Realimentación automática en evaluación por pares de respuestas abiertas mediante factorización de matrices

Jorge Díez, Oscar Luaces, and Antonio Bahamonde

Centro de Inteligencia Artificial
Universidad de Oviedo en Gijón
33204 – España

Resumen En este artículo presentamos un método para dar una realimentación para alumnos tras haber realizado un examen de respuesta abierta, es decir, que los alumnos pueden escribir un texto libremente que debe ser leído y comprendido para que ser evaluado. Supondremos además que esta evaluación ha sido realizada por otros compañeros; es decir, sometida a una evaluación por pares. La aplicación de este método incluye los MOOCs (cursos masivos abiertos online), pero también las colecciones de pruebas de evaluación continua que no se puedan evaluar de manera automática. En la evaluación por pares, cada alumno, tras hacer su examen, recibe varios exámenes que debe corregir siguiendo una guía que suministra el instructor (llamada rúbrica). En la evaluación, además de una nota global, en el método que aquí se presenta, se pedirá a los alumnos evaluadores que gradúen unas etiquetas que explicarían la nota global. Como cada examen recibe varias calificaciones es necesario unificarlas. El método propuesto utiliza para esta fase un algoritmo de aprendizaje automático multitarea. El artículo presenta los resultados obtenidos por el método en 3 conjuntos de datos originales obtenidos en cursos de la Universidad asturiana.

Keywords: peer assessment, feedback, preference learning, matrix factorization, multilabel classification

1 Introducción

La evaluación de exámenes es un problema cuando se trata de cursos que potencialmente puedan tener un elevado número de alumnos, como es el caso de los MOOCs o cuando se hagan muchas pruebas escritas que requieran una evaluación no automática. El recurso habitualmente empleado para abordar estas situaciones es utilizar la llamada *evaluación por pares*: son los propios alumnos quienes corregirán las respuestas de otros compañeros. Cada alumno corrige un grupo de respuestas y, así, cada respuesta recibirá varias evaluaciones. Naturalmente, esta solución necesita filtrar las notas que ponen los alumnos, pues pueden producirse unos sesgos importantes.

Hay una amplia colección de publicaciones que abordan este tema [5, 9–13]. Sin embargo, en estas publicaciones solamente se aborda la cuestión de la nota global que el alumno debe recibir por su respuesta. En este artículo vamos un paso más allá y proponemos un mecanismo eficiente para dar a los alumnos una realimentación (*feedback*) más amplia que la de un simple número.

Desde luego, tras someterse a una evaluación, la realimentación que el estudiante recibe es un elemento clave en el proceso de aprendizaje [3, 6, 14]. La nota que el evaluador asigna a la respuesta de un alumno es ya una realimentación: resume el nivel de competencia o destreza adquirido por el estudiante.

Sin embargo, en la mayoría de los casos, la nota global no es suficientemente informativa. Salvo cuando la puntuación tome un valor extremo, es decir, aquellos ejercicios que se consideren prácticamente perfectos o completamente incorrectos o en blanco, parece por tanto razonable acompañar la calificación con indicaciones adicionales que expliquen al estudiante los errores cometidos, o en qué aspectos debe incidir para mejorar su nivel de conocimientos. Estas indicaciones adicionales son las que tradicionalmente reciben los estudiantes cuando acuden a revisar sus exámenes buscando una justificación a su puntuación.

Para intentar generar de forma automática esta información adicional en un contexto de evaluación por pares se plantea el mismo problema que para asignar una puntuación: podemos tener una realimentación diferente por cada evaluador que corrija un mismo ejercicio. En este artículo proponemos extender la solución presentada en trabajos previos [2, 7] con el objetivo de obtener un modelo de consenso de los comentarios de realimentación utilizando un algoritmo de aprendizaje con unas peculiaridades específicas para esta tarea.

El resto del artículo está organizado así: en la Sección 2 se da una descripción general del método propuesto; la Sección 3 presenta las ecuaciones que gobiernan el método de aprendizaje propuesto; y finalmente mostramos los resultados obtenidos con 3 conjuntos de datos que proceden de exámenes realizados en la Universidad de Oviedo.

