ambos, s_u y s_v, junto con el sesgo del evaluador b_v.

difica dependencias condicionales entre variables aleatorias mediante un grafo, y usa un conjunto de factores para parametrizar la distribución de probabilidad. Concretamente, se usan redes Bayesianas, unos PGMs con grafos dirigidos y acíclicos, y cuyos factores son distribuciones de probabilidad condicionada de una variable aleatoria dadas las variables padre en el grafo.

Diferentes PGMs han sido propuestos específicamente para modelar la evaluación por pares [1, 8, 10, 11, 15]. En este trabajo, nos centramos en dos modelos propuestos por Piech et al. [10] (Figuras 1 y 2), elegidos por su simplicidad y la claridad con que plasman las hipótesis subyacentes. Nótese que, mientras Piech et al. [10] y otros trabajos previos usan estos modelos de manera estática para agregar una calificación final a partir de las calificaciones de los pares para cada actividad, en este estudio los modelos se usan y reajustan de manera dinámica para guiar al profesorado en la supervisión de la prueba y afinar la agregación de las calificaciones.

En el contexto de una clase de J estudiantes, el modelo PG_1 (Figura 1) asume que cada estudiante alcanza un nivel de destreza (τ_v) y desarrolla un sesgo (b_v) de manera individual como evaluador. Se asume que, si supiésemos la calificación real s_u de la actividad del estudiante u, podríamos estimar la calificación que le asignaría su par v aplicando sobre la nota real s_u el sesgo de éste, b_v, y considerando cierta variabilidad inversamente proporcional a su destreza como evaluador, τ_v. Las variables originalmente observadas (sólo la z, las calificaciones de los pares) se muestran sombreadas, mientras el resto son variables latentes. Los parámetros de las distribuciones a priori del modelo se representan por letras griegas y líneas discontinuas. Los recuadros indexados representan que las variables que agrupan se repiten tantas veces como se indica ($|S|$, el número total de estudiantes). Lo que cambia entre índices u y v es simplemente su consideración como estudiante evaluado o evaluador.

PG_3 (Figura 2) introduce una nueva hipótesis: el estudiante que realiza una buena prueba tiene más probabilidades de realizar mejores (más precisas) calificaciones a sus pares. En la práctica, esto supone que la calificación real del par evaluador s_v es la medida de precisión (sustitutiva de τ_v en PG_1). Se asume que, si supiésemos las calificaciones reales s_u y s_v, podríamos estimar la calificación que le asignaría al estudiante u su par v aplicando sobre la nota real s_u el sesgo de éste, b_v, y considerando cierta variabilidad inversamente proporcional a su calificación, s_v. Nótese que s_u y s_v representan copias de la misma variable. Se muestra ambas copias para enfatizar los dos usos de la variable: como calidad del estudiante examinado y como calidad del evaluador.

El otro modelo propuesto en [10], PG_2, no ha sido considerado ya que requiere más de una prueba evaluada por pares con el mismo grupo, un escenario que no se ha contemplado en este trabajo.

2.2. Técnicas de selección

Para sugerir al profesor cuál es la siguiente actividad que debería corregir, se pueden considerar diversas técnicas de selección de mayor o menor complejidad. En el ámbito del aprendizaje automático activo [9], se ha propuesto un gran cantidad de ellas. En este trabajo, exploramos el uso de dos técnicas diferentes:

RND: La selección es completamente aleatoria entre las pruebas que el docente no ha revisado todavía. Esta técnica nos servirá para trazar un resultado base para la comparación.

GrV: La selección se basa en el criterio de máxima va-

rianza: la siguiente actividad a corregir es, entre las pruebas que el docente no ha revisado todavía, aquella que ha recibido calificaciones de sus pares con una mayor varianza.

