Calificación de Calificadores en la Evaluación por Pares de Exámenes de Respuesta Abierta

Jorge Díez[1], Oscar Luaces[1], Amparo Alonso-Betanzos[2], Alicia Troncoso[3], and Antonio Bahamonde[1]

[1] Centro de Inteligencia Artificial, Universidad de Oviedo
[2] Dept. de Computación, Universidad de A Coruña
[3] Dept. de Lenguajes y Sistemas Informáticos, Universidad Pablo de Olavide

Resumen La evaluación de cuestiones de respuesta abierta en MOOCS es un reto hoy en día en el marco de Big Data. Recientemente, la evaluación por pares es un método que esta siendo explorado para resolver este problema. Un aspecto clave en el éxito de este método para la evaluación masiva es que los estudiantes que actúan de evaluadores estén motivados para llevar a cabo la tarea. En este trabajo, se propone un método tanto para calcular las notas a partir de las evaluaciones que los estudiantes asignan a cada respuesta como para evaluar a los propios evaluadores. Así, cada estudiante obtiene al final del proceso dos notas, una como autor y una como evaluador, y su nota final es una combinación lineal de ambas. Finalmente, se incluye también un estudio experimental del método a partir de datos reales de tres universidades públicas del Sistema Universitario Español.

Keywords: MOOCs, Evaluación por Pares, Factorización, Aprendizaje de Preferencias, Evaluación de Evaluadores

1. Introducción

El término *MOOC* (acrónimo en inglés de *Massive Open Online Course*), que podemos traducir como curso en línea masivo y abierto, fue acuñado en 2008, cuando en un curso abierto se registraron 2300 estudiantes. Desde entonces, el número de MOOCs ofertados ha continuado creciendo, atrayendo a miles de estudiantes de diferentes procedencias geográficas y sociales e implicando a las mejores universidades del mundo a través de plataformas como Coursera, edX o Udacity. Con los MOOCs asistimos también a la aparición de varios retos. Uno de ellos es la evaluación de exámenes con respuestas abiertas, que en numerosas ocasiones son necesarias para medir con precisión la adquisición de ciertos conocimientos. Obviamente, el método tradicional de corrección por parte de uno o varios instructores no es viable por el elevado número de trabajos a evaluar, y es necesario buscar alguna alternativa.

Una de las soluciones más utilizadas consiste en emplear la llamada *evaluación por pares* (*peer assessment* en inglés). Esto supone que a cada estudiante se

le asigna un pequeño subconjunto de pruebas realizadas por otros compañeros del mismo curso como en [9,2,5,10,6,1,4]. Es importante destacar que para intentar garantizar la calidad de la corrección, así como para intentar dar una mayor uniformidad a las evaluaciones, los criterios que se deben emplear tienen que estar claramente establecidos en un documento de reglas de evaluación detalladas, que en este contexto se llama *rúbrica*. También es conveniente que las respuestas sean anónimas para quien las califica.

La evaluación por pares tiene como ventaja, además de aliviar la carga de los instructores, el que supone por parte de los estudiantes una nueva lectura del material objeto de evaluación. Esta lectura se realiza con ojos críticos, tratando de medir los trabajos de otros compañeros con arreglo a los criterios de la rúbrica. Este proceso hace que los estudiantes-calificadores adquieran un conocimiento más profundo de la materia de la que trate el examen.

Por otra parte, la principal desventaja de la evaluación por pares es que los estudiantes no son evaluadores profesionales y sus calificaciones deben ser filtradas usando algún método de que nos permita *agregar las evaluaciones* recibidas para cada una de las respuestas. Un método estadístico sencillo consistiría en utilizar la media o la mediana de las notas recibidas por cada respuesta como nota final, [2]. Se ha demostrado que la inconsistencia de las calificaciones de los estudiantes se puede paliar con un número elevado de calificaciones por respuesta [5]. Pero solamente es viable disponer de unas pocas calificaciones para cada respuesta; de otra forma sobrecargaríamos en exceso a los estudiantes. Por tanto, es necesario buscar una función de agregación más allá de las medidas de centralidad estadísticas. Esta búsqueda es precisamente una de las aplicaciones más interesantes y complejas de *Big Data* en el campo de la Educación.