2 Descripción del método de filtrado de las calificaciones

Los alumnos al terminar de escribir sus respuestas actuarán como evaluadores; para ello recibirán lo siguiente:

- Un pequeño grupo de respuestas a evaluar.
- Una guía en la que se indican los criterios de valoración y quizás las lineas generales de las *respuestas correctas*; a este documento se le llama *rúbrica* y por supuesto será suministrado por el instructor.
- Una *plantilla de evaluación* con frases o *etiquetas* proporcionadas por el instructor. Estas etiquetas servirán para reflejar ciertos aspectos de la evaluación. De esta manera, el evaluador podrá indicar, mediante un grado ordinal, hasta qué punto cada una de esas etiquetas es aplicable al ejercicio que está corrigiendo. La calificación global del ejercicio es un caso particular de etiqueta de realimentación. La Figura 1 muestra un ejemplo de plantilla de evaluación.

Figura 1. Plantilla de corrección con frases de realimentación usada en el ejercicio de *Información Contable para el Comercio*

Tras las evaluaciones, cada respuesta habrá recibido calificaciones de distintos alumnos. En realidad, cada ejercicio recibirá varios vectores de calificaciones. Estos vectores tendrán una componente por cada etiqueta (incluyendo la nota global) y el valor será la nota o grado obtenido.

Consideramos de importancia capital el hecho de que existe una relación entre las etiquetas que se usan para evaluar las respuestas. Es razonable pensar, por ejemplo, que un ejercicio muy mal redactado tenga más faltas de ortografía que otro con muy buena redacción. También resulta clave el hecho de que cada respuesta solo recibe valoraciones de algunos alumnos, no de todos. Tendremos una *matriz de valoraciones* posibles, como sucede en los *sistemas de recomendación*, que en general contendrá valores solo para un pequeño porcentaje de componentes. Se tratará de matrices muy poco densas.

El planeamiento que proponemos en este artículo consiste en realizar los siguientes pasos:

1. Completaremos la matriz de valoraciones con valores *coherentes* con los que tenemos disponibles. Como cada valoración es un grado, calcularemos un real que nos permita ordenar las respuestas.
2. Una vez completada la matriz con los valores dados por el modelo podemos calcular la media de los valores para cada etiqueta en cada ejercicio. Este paso dará un ranking de respuestas para cada una de las etiquetas a evaluar.
3. Transformaremos los rankings en puntuaciones de cada etiqueta de tal forma que las notas sigan la misma distribución que las otorgadas por los alumnos.

La compleción de la matriz la realizará una función que debemos aprender a partir de la definición parcial que constituye la matriz de evaluación. Esta función deberá dar las notas o grados para cada respuesta y evaluador, es decir, un vector de tantas componentes como etiquetas tenga el examen que se está evaluando. Plantearemos ese paso clave como una multitarea de aprendizaje automático. Trataremos entonces de aprender un modelo que, de forma simultánea, sea capaz de hacer predicciones para todas las etiquetas (incluyendo la nota global).

La coherencia entre la matriz de partida y matriz completada de esta manera la evaluaremos en términos de orden relativo. Es decir, pretendemos que la matriz completa considere mejores las respuestas que para la mayoría de los alumnos (al actuar como evaluadores) lo eran.

Este aspecto es fundamental. En términos prácticos estamos diciendo que (en esta fase) de las evaluaciones no nos importa su valor numérico sino su valor relativo. Si un evaluador pone un 9 a una respuesta y un 4 a otra, no usaremos estos valores salvo para decir que la primera respuesta le parece *mejor que* la segunda. Este planteamiento es el del llamado punto de vista *ordinal*. La alternativa sería considerar las notas como un objetivo deseable para aprender directamente de ellas (en términos de aprendizaje estaríamos frente a una regresión), sería el punto de vista *cardinal*.

Las razones para optar por el planteamiento ordinal han sido estudiadas en el aprendizaje de preferencias en multitud de ocasiones, como por ejemplo en [1, 4, 8].

Sin embargo sí que tendremos en cuenta las notas que los alumnos han otorgado a las respuestas que han evaluado. En el tercer paso (ver la descripción anterior de los pasos a dar), la distribución de las notas dadas por los alumnos en cada etiqueta será la que se utilice para transformar el ranking del segundo paso en las calificaciones que otorgue el sistema.