2.3. Datos

Para el presente estudio, hemos recabado datos reales de dos entornos diferentes, ambos de estudios relacionados con la informática en el nivel universitario de máster. Los conjuntos están públicamente disponibles[2] tras la oportuna anonimización de los datos.
MAIB: En primer lugar, se recogieron datos de una clase de $J = 16$ alumnos sobre un examen de desarrollo. Tras el examen, cada estudiante recibió y evaluó $G = 3$ exámenes realizados por sus compañeros. El reparto se realizó de manera aleatoria garantizando que todos corrigen 3 pruebas y todos reciben 3 evaluaciones de su propia prueba. El Cuadro 1 muestra el resumen de las calificaciones obtenidas en este grupo.
MPD: En segundo lugar, se recogieron datos de una clase de $J = 16$ alumnos sobre un examen de desarrollo. Tras el examen, cada estudiante recibió y evaluó *todos* los exámenes realizados por sus compañeros. El Cuadro 2 muestra el resumen de las calificaciones obtenidas en este grupo.

En ambos casos, el alumnado disponía de una rúbrica desde el momento de la prueba, que les podría servir para saber cómo serían evaluados. Se les indicó que debían seguir la misma rúbrica para revisar la prueba de sus compañeros.

2.4. Métricas

Para la validación de la propuesta se compararán las calificaciones reales de las actividades presentadas por los estudiantes y las estimaciones obtenidas con el modelo. Asumiremos que las calificaciones reales son las asignadas por el docente (disponibles en ambos conjuntos de datos). Se usarán dos métricas distintas:
RECM: mide la diferencia entre las calificaciones reales, s_u, y las estimadas por el modelo, $\hat{s}_u$, como la raíz cuadrada del valor medio del cuadrado de las diferencias individuales (por actividad/estudiante):

$$RECM = \sqrt{\frac{1}{J}\sum_{u=1}^{J}(\hat{s}_u - s_u)^2} \qquad (1)$$

El RECM es siempre no negativo, donde un error de 0 identifica una estimación perfecta, y es sensible a valores atípicos que producen grandes errores (no apropiado para comparaciones de datos en diferente escala). En este caso, comparamos calificaciones entre 0 y 10, y el RECM nos proporciona una medida de la diferencia *numérica* entre las calificaciones reales y las estimadas.
Coeficiente τ de Kendall [6]: es una medida de correlación de rango de dos muestras que compara, tras ordenarlas, en qué grado se parecen los rankings obtenidos. Es decir, no tiene en cuenta los valores absolutos, sino relativos. La fórmula que define esta métrica es:

$$\tau = \frac{2}{J(J-1)} \sum_{\substack{t<u: \\ t,u\in\{1..J\}}} sgn(s_t - s_u)sgn(\hat{s}_t - \hat{s}_u) \qquad (2)$$

donde $sgn(exp)$ es la función que devuelve el signo de la expresión exp. Intuitivamente, τ es 1 cuando el orden de las actividades es el mismo de acuerdo con las calificaciones reales o con las estimaciones, y -1 cuando el orden es el inverso. Un valor de 0 representaría dos muestras con una coincidencia que podría ser obtenida con un orden aleatorio. En nuestro caso, el coeficiente de correlación τ de Kendall nos proporciona información relevante sobre la ordenación de las actividades según la calificación estimada. Sería razonable aceptar un pequeño error numérico entre calificaciones reales y estimadas, si se mantiene una relación de ordenación justa: se da mayor calificación a quien mejor lo haya hecho. Esta métrica nos permite analizar este comportamiento sin entrar a valorar la calificación numérica exacta asignada.

3. Experimentos

El procedimiento propuesto se ha validado en los dos conjuntos de datos presentados en la Sección 2.3. Para simular el proceso de revisión que llevaría a cabo un docente, inicialmente no se ajusta el modelo con las calificaciones (reales) de éste. Se asume que el docente siempre atenderá a la sugerencia del método sobre qué actividad revisar a continuación. Así, a cada paso, se reajusta el modelo con la calificación que le asignaría el docente a la actividad que sugiere nuestro procedimiento. Con el nuevo ajuste del modelo, se reestima la calificación de las pruebas que el docente todavía no ha revisado.