En este artículo presentamos un sistema para *calificar a los calificadores* en un procedimiento de evaluación por pares. Esto es importante para motivar a los estudiantes-calificadores a realizar su trabajo correctamente. Por tanto, antes de comenzar todo el proceso se ha de anunciar que la calificación final de un estudiante se calculará como una combinación lineal de sus notas como autor y como calificador de un grupo de respuestas de otros compañeros. El procedimiento de calificación, como veremos, será el resultado de una tarea de aprendizaje automático íntimamente ligada a la de la agregación de las calificaciones que recibe una respuesta.

Como los estudiantes dan calificaciones que no son consistentes, utilizaremos no sus notas sino el orden en que colocan las respuestas que califican. Así, cada par de respuestas con notas diferentes lo vamos a interpretar como una indicación de que una de ellas es *preferible* a otra a juicio del calificador. La agregación de calificaciones será un algoritmo de extensión de los rankings de cada calificador para obtener un ranking de todos los estudiantes con el mayor consenso posible. En otras palabras, la agregación se presentará como una tarea de aprendizaje de preferencias.

El planteamiento de este artículo será el de un sistema de *factorización de matrices*, como los empleados en otras tareas de compleción de tablas de datos como son los *Sistemas de Recomendación*. Usaremos el planteamiento básico de

[1,4] que se trata de un procedimiento *escalable* a grandes volúmenes de datos pues supone resolver un problema de optimización convexa con un algoritmo de descenso de gradiente. Finalmente, la calificación de los calificadores será una medida de la similitud del ranking que construiría un calificador con todos las respuestas y el ranking de consenso de todos los calificadores.

Tras presentar el trabajo relacionado y la descripción formal del método de calificación, en este artículo presentamos los resultados de un experimento real. A los alumnos de la asignatura de *Sistemas Inteligentes* de 3 universidades españolas, A Coruña, Pablo de Olavide en Sevilla y Oviedo en Gijón, se les puso la misma prueba que calificaron tanto los estudiantes como los instructores de las tres universidades. En la sección de resultados experimentales mostramos que las discrepancias entre el ranking del sistema y la instructora de Sevilla (considerada la mejor calificadora por los demás instructores) es menor que la discrepancia entre esta instructora y las calificaciones de los instructores de las otras dos universidades. También mostramos que las calificaciones de los calificadores por el método que aquí se presenta son comparables a los que se obtienen en caso de comparar sus rankings con el de la instructora andaluza.

2. Trabajo Relacionado

Para abordar el reto de la evaluación de respuestas abiertas hay otras alternativas a la evaluación por pares desde el Aprendizaje Automático en Big Data. La más directa consiste en tratar la calificación de una prueba como una tarea de clasificación o regresión (ordinal o métrica). Para esto debemos disponer de una descripción de las respuestas como un vector de características y también de un subconjunto de respuestas evaluadas por el instructor que sean generalizables por un algoritmo de aprendizaje. Para el caso de cursos de programación, en [12] se plantea una descripción de los programas de los estudiantes que está basada en la sintaxis de sus códigos.

En este artículo usamos evaluación por pares. En la introducción se citaron algunas referencias en las que no profundizaremos aquí por falta de espacio. Sin embargo, glosaremos a continuación algunas referencias bibliográficas relevantes sobre la calificación de calificadores que es el foco de este artículo. La evaluación por pares se usa intensamente en los procesos de selección de artículos en revistas o congresos y también en los proyectos de investigación. En cada uno de estos campos hay precedentes de métodos que tratan de calificar a los calificadores.

En [11] se propone un modelo probabilístico para los revisores de artículos que se aprende con un algoritmo EM (*Expectation-Maximization*) que estima la media y la varianza de cierta distribución para cada revisor. Se usa explícitamente la hipótesis de que el sesgo de cada revisor es independiente de los artículos que revisa, lo cual no es siempre exacto pues las revisiones dependerán del conocimiento que los revisores tengan sobre la materia del artículo que evalúan. Aunque el objetivo fundamental de [11] es mejorar la evaluación de los artículos, se establece implícitamente un ranking de los revisores en función de sus sesgos.