3 Formalización del método

Sea $\mathcal{G}$ un conjunto de evaluadores, $\mathcal{A}$ un conjunto de ejercicios a evaluar y $\mathcal{L}$ el conjunto de aspectos evaluables (etiquetas). Tras la evaluación de los alumnos tendremos una matriz de valoraciones, $\boldsymbol{M}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{a})$, de triple entrada, que contiene la valoración dada por el evaluador $\boldsymbol{g} \in \mathcal{G}$ para el ejercicio $\boldsymbol{a} \in \mathcal{A}$ en lo que se refiere al aspecto o etiqueta $\boldsymbol{l} \in \mathcal{L}$ (incluida la nota global).

Como habíamos comentado en la sección anterior, esta matriz tendrá una dispersión muy alta, puesto que cada evaluador, $\boldsymbol{g}$, sólo va a evaluar un pequeño subconjunto $\mathcal{A}_{\boldsymbol{g}} \in \mathcal{A}$ de ejercicios.

Para completar la matriz generalizamos, usando aprendizaje de preferencias, los valores disponibles en $\boldsymbol{M}$. En concreto partimos de un conjunto de entrenamiento

$$\mathcal{D} = \{(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{\beta}, \boldsymbol{\omega}) \;/\; \boldsymbol{M}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{\beta}) > \boldsymbol{M}(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{\omega})\}, \tag{1}$$

donde $\boldsymbol{g} \in \mathcal{G}$, $\boldsymbol{l} \in \mathcal{L}$ y $\boldsymbol{\beta}, \boldsymbol{\omega} \in \mathcal{A}_{\boldsymbol{g}}$, indican que, según el criterio del evaluador $\boldsymbol{g}$, para la etiqueta $\boldsymbol{l}$, el ejercicio $\boldsymbol{\beta}$ merece mayor valoración que el ejercicio $\boldsymbol{\omega}$. De esta manera, en este conjunto de *juicios de preferencia* queda registrado el orden relativo de los ejercicios respecto a cada etiqueta, pero eliminando las valoraciones numéricas dadas por los evaluadores para ser fieles al planteamiento ordinal frente al cardinal.

El mecanismo de aprendizaje parte de una representación vectorial de los ejemplos de entrada. Así, para los ejercicios utilizaremos una representación basada en el contenido conocida como *bolsa de palabras (bag of words)*. Esta representación requiere que obtengamos previamente un *corpus* con todas las

palabras que se hayan usado en el conjunto $\mathcal{A}$ de ejercicios. Cada ejercicio se representará por un vector en el que cada componente se corresponde con un término del corpus: su valor será 1 sólo si el término aparece en el ejercicio, y 0 en caso contrario.

Los identificadores de los evaluadores y de las etiquetas se representarán también de forma vectorial. el objeto i-ésimo (evaluador o etiqueta) será un vector donde la única componente distinta de cero será la i-ésima. Además, puesto que queremos aprender a evaluar simultáneamente distintos aspectos de cada ejercicio, es necesario distinguir para cada ejemplo cuál es el aspecto que está valorando el evaluador. Para ello construimos una representación mediante la concatenación vectorial (suma directa) $\boldsymbol{h} = (\boldsymbol{g} \oplus \boldsymbol{l})$, de manera que $\boldsymbol{h}$ representa al evaluador $\boldsymbol{g}$ cuando evalúa el aspecto $\boldsymbol{l}$.

Los elementos que conforman los ejemplos de entrenamiento son proyectados desde sus respectivos espacios de entrada a un espacio Euclídeo común $\mathbb{R}^k$,

$$\mathbb{R}^{|\mathcal{G}|+|\mathcal{L}|} \mapsto \mathbb{R}^k, \qquad \boldsymbol{h} \mapsto \boldsymbol{W}\boldsymbol{h} = \boldsymbol{W}(\boldsymbol{g} \oplus \boldsymbol{l}), \tag{2}$$

$$\mathbb{R}^{|\operatorname{corpus}(\mathcal{A})|} \mapsto \mathbb{R}^k, \qquad \boldsymbol{a} \mapsto \boldsymbol{V}\boldsymbol{a}. \tag{3}$$

Nótese que la dimensión del espacio de entrada de los ejercicios depende del tamaño del corpus de palabras extraído de $\mathcal{A}$, mientras que la dimensión del espacio de entrada de evaluadores concatenados con etiquetas será la suma del número de evaluadores más el número de aspectos evaluados.