3.1. Experimentos con MAIB

Diseño experimental: Se han considerado los dos modelos, PG_1 y PG_3, presentados en la Sección 2.1, propuestos en [10]. Entre ambos, la principal diferencia es que PG_3 aprovecha que el estudiante es también evaluador y asume que quien realiza bien la actividad evaluará de manera más precisa. Para la selección de los valores de los hiperparámetros de las distribuciones de probabilidad a priori se buscaba imponer el uso de distribuciones poco informativas que afecten lo mínimo

[2] https://jhernandezgonzalez.github.io/supp_data_peer.html

		Alumnos															
Revisor		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Docente		6,90	8,73	9,03	9,47	9,77	8,93	8,73	7,93	9,37	8,40	8,70	7,67	8,33	7,83	7,40	8,13
Pares	Media	8,74	9,17	9,61	9,63	9,66	9,52	9,29	8,84	9,70	9,50	9,32	8,01	9,74	9,09	8,59	9,24
	Desv.	0,69	0,80	0,48	0,35	0,36	0,54	0,30	0,93	0,29	0,40	0,34	1,09	0,19	0,51	1,95	0,53

Cuadro 1: Calificación asignada por el docente a los alumnos, así como por sus pares (valor medio y desviación típica, Desv.), en el conjunto de datos MAIB.

		Alumnos															
Revisor		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Docente		9,00	5,00	7,00	7,00	6,00	7,00	8,00	9,00	5,00	5,00	6,00	7,00	7,00	8,00	9,00	6,00
Pares	Media	9,00	6,13	7,27	8,80	7,00	7,67	7,67	9,00	6,33	6,60	6,47	7,53	7,80	7,93	8,60	6,33
	Desv.	0,97	0,88	0,85	1,05	1,15	0,87	1,45	1,03	1,53	1,40	1,36	1,09	0,75	0,85	1,08	1,45

Cuadro 2: Calificación asignada por el docente a los alumnos, así como por sus pares (valor medio y desviación típica, Desv.), en el conjunto de datos MPD.

posible la discusión sobre el resto de factores condicionantes en este problema, y que permitan un aprendizaje a partir de los datos efectivo y no completamente condicionado. Así, se han utilizado valores medios estimados de los datos disponibles para fijar estos valores. Concretamente, se eligieron los siguientes valores para los hiperparámetros $\{\alpha_0 = 10; \beta_0 = 10; \eta_0 = 1; \mu_0 = 7{,}5; \gamma_0 = 1; \theta_0 = 0{,}5; \theta_1 = 0{,}2\}$

También se comparan los resultados con las dos técnicas de selección, RND y GrV, presentadas en la Sección 2.2. Para dar cuenta de la varianza de la selección aleatoria de RND, los resultados mostrados a continuación son el valor medio de 10 repeticiones de la simulación completa. Para evaluar el rendimiento del procedimiento, se comparan las calificaciones reales y las estimadas usando el RECM y el coeficiente τ de Kendall, tal y como fueron definidos en la Sección 2.4.

Resultados: En la Figura 3 se muestran los resultados de la simulación del proceso de revisión usando dos modelos diferentes para la clase y siguiendo dos técnicas de selección de la siguiente actividad a corregir. En el eje horizontal se simula el avance temporal de la revisión de la prueba en número de actividades ya calificadas por el docente, hasta un total de $J = 16$, el número de alumnos que participaron en la prueba donde se recogieron los datos MAIB.

Se observa que inicialmente el modelo PG_1 obtiene mejores resultados que el modelo PG_3 (siguiendo ambas técnicas de selección). Cuando se alcanzan las 4-5 actividades corregidas ($\sim 25\,\%$) y tras una rápida estabilización, los resultados de PG_3 ya son comparables a los de PG_1. Siguiendo la selección aleatoria de RND, las estimaciones de las calificaciones de las actividades restantes de ambos modelos son similares. Sin embargo, siguiendo la selección basada en la varianza de las calificaciones de los pares, GrV, los resultados mejoran con respecto a la selección RND: PG_1 mejora en términos de la τ de Kendall, aunque no en términos de RECM; PG_3, una vez superado el período inicial de estabilización, muestra resultados sustancialmente mejores que los obtenidos con PG_1 y con RND.