Sobre la evaluación de proyectos encontramos el trabajo de Kurokawa y otros [3]. El artículo está motivado por una experiencia piloto que recientemente ha llevado a cabo el NSF (*National Science Foundation*) estadounidense. En ella se pedía a los Investigadores Principales de las propuestas de una convocatoria que evaluasen otras propuestas. Este mecanismo se presta a comportamientos poco éticos que podrían producir un importante sesgo en la selección de propuestas. En [3] sus autores proponen un método imparcial en este contexto. La propuesta se basa en la selección de proyectos específicos para cada evaluador cuyas características entran por tanto en juego de manera crucial.

Dentro del mundo académico, en [13] se propone un método de evaluación por pares que considera la calidad de los calificadores para hacer las asignaciones de nuevos trabajos a evaluar. Cada calificación es calificada por el estudiante que la recibe y esto modifica las calificaciones como calificadores, la cual inicialmente es un valor arbitrario pero alto. La elección de estas calificaciones iniciales es fundamental en todo el proceso, no resultando sencillo encontrar los valores óptimos, lo que es una debilidad de esta propuesta.

Una propuesta más similar a la que aquí presentamos es la de Joachims y otros en [6,7]. En estos artículos se usan distintos modelos probabilísticos que permiten estimar, además del ranking de las respuestas, una calificación de los calificadores. La diferencia con nuestro artículo está en el método de estimación de esta calificación y en que nosotros medimos experimentalmente la precisión de estas calificaciones.

3. Marco Formal y Procedimiento de Aprendizaje

Sea $\mathcal{G}$ un conjunto de *calificadores* y $\mathcal{A}$ un conjunto de respuestas a calificar. A cada calificador se le asignará un subconjunto de respuestas para que las califique. Recogeremos en una *matriz de calificaciones* $\boldsymbol{M}$ los resultados: las filas tendrán índices en $\mathcal{G}$ y columnas en $\mathcal{A}$. La componente $M(g, a) \in [0, 10]$ es la nota que el calificador $\boldsymbol{g} \in \mathcal{G}$ le asigna a la respuesta $\boldsymbol{a} \in \mathcal{A}$. Es importante destacar que $\boldsymbol{M}$ es una matriz muy poco densa pues cada calificador califica solo un número reducido de respuestas. El objetivo es rellenar las componentes de $\boldsymbol{M}$ de la misma forma que se hace en los *Sistemas de Recomendación*. Con la matriz completa tomaremos como nota final de cada respuesta la media de las calificaciones que le hayan correspondido. Al mismo tiempo buscaremos la manera de calificar a los calificadores. La estrategia que usaremos se conoce como *factorización de matrices*.

El primer paso consiste en proyectar tanto los calificadores como las respuestas a un mismo espacio Euclídeo. Para esto usaremos una representación vectorial. Caben distintas alternativas en este punto, por ejemplo usar características de los calificadores y de los exámenes. En este artículo, emplearemos la versión más sencilla, que usa vectores binarios que tienen por conjuntos de índices $\mathcal{G}$ y $\mathcal{A}$, como en las líneas de $\boldsymbol{M}$; cada elemento tendrá asociado un vector con todas las componentes nulas salvo la que corresponde con su índice que valdrá 1. Usaremos el símbolo $\boldsymbol{g}$ (respectivamente $\boldsymbol{a}$) tanto para un calificador

(respectivamente una respuesta) como para su representación vectorial. Con esta codificación, las proyecciones (embebimiento) de las que partimos serán de la forma

$$\mathbb{R}^{|\mathcal{G}|} \to \mathbb{R}^k, \quad \boldsymbol{g} \mapsto \boldsymbol{W}\boldsymbol{g}; \qquad \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}^k, \quad \boldsymbol{a} \mapsto \boldsymbol{V}\boldsymbol{a}. \tag{1}$$

El objetivo es estimar la *calificación individual* del calificador $\boldsymbol{g}$ a la respuesta $\boldsymbol{a}$ como

$$\mathrm{f}(\boldsymbol{g}, \boldsymbol{a}) = \langle \boldsymbol{W}\boldsymbol{g}, \boldsymbol{V}\boldsymbol{a} \rangle . \tag{2}$$