Para estimar la valoración para la etiqueta $\boldsymbol{l}$ del ejercicio $\boldsymbol{a}$ según el criterio del evaluador $\boldsymbol{g}$, definimos ahora una *función de utilidad* como el producto escalar entre las imágenes de $\boldsymbol{h}$ y $\boldsymbol{a}$:

$$f(\boldsymbol{h}, \boldsymbol{a}) = \langle \boldsymbol{W}\boldsymbol{h}, \boldsymbol{V}\boldsymbol{a} \rangle = \langle \boldsymbol{W}(\boldsymbol{g} \oplus \boldsymbol{l}), \boldsymbol{V}\boldsymbol{a} \rangle = \boldsymbol{h}^T \boldsymbol{W}^T \boldsymbol{V}\boldsymbol{a}. \tag{4}$$

En esta ecuación, las matrices $\boldsymbol{W}^T$ y $\boldsymbol{V}$ son *factores* de la matriz de pesos de los productos de las componentes de $\boldsymbol{h}$ y de $\boldsymbol{a}$. Por esta razón, la técnica de aprendizaje que empleamos se llama de *factorización de matrices*.

Con esta función podemos completar la matriz de valoraciones $\boldsymbol{M}$ y calcular finalmente una única valoración para cada ejercicio en cada aspecto, como la media de las valoraciones que se estima que darían todos los evaluadores para ese aspecto si hubiesen evaluado ese ejercicio. Es decir con

$$f((\bar{\boldsymbol{g}} \oplus \boldsymbol{l}), \boldsymbol{a}) = \langle \boldsymbol{W}(\bar{\boldsymbol{g}} \oplus \boldsymbol{l}), \boldsymbol{V}\boldsymbol{a} \rangle \tag{5}$$

donde $\bar{\boldsymbol{g}}$ representa al *evaluador medio*, $\bar{\boldsymbol{g}} = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{g} \in \mathcal{G}} \boldsymbol{g}$.

Para resolver la tarea de clasificación multietiqueta deberemos aprender los parámetros $\boldsymbol{W}$ y $\boldsymbol{V}$ para que las valoraciones del calificador medio sean coherentes con las de los evaluadores. Pero como estas matrices también permiten hacer estimaciones de cada valoración de cada evaluador, buscaremos que también estas estimaciones sean coherentes. Formalmente, pediremos que se minimice la *función de perdida*

$$\operatorname{err}(\boldsymbol{W}, \boldsymbol{V}) = \sum_{(\boldsymbol{g}, \boldsymbol{l}, \boldsymbol{\beta}, \boldsymbol{\omega}) \in \mathcal{D}} \max\{0, 1 - f((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l}, \boldsymbol{\beta}) + f((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l}, \boldsymbol{\omega})\}. \tag{6}$$

Para minimizar la función de pérdida utilizaremos un Descenso Estocástico de Gradiente (*SGD*), que en cada iteración modifica los parámetros del modelo, Θ, que en nuestro caso son las matrices $\boldsymbol{W}$ y $\boldsymbol{V}$, de la siguiente manera:

$$\Theta \leftarrow \Theta - \gamma \left(\frac{\partial \operatorname{err}(\Theta)}{\partial \Theta} + \nu \cdot \frac{\partial \|\Theta\|_F^2}{\partial \Theta} \right) \tag{7}$$

donde $\|\cdot\|_F^2$ es la norma Frobenius que se incluye a modo de *regularización*, γ es la *tasa de aprendizaje* y ν es el *factor de regularización*. Es habitual que γ varíe durante la ejecución del SGD, dependiendo su valor del número de iteraciones realizadas. En los experimentos que se describen al final de este artículo, para determinar el valor de γ en la iteración i-ésima hemos utilizado la expresión

$$\gamma = \frac{1}{1 + \gamma_s \cdot i}. \tag{8}$$

Las derivadas parciales utilizadas en el SGD quedan como sigue:

$$\frac{\partial \operatorname{err}(\Theta)}{\partial \boldsymbol{W}} = \boldsymbol{V}(\boldsymbol{\omega} - \boldsymbol{\beta})((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l})^{\mathrm{T}} \tag{9}$$

$$\frac{\partial \operatorname{err}(\Theta)}{\partial \boldsymbol{V}} = \boldsymbol{W}((\bar{\boldsymbol{g}} + \boldsymbol{g}) \oplus \boldsymbol{l})(\boldsymbol{\omega} - \boldsymbol{\beta})^{\mathrm{T}} \tag{10}$$