3.2. Experimentos con MPD

Diseño experimental: Siguiendo un diseño similar al presentado anteriormente para los datos MAIB, se comparan los dos modelos, PG_1 y PG_3, usando los valores de hiperparámetros siguientes, $\{\alpha_0 = 7; \beta_0 = 7; \eta_0 = 2; \mu_0 = 7; \gamma_0 = 0{,}5; \theta_0 = 0{,}04; \theta_1 = 0{,}15\}$, seleccionados con el mismo procedimiento que pretende imponer igualmente distribuciones a priori poco informativas que permitan un aprendizaje efectivo a partir de los datos. También se usan las dos técnicas de selección, RND y GrV, y los resultados se evalúan con el RECM y el coeficiente τ de Kendall.

En la prueba donde se recogieron los datos MPD, todos los estudiantes ($J = 16$) evaluaron a todos sus pares ($G = 15$). Como se explicaba anteriormente, éste no es un escenario habitual, pero nos permite explorar la importancia de otros factores, como la importancia del número de evaluaciones de pares, G. Cuando se fija un valor de G experimental menor que el real (p.ej., $G = 3$), se eligen aleatoriamente, entre las disponibles, G calificaciones por actividad y G por estudiante como evaluador (p.ej., cada estudiante realiza $G = 3$ evaluaciones a sus pares, y recibe $G = 3$ calificaciones de sus pares). El comportamiento del procedimiento se evalúa en el subconjunto resultante. Para dar cuenta de la varianza de este submuestreo aleatorio, los resultados mostrados a continuación son el valor medio de 10

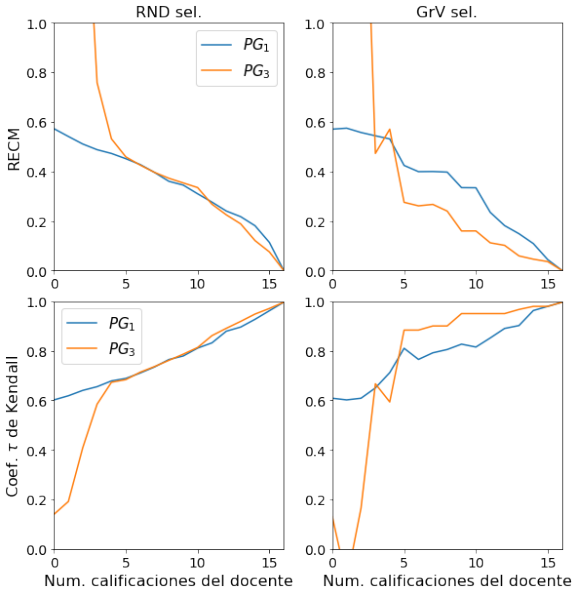


Figura 3: Comparación del rendimiento de los modelos PG_1 y PG_3 sobre los datos MAIB. Se muestra, para las técnicas de selección RND y GrV (por columnas), resultados en términos de RECM y el coeficiente τ de Kendall (por filas). Cada línea en las gráficas muestra la evolución del rendimiento a medida que el docente califica más actividades. En MAIB, $J = 16$ estudiantes revisan $G = 3$ actividades cada uno.

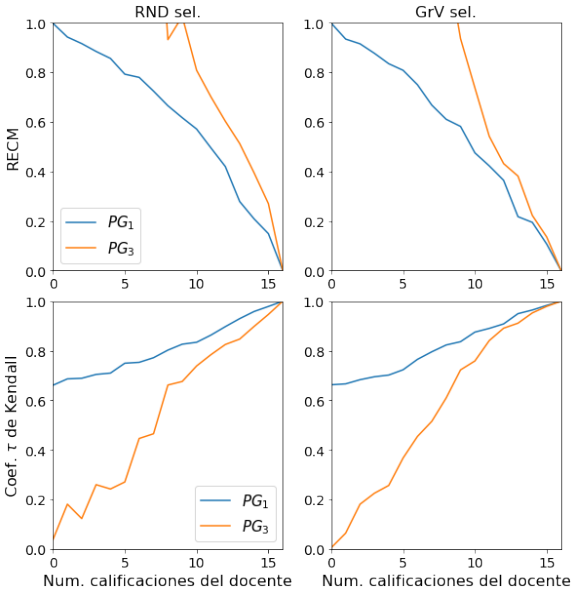


Figura 4: Comparación del rendimiento de los modelos PG_1 y PG_3 sobre los datos MPD. Se muestra, para las técnicas de selección RND y GrV (por columnas), resultados en términos de RECM y el coeficiente τ de Kendall (por filas). Cada línea en las gráficas muestra la evolución del rendimiento a medida que el docente califica más actividades. Se muestra el caso en que $J = 16$ estudiantes revisan $G = 3$ actividades cada uno.

repeticiones de la simulación completa con diferente muestreo.