Como habíamos adelantado, la estimación de la *calificación final* de una respuesta $\boldsymbol{a}$ será la media de las estimaciones de las calificaciones individuales,

$$\frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathrm{f}(\boldsymbol{g}, \boldsymbol{a}) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \langle \boldsymbol{W}\boldsymbol{g}, \boldsymbol{V}\boldsymbol{a} \rangle = \left\langle \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \boldsymbol{W}\boldsymbol{g}, \boldsymbol{V}\boldsymbol{a} \right\rangle = \langle \boldsymbol{W}\bar{\boldsymbol{g}}, \boldsymbol{V}\boldsymbol{a} \rangle = \mathrm{f}(\bar{\boldsymbol{g}}, \boldsymbol{a}), \tag{3}$$

donde, evidentemente $\bar{\boldsymbol{g}}$ es el vector media de todos los calificadores,

$$\bar{\boldsymbol{g}} = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{g} \in \mathcal{G}} \boldsymbol{g}.$$

Con estas estimaciones definimos la *calificación del calificador* $\boldsymbol{g}$ como la similitud entre el modelo de sus calificaciones individuales y las finales. Observamos que el orden que induce $\boldsymbol{g}$ (2) en las respuestas solo depende de la orientación de su vector asociado en $\mathbb{R}^k$, es decir, $\boldsymbol{W}\boldsymbol{g}$. Lo mismo sucede con el vector media para las calificaciones finales, $\boldsymbol{W}\bar{\boldsymbol{g}}$. Entonces definimos la calificación de un calificador (CC) como el coseno del ángulo que forman en el espacio Euclídeo común,

$$\mathrm{CC}(\boldsymbol{g}) = \cos(\boldsymbol{W}\boldsymbol{g}, \boldsymbol{W}\bar{\boldsymbol{g}}) = \frac{\langle \boldsymbol{W}\boldsymbol{g}, \boldsymbol{W}\bar{\boldsymbol{g}} \rangle}{\|\boldsymbol{W}\boldsymbol{g}\| \cdot \|\boldsymbol{W}\bar{\boldsymbol{g}}\|}. \tag{4}$$

Por tanto, para estimar la calidad de los calificadores, necesitamos las matrices de proyección (desde luego $\boldsymbol{W}$, pero va ligada a $\boldsymbol{V}$). La propuesta clave de este artículo consiste en obtener las matrices de proyecciones optimizando la función suma de las calificaciones individual (2) y final (3),

$$\mathrm{f}(\bar{\boldsymbol{g}}, \boldsymbol{a}) + \mathrm{f}(\boldsymbol{g}, \boldsymbol{a}) = \langle \boldsymbol{W}\bar{\boldsymbol{g}}, \boldsymbol{V}\boldsymbol{a} \rangle + \langle \boldsymbol{W}\boldsymbol{g}, \boldsymbol{V}\boldsymbol{a} \rangle = \langle \boldsymbol{W}(\bar{\boldsymbol{g}} + \boldsymbol{g}), \boldsymbol{V}\boldsymbol{a} \rangle = \mathrm{f}(\bar{\boldsymbol{g}} + \boldsymbol{g}, \boldsymbol{a}). \tag{5}$$

La optimización de esta función la entendemos en el sentido de que ordene tanto las respuestas como los calificadores en la medida de lo posible. La formalización de esta idea se consigue utilizando el conjunto de *juicios de preferencias* $\mathcal{D}$ definido como los triples formados por un calificador $\boldsymbol{g}$, y un par de respuestas $(\boldsymbol{a}_m, \boldsymbol{a}_p)$ a las que calificó en ese orden,

$$\boldsymbol{M}(\boldsymbol{g}, \boldsymbol{a}_m) > \boldsymbol{M}(\boldsymbol{g}, \boldsymbol{a}_p) \Rightarrow [\boldsymbol{g}, \boldsymbol{a}_m, \boldsymbol{a}_p] \in \mathcal{D}. \tag{6}$$

Nótese que solo se consideran los pares de respuestas con calificaciones diferentes de cada calificador; los empates se descartan al generar el conjunto de juicios de preferencias.