Con las expresiones obtenidas en (9) y (10) junto con las derivadas parciales de la regularización, que en este caso son respectivamente $2\boldsymbol{W}$ y $2\boldsymbol{V}$, calculamos la modificación que hay que aplicar a las matrices $\boldsymbol{W}$ y $\boldsymbol{V}$ en cada iteración del algoritmo SGD. El proceso iterativo de aprendizaje se detiene cuando se ha alcanzado un número de iteraciones máximo.

4 Experimentación

Hemos llevado a cabo tres experimentos con datos reales, recabados en tres exámenes de la Universidad de Oviedo en las siguientes asignaturas:

- *Información Contable para el Comercio*, en la Facultad de Comercio, Turismo y Ciencias Sociales 'Jovellanos' de Gijón.
- *Derecho Constitucional*, en la Facultad de Derecho de Oviedo.
- *Economía Española*, en la Facultad de Economía y Empresa de Oviedo.

Para realizar estos tres ejercicios hemos utilizado una instalación *ad hoc* de la plataforma de aprendizaje Moodle (moodle.org). Esta plataforma dispone de una herramienta denominada *Taller* (*workshop*), diseñada especialmente para realizar evaluación por pares, que facilita la distribución de ejercicios de forma anónima entre los participantes, la definición de criterios de evaluación múltiples y la recogida en su base de datos de las calificaciones asignadas por los evaluadores para cada criterio de evaluación. Nosotros suplantamos el mecanismo implementado en Moodle, basado en el cálculo de medias ponderadas, por

Tabla 1. Características principales de los conjuntos de datos utilizados en el experimento de evaluación por pares. La dispersión indica el porcentaje vacío de la matriz de valoraciones $\boldsymbol{M}$.

	Información Contable	Derecho Constitucional	Economía Española
Núm. ejercicios	119	66	111
Núm. evaluadores	112	66	108
Núm. evaluaciones	1120	660	1065
Dispersión (%)	92.09	84.85	91.36
Núm. evaluaciones por ejercicio, media	9.41 ± 0.71	10 ± 0	9.59 ± 0.67
Núm. evaluaciones por evaluador, media	10 ± 0	10 ± 0	9.86 ± 0.99

nuestro método basado en factorización de matrices para obtener los resultados que se discuten en esta sección.

Finalmente, es importante resaltar que en todos los casos las evaluaciones se efectuaron de forma anónima tanto para el evaluador como para los evaluados, garantizando además que en ningún caso un estudiante evaluase su propio ejercicio.

La Tabla 1 muestra las características de los tres conjuntos de datos. Todos los experimentos se establecieron de forma que prácticamente todos los evaluadores evaluaron 10 ejercicios (en el caso de Derecho Constitucional esto se cumplió estrictamente), lo que produjo unas matrices de valoración muy dispersas: en torno al 90% de las matrices están vacías. También cabe destacar que algunos estudiantes no han participado como evaluadores en los ejercicios de Información Contable y en Economía Española. Esto no presenta ningún problema a nuestro método.

Por otra parte, en la Figura 2 se muestran las distribuciones de los valores otorgados por los evaluadores para las etiquetas en los tres ejercicios, siendo los gráficos de la última columna los correspondientes a las calificaciones globales.

4.1 Resultados

Como ya se explicó en la Sección 2, nuestro objetivo es obtener un ranking de ejercicios que se ajuste lo más posible a los rankings dados por los evaluadores, por lo que mediremos el error de nuestros modelos en términos de número de pares ordenados de forma diferente a la ordenación dada por los evaluadores.

El proceso de entrenamiento de nuestro método depende de algunos parámetros, k, ν y γ_s (8), cuyos valores hemos seleccionado mediante una validación cruzada de 2 particiones y 5 repeticiones explorando todas las combinaciones posibles con los valores siguientes: $k \in \{2, 10, 20, 50, 100\}, \nu \in \{10^e : e = -3, \ldots + 1\}, \gamma_s = \{10^e : e = -3, \ldots, -1\}$.