Resultados: Siguiendo el mismo estilo de gráfica que en la subsección anterior, en la Figura 4 se muestran los resultados de la simulación del proceso de revisión, con los datos MPD y fijando $G = 3$, para los dos modelos y dos técnicas de selección.

Se observa que el modelo PG_1 obtiene mejores resultados que el modelo PG_3 (siguiendo ambas técnicas de selección) a lo largo de casi todo el procedimiento/simulación. Siguiendo la selección aleatoria de RND, PG_3 nunca alcanza el rendimiento de PG_1 ni en términos de RECM ni en términos de τ de Kendall. Siguiendo la selección basada en la varianza de las calificaciones de los pares, GrV, los resultados mejoran con respecto a la selección RND principalmente para PG_3, que a partir de $10-11$ actividades revisadas por el docente ($65-75\,\%$), obtiene resultados comparables a los obtenidos con PG_1.

Con un estilo de gráfica similar, la Figura 5 muestra los resultados de la simulación del proceso de revisión con los datos MPD a medida que aumenta el número de revisiones de los pares que recibe cada actividad ($G = \{1, 3, 5, 7, 10, 15\}$), para cada modelo por separado.

Con ambos modelos y en términos de ambas métricas, los resultados tienden a mejorar a medida que aumenta el número de revisiones que cada actividad recibe. La mejora es sustancial cuando el número de revisiones es bajo (p.ej., al pasar de $G = 1$ a 3 revisiones), pero muestra un comportamiento asintótico y la ganancia se reduce a medida que aumenta el valor de G. Entre modelos, PG_1 muestra nuevamente su comportamiento estable en comparación con el de PG_3.

4. Discusión y conclusiones

El procedimiento que se propone en este trabajo guía al profesorado en la revisión de una prueba de evaluación por pares para calificar dicha prueba. Usa un modelo gráfico probabilístico para estimar las calificaciones del alumnado en base a las calificaciones de los pares y otras características personales (precisión, sesgo) estimadas de los datos. El modelo también permite obtener una medida de incertidumbre asociada al modelo. Esto, combinado con una técnica de selección que sugiere al docente qué actividad revisar a continuación, le permitirá a éste corregir tantas actividades como estime oportuno. Así, cada docente establecerá su propio compromiso entre carga de trabajo e incertidumbre tolerable.

Dados los resultados de la sección anterior, las estimaciones del modelo PG_1 parecen ser más fiables que las de PG_3 si el docente tiene contemplado revisar un

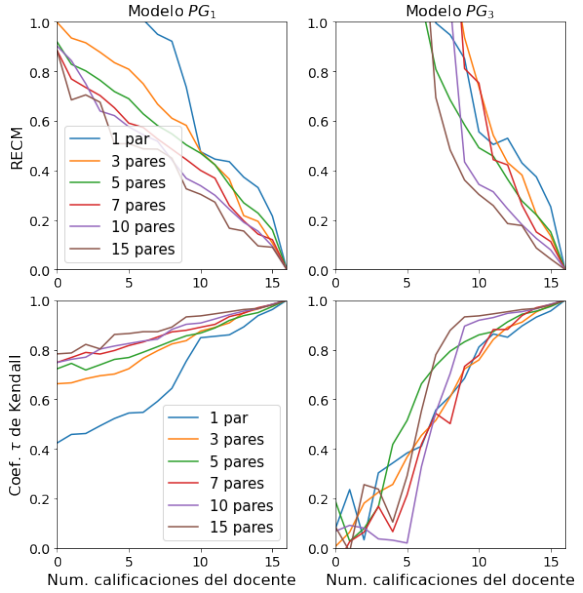


Figura 5: Análisis de la influencia del número de revisiones por actividad en el rendimiento de los modelos PG_1 y PG_3 sobre los datos MPD siguiendo la estrategia de selección GrV. Se muestra, para los 2 modelos (por columnas), resultados en términos de RECM y el coeficiente τ de Kendall (por filas). Cada línea en las gráficas muestra la evolución del rendimiento a medida que el docente califica más actividades. En MPD, el número de estudiantes es $J = 16$.