En varias ocasiones vamos a necesitar comparar dos órdenes de un mismo conjunto de objetos. El grado en que coinciden dos órdenes se puede medir como la proporción de veces que coincide el orden de cada par de objetos. Esta proporción es el área bajo la curva ROC (AUC), que también se conoce por otros nombres como *índice de concordancia* (*C-index*) o el índice de τ de *Kendall*.

Por ejemplo, la coincidencia entre el orden dado por una función de ordenación h y el dado por los calificadores en $\mathcal{D}$ se puede calcular como

$$\mathrm{AUC}(\mathrm{f}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{g}, \boldsymbol{a}_m, \boldsymbol{a}_p) \in \mathcal{D}} \mathbb{I}_{h(\boldsymbol{g}, \boldsymbol{a}_m) > h(\boldsymbol{g}, \boldsymbol{a}_p)} + \frac{1}{2} \mathbb{I}_{h(\boldsymbol{g}, \boldsymbol{a}_m) = h(\boldsymbol{g}, \boldsymbol{a}_p)}. \tag{7}$$

Este índice no es una medida simétrica. Por tanto, cuando se comparan dos órdenes, debe considerarse a uno de ellos como el *verdadero*. En la fórmula anterior se considera como verdadero el orden que los calificadores dan a las respuestas.

Teniendo en cuenta lo anterior, buscamos una función de pérdida que sea derivable y convexa y que pueda sustituir al complementario del AUC (que no lo es). Usaremos el planteamiento de *maximizar el margen*, como en las Máquinas de Vectores Soporte (SVM) de clasificación. Así, para cada $(\boldsymbol{g}, \boldsymbol{a}_m, \boldsymbol{a}_p) \in \mathcal{D}$, tratamos de asegurar que la diferencia de valoraciones por f para las respuestas sea de al menos un punto. De esta forma buscamos matrices $\boldsymbol{W}$ y $\boldsymbol{V}$ que minimicen

$$\mathrm{err}(\boldsymbol{W}, \boldsymbol{V}) = \sum_{(\boldsymbol{g}, \boldsymbol{a}_m, \boldsymbol{a}_p) \in \mathcal{D}} \mathrm{máx}(0, 1 - \mathrm{f}(\bar{\boldsymbol{g}} + \boldsymbol{g}, \boldsymbol{a}_m) + \mathrm{f}(\bar{\boldsymbol{g}} + \boldsymbol{g}, \boldsymbol{a}_p)). \tag{8}$$

La solución puede encontrarse usando un algoritmo de *Descenso Estocástico del Gradiente (SGD)* que iterativamente modifica

$$\Theta \leftarrow \Theta - \gamma \cdot \frac{\partial \mathrm{err}(\Theta)}{\partial \Theta}, \tag{9}$$

donde Θ es alternativamente $\boldsymbol{W}$ y $\boldsymbol{V}$, mientras que γ es la *tasa de aprendizaje* que se va modificando en cada iteración.

La derivada parcial con respecto a $\boldsymbol{W}$, cuando $(1 - \mathrm{f}(\bar{\boldsymbol{g}} + \boldsymbol{g}, \boldsymbol{a}_m) + \mathrm{f}(\bar{\boldsymbol{g}} + \boldsymbol{g}, \boldsymbol{a}_p) > 0)$ (en otro caso es 0), es

$$\frac{\partial \mathrm{err}(\boldsymbol{W}, \boldsymbol{V})}{\partial \boldsymbol{W}} = \boldsymbol{V}\boldsymbol{a}_p(\bar{\boldsymbol{g}} + \boldsymbol{g})^{\mathrm{T}} - \boldsymbol{V}\boldsymbol{a}_m(\bar{\boldsymbol{g}} + \boldsymbol{g})^{\mathrm{T}} = \boldsymbol{V}(\boldsymbol{a}_p - \boldsymbol{a}_m)(\bar{\boldsymbol{g}} + \boldsymbol{g})^{\mathrm{T}}. \tag{10}$$

Una ecuación análoga se tiene para la derivada parcial con respecto a $\boldsymbol{V}$.