Una vez seleccionados los valores más adecuados y dado que obtenemos los modelos con un proceso estocástico, hemos calculado 10 modelos diferentes por

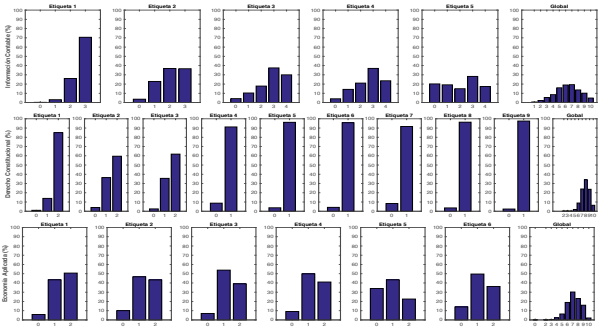

Figura 2. Distribuciones de los valores otorgados por los evaluadores para todas las etiquetas

cada conjunto de entrenamiento. Los resultados de la Tabla 2 muestran los valores medios obtenidos por los 10 modelos de cada experimento. La información se ha organizado de manera que en cada fila se muestran los resultados obtenidos para cada etiqueta de realimentación y en las columnas se muestra lo siguiente:

- *# JP*: número de juicios de preferencia (ejemplos de entrenamiento) del conjunto para la etiqueta correspondiente. Debemos tener en cuenta que los juicios de preferencia asociados a una etiqueta se generan por cada par de ejercicios evaluados que reciben un valor distinto y que permiten, por tanto, establecer un orden relativo entre ellos. Así pues, dos ejercicios que reciban idéntica valoración no dan lugar a ningún juicio de preferencia en esa etiqueta. Por esta razón el número de juicios de preferencia por etiqueta puede ser distinto dentro de un mismo conjunto de entrenamiento, como puede apreciarse en las tablas.
- *Discrepancias*: Porcentaje de pares de ejercicios cuya evaluación produce una ordenación contradictoria a la mayoritaria. Así, si para la etiqueta $\boldsymbol{l}$ hay 3 evaluadores que opinan que el ejercicio x es mejor que el y, pero otros 2 evaluadores opinan lo contrario se contabilizan 2 discrepancias. Por tanto, es imposible que un modelo pueda cometer menos errores que el número de discrepancias, es una cota inferior del error.
- *Error*: Porcentaje de pares de ejercicios que el modelo ordena de forma contraria a los evaluadores. Este valor es la discrepancia entre el modelo aprendido y los datos de entrenamiento. Se muestra el valor medio de 10 ejecuciones del algoritmo $\pm$ la desviación estándar.
- *Etiqueta*: Indica el aspecto evaluado. Entre paréntesis se muestra un número que concuerda con el utilizado en la Figura 2 para poder identificar cada uno de los gráficos con su etiqueta.

Tabla 2. Resultados

Información Contable para el Comercio

#JP	Disc. (%)	Error (%)	Etiqueta
4233	5.20	17.94 ± 0.3	Calificación global
1603	3.74	22.90 ± 0.5	(1) El trabajo contiene faltas de ortografía
3068	5.05	22.10 ± 0.4	(2) Calidad de la redacción del trabajo
3043	4.24	17.95 ± 0.3	(3) El análisis financiero *a corto plazo* es
3187	4.61	18.72 ± 0.2	(4) El análisis financiero *a largo plazo* es
3455	4.08	16.38 ± 0.3	(5) El análisis económico es

Derecho Constitucional

#JP	Disc. (%)	Error (%)	Etiqueta
2158	9.73	23.33 ± 0.2	Calificación global
570	2.98	21.21 ± 1.2	(1) El trabajo contiene faltas de ortografía
1273	6.44	19.63 ± 0.3	(2) Calidad de la redacción del trabajo
1172	4.95	18.93 ± 0.2	(3) Argumentación utilizada en el trabajo
378	2.91	21.61 ± 1.4	(4) El trabajo no cita los artículos aplicables
171	0.00	16.49 ± 2.7	(5) No sabe lo que es la moción de censura
218	2.29	21.83 ± 1.4	(6) No sabe lo que es la cuestión de confianza
369	1.90	15.12 ± 1.3	(7) No sabe las competencias del Rey
184	0.00	16.74 ± 2.3	(8) No sabe cómo se nombra al Presidente
112	2.68	26.96 ± 3.8	(9) No sabe las competencias del Presidente