número pequeño de actividades ($< 25\,\%$). De entrada, las estimaciones de PG_1 tienen una alta correlación (coef. τ de Kendall) de al menos 0,6. Las estimaciones de PG_3 tienen inicialmente una baja correlación (cercana a 0) aunque, tras un período de estabilización, la correlación llega a niveles muy altos (0,9) con sólo un $25\,\%$ aproximadamente de las actividades revisadas en el conjunto de datos MAIB. Esto sería, de acuerdo con dichos resultados, una indicación de que el modelo PG_3 es una mejor opción si el docente está dispuesto a revisar más de un $25\,\%$ de las pruebas. Sin embargo, aunque el comportamiento robusto de PG_1 se sigue observando en los experimentos con el conjunto de datos MPD, las estimaciones de PG_3 sólo llegan a compararse con las de PG_1 (nunca superarlas) tras la incorporación de las calificaciones reales de un gran número de actividades por el docente. Los resultados en términos de RECM (diferencias numéricas) y coeficiente τ de Kendall (similitud de rango) son equivalentes en la mayoría de casos, aunque hacemos énfasis en la segunda métrica por ajustarse mejor a los objetivos del procedimiento: dados dos estudiantes, no se debería asignar mejor calificación a quien realmente obtendría una peor.

Con respecto a la técnica de selección de la siguiente actividad a revisar (sugerida al docente), sorprendentemente la técnica RND de selección completamente aleatoria pone una base bastante competitiva. Aún así, la técnica GrV basada en la varianza de las calificaciones mejora los resultados del procedimiento: inequívocamente cuando se usa el modelo PG_3, pero de manera más leve si se usa PG_1 (principalmente en términos de τ de Kendall). Esto parecería indicar que la técnica de selección que se emplea es realmente determinante, y que considerar los datos para tomar estas decisiones parece una decisión acertada.

El diseño de la prueba de evaluación por pares del conjunto MPD nos ha permitido también explorar el comportamiento del procedimiento según el número de calificaciones de los pares que se recogen. Como es habitual en entornos donde se combina la contribución de multiples actores de fiabilidad cuestionable (*crowdsourcing*, en inglés), a mayor número de contribuyentes, mejor resultado [5]. Se ha podido verificar que un mayor número de calificaciones de pares (mayor G) lleva a mejores estimaciones del modelo. Se trata de un resultado esperable, ya que se dispone de una mayor cantidad de datos para ajustar mejor los modelos. La mejoría al aumentar el número de evaluaciones de los pares es más relevante cuando el número absoluto es inicialmente corto. Cuando ya muchos pares han revisado una actividad, la aportación de uno nuevo no supone un gran cambio. Estos resultados sugieren que es interesante disponer del mayor número posible de evaluaciones de pares, pero que un gran número de ellas tampoco sería relevante. Esta conclusión es realmente apropiada para la práctica real, donde es poco razonable exigir al alumnado revisar un gran número de actividades.

Asimismo, es oportuno reflexionar sobre el ajuste de los hiperparámetros del modelo. Aparte de la elección de la técnica de selección y del propio modelo, fijar los hiperparámetros es el único requisito para poner en funcionamiento el procedimiento propuesto. Fijar los valores de los hiperparámetros de manera oportuna, ayudando a evitar el problema de la escasez de datos pero sin sesgar irremediablemente las estimaciones, no parece trivial. De hecho, una mala elección de hiperparámetros podría estar detrás de la desviación del comportamiento del modelo PG_3 con los datos MPD. En este se han utilizado valores medios estimados de los datos disponibles para fijar los valores de los hiperparámetros de ambos modelos. En el futuro será necesario estudiar en concreto el efecto de los hiperparámetros. Como resultado de dicho estudio futuro, sería deseable proveer una serie de indicaciones sobre cómo fijarlos que se puedan trasladar de manera intuitiva al docente-usuario final del método.