4. Evaluación Experimental

En esta sección mostramos los resultados obtenidos en un experimento piloto diseñado para estimar la precisión de los algoritmos descritos en los apartados

anteriores. En el experimento participaron estudiantes de 3 universidades españolas: A Coruña (UDC), Pablo de Olavide en Sevilla (UPO) y Oviedo (Uniovi) en Gijón; los acrónimos que aparecen tras el nombre de cada universidad serán utilizados en la Figura 1 de esta sección. A todos los estudiantes matriculados en la asignatura *Sistemas Inteligentes* en estas tres universidades se les pedía que formalizasen tres problemas de búsqueda clásicos tomados del libro [8]. Es decir, estamos ante una pregunta con respuesta abierta.

Cada alumno tuvo que subir su trabajo a un evento registrado en EasyChair con el nombre *JRLO2014 (Joint Research in Learning to Order 2014)*. Los trabajos eran documentos de texto anónimos. Utilizamos para la experimentación los trabajos entregados por 175 alumnos de las tres universidades, 160 de los cuales también actuaron como calificadores. Cada calificador recibió una media de 8.29 respuestas para calificar repartidos de forma aleatoria, cuidando que no se calificase la propia respuesta. Para tratar de hacer uniformes las calificaciones se les proporcionó a los estudiantes-calificadores una *rúbrica* detallada donde se especificaban las reglas a aplicar en la calificación. Cada respuesta fue calificada una media de 7.58 veces. Las notas eran enteros entre 0 y 10 y registramos 1326; es decir, disponíamos únicamente de un 4.74 % de las 28000 (160×175) calificaciones posibles.

Por otra parte, en el SGD empleado en el experimento para optimizar la (8) se utilizó una tasa de aprendizaje γ (9) que se actualiza en función de la iteración i mediante la expresión $\gamma \leftarrow \dfrac{1}{(\gamma_s \cdot i) + 1}$. Los valores de los parámetros finalmente utilizados fueron los resultantes de una búsqueda (separando ejemplos para test) en los siguientes rangos de valores: $k \in \{2, 50, 100\}, \gamma_s \in \{10^e : e = -7, \ldots, 0\}$. Recuérdese que k es la dimensión del espacio Euclídeo común donde se proyectan tanto los calificadores como las respuestas (1).

Para poder establecer una comparación, los instructores que imparten los cursos de Sistemas Inteligentes de las universidades participantes calificaron las respuestas de *todos* los estudiantes.

4.1. Resultados del Experimento

El rendimiento de los modelos aprendidos fue evaluado en términos de AUC (7). En todos los casos consideramos como *ranking de referencia* el suministrado por la profesora de la UPO, que fue la que mejor se ajustó a la rúbrica suministrada a los estudiantes a juicio de los otros instructores.

Para poder apreciar los resultados obtenidos por el método propuesto en este artículo, hemos añadido los resultados que se obtendrían con un método simple y que, en general es inadecuado: el ranking que se obtiene con la *media* de las notas recibidas por cada respuesta. También analizamos la sensibilidad del método propuesto en relación con el número de evaluaciones efectuadas por cada evaluador. Para ello, construimos nuevos conjuntos muestreando a partir del conjunto original, simulando que a los calificadores se les pidió que evaluaran un máximo de n respuestas ($n = 5, \ldots, 11$).

Los resultados se muestran en la Figura 1. El AUC respecto la profesora de la UPO (referencia) varía entre 0.622 y 0.677 para la media y entre 0.683 y 0.715 para las valoraciones de (3). En la figura se aprecia que cuando se tiene un número reducido de evaluaciones por respuesta el rendimiento de la media y de nuestro sistema empeora. En cualquier caso, los modelos aprendidos por nuestra propuesta tienen un rendimiento siempre mejor que el de la media y, sorprendentemente, mejor que el de los instructores de Gijón y A Coruña.