Economía Española

#JP	Disc. (%)	Error (%)	Etiqueta
3736	8.00	27.26 ± 0.4	Calificación global
2318	5.95	27.07 ± 0.4	(1) Capacidad para comprender y exponer los…
2331	5.19	27.73 ± 0.6	(2) Capacidad para distinguir adecuadamente…
2329	5.84	25.09 ± 0.4	(3) Capacidad para mostrar el balance global…
2544	6.29	26.61 ± 0.4	(4) Capacidad para exponer razonadamente las…
2735	5.45	21.26 ± 0.2	(5) Se valorará la inclusión de referencias…
2648	6.50	27.69 ± 0.3	(6) Se valorarán igualmente los aspectos…

5 Conclusiones

En este artículo abordamos el problema de la elaboración automática de información de realimentación para ser suministrada a alumnos tras un proceso de evaluación por pares. Nuestra propuesta consiste en plantear el problema desde el enfoque de la clasificación multietiqueta graduada, para resolverlo luego mediante factorización de matrices, optimizando de forma simultánea los modeos para todas las etiquetas. Las aportaciones de este artículo pueden resumirse en lo siguiente:

- Presenta un problema inédito en la evaluación por pares: el suministro de frases de realimentación que expliquen la nota que recibe cada ejercicio.
- Un nuevo algoritmo para clasificación multietiqueta que utiliza factorización de matrices. En este artículo el aprendizaje de las etiquetas se hace utilizando un planteamiento ordinal, pero se puede modificar sin dificultad para que se pueda aplicar a un planteamiento cardinal.
- Utiliza 3 conjuntos de datos reales que se presentan por primera vez en este artículo. Corresponden a exámenes de asignaturas de distintos Grados de la Universidad asturiana.

Acknowledgments

Esta investigación ha sido subvencionada en parte por el Ministerio de Economía y Competitividad y por FEDER (Fondo Europeo de Desarrollo Regional) mediante el proyecto TIN2015-65069-C2-2-R (MINECO/FEDER). También queremos agradecer a los alumnos y profesores que colaboraron con nosotros en los exámenes en la Universidad de Oviedo en las asignaturas de Economía Española (Juan Vázquez), Derecho Constitucional (Francisco Bastida) e Información Contable para el Comercio (Mónica Álvarez Pérez).

Referencias

1. Bahamonde, A., Bayón, G.F., Díez, J., Quevedo, J.R., Luaces, O., del Coz, J.J., Alonso, J., Goyache, F.: Feature subset selection for learning preferences: A case study. In: Procs. ICML '04. pp. 49–56. (2004)
2. 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)
3. Gielen,S.,Peeters,E.,Dochy,F.,Onghena,P.,Struyven,K.: Improving the effectiveness of peer feedback for learning. Learning and Instruction 20(4), 304–315 (2010).
4. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD) (2002)
5. Kulkarni, C., Wei, K.P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., Klemmer, S.R.: Peer and self assessment in massive online classes. In: Design Thinking Research, pp. 131–168. Understanding Innovation, Springer (2015)
6. Liu, N.F., Carless, D.: Peer feedback: the learning element of peer assessment. Teaching in Higher education 11(3), 279–290 (2006)
7. 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 85, 322 – 328 (2015)
8. Luaces, O., Díez, J., Joachims, T., Bahamonde, A.: Mapping preferences into euclidean space. Expert Systems with Applications 42(22), 8588 – 8596 (2015)
9. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., Koller, D.: Tuned models of peer assessment in MOOCs. In: Procs. of EDM'13. pp. 153–160. (2013)
10. Raman, K., Joachims, T.: Methods for ordinal peer grading. In: ACM Conference on Knowledge Discovery and Data Mining (KDD) (2014)
11. Raman, K., Joachims, T.: Bayesian ordinal peer grading. In: Proceedings of the ACM Conference Learning @ Scale '15. pp. 149–156. (2015)
12. Sadler, P.M., Good, E.: The impact of self-and peer-grading on student learning. Educational Assessment 11(1), 1–31 (2006)
13. 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)
14. Tseng, S.C., Tsai, C.C.: On-line peer assessment and the role of the peer feedback: A study of high school computer course. Computers & Education 49(4), 1161 – 1174 (2007)