Nuestra propuesta se ha validado con dos conjuntos de datos reales de dos experiencias concretas en dos grupos clase de tamaño reducido ($J = 16$) en la eta-

pa universitaria. Esta validación ha permitido adquirir algunas intuiciones sobre la potencial contribución al docente para la gestión de la carga de trabajo extra que supone la supervisión de una tarea de evaluación por pares. Sin embargo, y aunque la metodología podría aplicarse en cualquier grupo-clase, un estudio más extensivo es necesario para evaluar la generalización de estos resultados a otros contextos o etapas educativas.

En el presente trabajo, el uso del modelo gráfico probabilístico tiene un uso concreto: estimar la calificación de aquellas actividades que el profesor todavía no ha revisado, y estimar la incertidumbre del modelo. El siguiente paso será usar el modelo para calcular cuál es la actividad que, si fuese corregida por el docente a continuación, supondría el mayor descenso de incertidumbre en el modelo. De esta manera, se usaría una técnica de selección basada en el modelo y no exclusivamente en los datos, como hasta ahora.

También se podría plantear llevar a cabo otras asignaciones del proceso de la evaluación por pares aprovechando el modelo; por ejemplo, escoger quién (o quiénes) es la persona más apropiada para corregir el trabajo de cierto estudiante de acuerdo a las características de ambos (evaluador y evaluado).

Agradecimientos

JHG es un profesor Serra Húnter. Este trabajo se enmarca en las actividades de los grupos de innovación IEData-UNED (GID2016-6) y ADALID-URJC. Ha sido financiado, en parte, por la Comunidad de Madrid y los Fondos Estructurales de la UE a través del proyecto S2018/NMT-4331. Los autores agradecen a Jesús Cerquides y Alejandra López de Aberasturi (IIIA-CSIC) las discusiones sobre este estudio, así como la colaboración del alumnado en las pruebas donde se recogieron los datos.

Referencias

[1] Yoram Bachrach, Tom Minka, John Guiver, y Thore Graepel. How To Grade a Test Without Knowing the Answers — A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. En *Proceedings of the 29th International Conference on Machine Learning*, p. 8, 2012.

[2] Clare Brindley y Susan Scoffield. Peer assessment in undergraduate programmes. *Teaching in Higher Education*, 3(1):79–90, 1998.

[3] Javier Gil Flores y María Teresa Padilla Carmona. La participación del alumnado universitario en la evaluación del aprendizaje. *Educación XX1*, 2009.

[4] Tasos Hovardas, Olia E. Tsivitanidou, y Zacharias C. Zacharia. Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers and Education*, 71:133–152, 2014.

[5] Jeff Howe. The rise of crowdsourcing. *Wired Magazine*, 15(6):1–4, 2006.

[6] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[7] Daphne Koller y Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[8] Igor Labutov y Christoph Studer. JAG: A crowdsourcing framework for joint assessment and peer grading. En *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 1010–1016, 2017.

[9] Yoad Lewenberg, Yoram Bachrach, Ulrich Paquet, y Jeffrey S. Rosenschein. Knowing what to ask: A Bayesian active learning approach to the surveying problem. En *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 1396–1402, 2017.

[10] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, y Daphne Koller. Tuned models of peer assessment in MOOCs. En *Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013*, 2013.

[11] Karthik Raman y Thorsten Joachims. Methods for ordinal peer grading. En *Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1037–1046, 2014.

[12] Daniel Reinholz. The assessment cycle: a model for learning through peer assessment. *Assessment and Evaluation in Higher Education*, 41(2):301–315, 2016.

[13] Philip M. Sadler y Eddie Good. The Impact of Self- and Peer-Grading. *Educational Assessment*, 11(1):1–31, 2006.

[14] Neus Sanmartí. *Avaluar i aprendre: un únic procés*. Octaedro Editorial, 2019.

[15] Bar Shalem, Yoram Bachrach, John Guiver, y Christopher M. Bishop. Students, Teachers, Exams and MOOCs: Predicting and Optimizing Attainment in Web-Based Education Using a Probabilistic Graphical Model. En *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 82–97, 2014.

[16] Stan Development Team. Stan modeling language users guide and reference manual, 2.26. Software, 2021.

[17] Keith Topping. Peer assessment. *Theory into Practice*, 48(1):20–27, 2009.