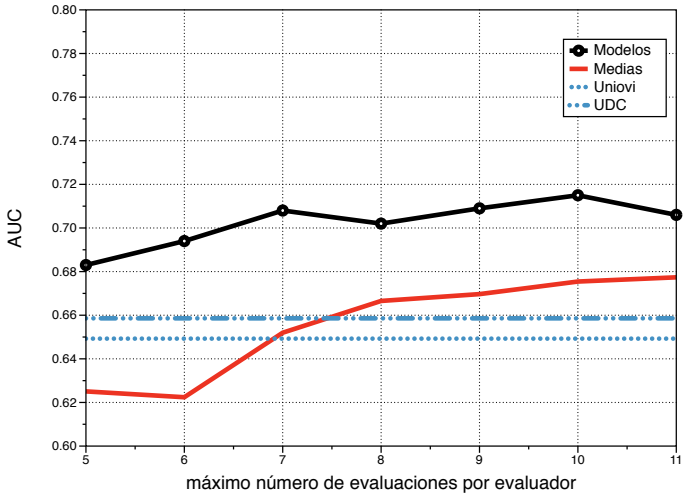

Figura 1. AUCs tomando como referencia a la instructora de la UPO. En el eje horizontal se representa el número máximo de respuestas evaluadas por cada estudiante-calificador. Estas calificaciones se muestrearon aleatoriamente dentro del conjunto de calificaciones reales que los estudiantes hicieron

Finalmente, evaluamos el rendimiento de la medida CC definida en (4) para calificar a los calificadores. Esencialmente, debemos comprobar cuánto se parece el ranking de calificadores dado por esta medida de evaluación, R_{CC}, frente al orden que estableceríamos si pudiésemos comparar cada calificador con un calificador de referencia, del que no dispondremos en un sistema en producción. En nuestro experimento la profesora de la UPO actúa como referencia, así que ordenaremos los calificadores atendiendo al AUC entre las calificaciones otorgadas por ese calificador según el modelo, $\mathtt{f}(\boldsymbol{g}, \cdot)$, y las otorgadas por la profesora. Si denominamos R_{ref} al ranking de calificadores así obtenido entonces podemos medir la calidad de R_{CC} como $\mathrm{AUC}(R_{\mathrm{CC}}, R_{\mathrm{ref}})$.

La Tabla 1 muestra la calidad del ranking R_{CC} (en términos de AUC respecto a R_{ref}). Podemos apreciar que hay una gran coincidencia, con valores de AUC por encima de 0.70, en ambos órdenes cuando el número de respuestas calificadas por estudiante es mayor que 7. Obviamente, cuanto menor sea el número de evaluaciones efectuadas por un calificador más difícil será determinar su calidad.

Tabla 1. Calificación de calificadores. La primera columna indica el máximo número de evaluaciones por evaluador, la segunda es el AUC del ranking de calificadores dado por CC (4) y el ranking obtenido comparando las calificaciones de los calificadores con la profesora de la UPO. Las tres últimas columnas muestran el porcentaje de coincidencia de calificadores en ambos rankings en cada tercil

Max. eval.	AUC($R_{\mathrm{CC}}, R_{\mathrm{ref}}$)	% de coincidencias en T1	T2	T3
5	0.617	47.6	45.8	49.6
6	0.649	45.3	43.6	55.5
7	0.687	52.4	50.4	58.0
8	0.730	54.9	50.9	69.8
9	0.728	52.7	48.9	68.4
10	0.742	57.5	53.3	69.6
11	0.808	65.5	60.7	80.2

Para dar otra perspectiva de los rankings de calificadores, hemos calculado el porcentaje de calificadores que coinciden en el mismo tercil en los rankings R_{CC} y R_{ref}. Los resultados están en las 3 últimas columnas de la Tabla 1.

5. Conclusiones

En este trabajo hemos presentado un mecanismo de evaluación de respuestas abiertas mediante la aplicación de técnicas de factorización de matrices a los datos obtenidos a partir de las calificaciones otorgadas en un proceso de evaluación por pares. Nuestro método aprende una función de calificación a partir de juicios de preferencias, órdenes parciales de ejercicios según cada calificador. En los juicios de preferencias se descarta el valor absoluto de la calificación para anular el efecto subjetivo del proceso. La función de calificación se obtiene mediante la optimización de una aproximación convexa del AUC que utiliza el descenso de gradiente estocástico (SGD). Por tanto, el algoritmo que presentamos es perfectamente escalable a grandes volúmenes de datos, como los que se generan en los MOOCs, donde el número de alumnos puede ascender a decenas de miles.

Los resultados obtenidos sobre datos recogidos en un experimento de evaluación por pares realizado en tres universidades españolas muestran que el método ordena los ejercicios evaluados mejor que un procedimiento que calcula la media de las calificaciones dadas por los calificadores. Incluso vemos que los resultados son comparables a los de los instructores profesionales.

Por todo lo anterior, podemos concluir que es posible desarrollar métodos escalables para producir un ranking fiable de respuestas abiertas a partir de un reducido número de calificaciones dadas por evaluadores no expertos. Nuestro método permite, además, calificar a los calificadores en función de la calidad de sus evaluaciones. Esto es importante puesto que los alumnos, que participan en el proceso con el doble rol de estudiante y evaluador, pueden ser calificados por ambos aspectos. Este hecho resulta ser un buen incentivo para que los estudiantes

realicen su labor de evaluación con interés, puesto que su nota final dependerá en parte de su calidad como evaluador.

Agradecimientos La investigación que se presenta en este artículo ha sido subvencionada en parte por los proyectos TIN2011-23558 y TIN2012-37954 del Ministerio de Economía y Competitividad (MINECO) y el proyecto P12-TIC-1728 de la Junta de Andalucía; todos estos proyectos han sido parcialmente subvencionados por fondos FEDER. Además queremos agradecer a nuestros estudiantes su colaboración calificando las respuestas de sus colegas de las universidades A Coruña, Pablo de Olavide y Oviedo.

Referencias

1. Díez, J., Luaces, O., Alonso-Betanzos, A., Troncoso, A., Bahamonde, A.: Peer Assessment in MOOCs Using Preference Learning via Matrix Factorization. In: NIPS Workshop on Data Driven Education (2013)
2. Kulkarni, C., Pang-Wei, K., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., Klemmer, S.R.: Peer and self assessment in massive online classes. Tech. rep., Stanford University (2013)
3. Kurokawa, D., Lev, O., Morgenstern, J., Procaccia, A.: Impartial peer review. In: IJCAI (2015)
4. Luaces, O., Díez, J., Alonso-Betanzos, A., Troncoso, A., Bahamonde, A.: A factorization approach to evaluate open-response assignments in moocs using preference learning on peer assessments. Knowledge-Based Systems (2015)
5. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D.: Tuned models of peer assessment in MOOCs. In: Proceedings of the 6^{th} International Conference on Educational Data Mining (EDM'13). pp. 153–160. International Educational Data Mining Society (2013)
6. Raman, K., Joachims, T.: Methods for ordinal peer grading. In: ACM Conference on Knowledge Discovery and Data Mining (KDD) (2014)
7. Raman, K., Joachims, T.: Bayesian ordinal peer grading. In: Proceedings of the Second (2015) ACM Conference on Learning @ Scale. pp. 149–156. L@S '15, ACM, New York, NY, USA (2015)
8. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson Education (2010)
9. Sadler, P.M., Good, E.: The impact of self-and peer-grading on student learning. Educational assessment 11(1), 1–31 (2006)
10. Shah, N.B., Bradley, J.K., Parekh, A., Wainwright, M., Ramchandran, K.: A case for ordinal peer-evaluation in MOOCs. In: NIPS Workshop on Data Driven Education (2013)
11. Spalvieri, A., Mandelli, S., Magarini, M., Bianchi, G.: Weighting peer reviewers. In: Privacy, Security and Trust (PST), 2014 Twelfth Annual International Conference on. pp. 414–419. IEEE (2014)
12. Srikant, S., Aggarwal, V.: A system to grade computer programming skills using machine learning. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1887–1896. ACM (2014)
13. Xiao, Y., Dörfler, F., van der Schaar, M.: Incentive design in peer review: Rating and repeated endogenous matching. arXiv:1411.2139 (2014)