Actas del Simposio-Taller XXII Jenui. Almería, 5 de julio 2016
ISBN: 978-84-16642-29-8
Páginas: 19-26

# ¿Por qué no evaluamos la evaluación? Un esbozo para un sistema de evaluación entre iguales

Rosana Satorre-Cuerda, Patricia Compañ-Rosique, Carlos Villagrá-Arnedo, Francisco Gallego-Durán, Rafael Molina-Carmona, Faraón Llorens-Largo
Departamento de Ciencia de la Computación e Inteligencia Artificial
Universidad de Alicante
Alicante
[rosana.satorre, patricia.company, villagra, fjgallego, rmolina, faraon.llorens]@ua.es

## Resumen

Existen múltiples maneras de evaluar y distintos objetivos de la evaluación. Y el resultado de la evaluación es el que califica la calidad del trabajo realizado. Por tanto cualquier reflexión sobre la propia evaluación permitirá obtener mejores resultados, al juzgar los trabajos de la forma más justa posible. La evaluación entre iguales es una estrategia evaluativa cada vez más utilizada en el entorno universitario. Actualmente se está empleando en el ámbito de la docencia, pero cuenta con una amplia trayectoria en el campo de la investigación. Con este método, un autor revisa, generalmente de forma anónima, el trabajo de sus colegas que, a su vez, pueden convertirse en revisores de la obra de ese autor. Este sistema de evaluación entre iguales enriquece la propia evaluación, incluso puede llegar a ser una de las pocas maneras de evaluar al no existir una autoridad jerárquica en el tema. En esta investigación consideramos el proceso de evaluación entre iguales como un proceso de clasificación, en el que disponemos de varios clasificadores (los revisores) que, ante una entrada (el trabajo a revisar), deben asignar una determinada clase (calificación del trabajo). En este análisis se propone una métrica para valorar el grado de bondad de los revisores, con el objetivo de contribuir a mejorar la calidad de los procesos de evaluación entre iguales proporcionando una valoración más objetiva de la labor de los revisores.

## Abstract

There are many different ways to assess and different objectives of this assessment. The result of the evaluation process allows the grading of the work quality. Therefore any reflection on the evaluation itself allow for better outcomes, judging the work in the fairest possible way. Peer review is an evaluation strategy increasingly used in the university environment. It has now a central role in the field of teaching, but there is an extensive experience in the field of research. This evaluation method implies an author reviewing, usually anonymously, the work of his or her colleagues who, in turn, can become reviewers of the author's work. This peer evaluation system enriches the evaluation itself, and it can even become one of the few ways to assess the works in absence of a hierarchical authority on the subject. In this research we consider the process of peer review as a classification process, in which we have several classifiers (reviewers) that, before an input (work to revise) should assign a particular class (work grading). In this analysis a metric is proposed to assess the degree of goodness of the reviewers, in order to help improve the quality of peer review processes providing a more objective assessment of the reviewers task.

## Palabras clave

Clasificador, matriz de confusión, métrica.

## 1. Introducción

La evaluación entre iguales es una forma utilizada en distintos ámbitos para medir la calidad de un trabajo. Este tipo de evaluación se basa en someter el trabajo a la revisión de expertos del área en la que se enmarca el mismo. De esta manera, un experto revisa, generalmente de forma anónima, el trabajo de sus colegas que, a su vez, pueden convertirse en revisores de su propia obra. Se usa tradicionalmente en los procesos de aceptación de artículos en publicaciones y congresos o en la evaluación de proyectos de investigación y, a pesar de las críticas, se considera un método fiable y efectivo [4]. Más recientemente, este tipo de revisión entre iguales se ha aplicado en el ámbito educativo, resultando ser un instrumento de aprendizaje muy efectivo [2,3,5,6]. Esto se debe a la consideración de que el estudiante refuerza sus conocimientos en el momento en el que se ve con la obli-

gación de evaluar un trabajo. Con esta metodología el docente valora no sólo el trabajo que realiza el estudiante sino la capacidad de evaluación del trabajo de otro estudiante.

Para que una evaluación entre iguales sea efectiva, es necesario proporcionar una rúbrica bien definida. De esta manera, todos los revisores disponen de unos criterios de evaluación claros. Sin embargo, la realidad es que en toda revisión aparece una componente subjetiva que se debe tener en cuenta debido a las propias ideas que el revisor tiene sobre lo correcto o incorrecto, el grado de conocimiento del área en la que se enmarca el trabajo o la dedicación del revisor a la tarea encomendada. Ello nos lleva a plantearnos si es posible diseñar una herramienta que permita establecer criterios objetivos sobre los revisores, analizando cuestiones como el nivel de confianza del revisor, la casuística de aciertos/fallos con respecto al resto de compañeros, etc.

El proceso de evaluación entre iguales puede verse como un proceso de clasificación, en el que disponemos de varios clasificadores (los revisores) que, ante una entrada (el trabajo a revisar), deben asignar una determinada clase (por ejemplo, en el caso de un proceso de revisión para una publicación las clases serían “aceptar”, “aceptar con cambios” o “rechazar”; en el caso de una evaluación entre estudiantes sería la nota asignada) en base a unos determinados algoritmos de clasificación (los criterios establecidos en la rúbrica). Teniendo en cuenta este punto de vista, resultaría interesante intentar aprovechar algunas de las herramientas que tradicionalmente se emplean para valorar la precisión de un clasificador. Entre ellas, se pueden citar las matrices de confusión [7]. Una matriz de confusión nos permite visualizar mediante una tabla de contingencia la distribución de errores cometidos por un clasificador. Además de las herramientas que se emplean habitualmente en los clasificadores, también se puede incorporar a esa métrica otros factores como puede ser el grado de acuerdo que presenta un revisor con el resto de correctores que han valorado el mismo trabajo.

Este trabajo trata de contribuir a mejorar la calidad de los procesos de evaluación entre iguales proporcionando una valoración más objetiva del trabajo de los revisores y en consecuencia del trabajo del estudiante. Es necesario evaluar la propia evaluación entre iguales porque su utilidad no se limita a la labor docente complementando la tarea del profesor, sino que se extiende a otros entornos como la investigación donde se convierte frecuentemente en la única forma de evaluación.

## 2. Evaluación entre iguales

La labor del docente es enseñar a sus estudiantes, lo cual implica no sólo la impartición de conocimientos sino la evaluación de la adquisición de esos conocimientos. Para ello y con la intención de ser lo más objetivo posible se realizan exámenes, recogida de trabajos individuales, de trabajos colectivos, proyectos, prácticas y lo que nos concierne en estos momentos, corrección de trabajos por parte de nuestros estudiantes, es decir, evaluaciones entre iguales. En la evaluación entre iguales se valora no sólo los conocimientos sobre la asignatura que cada estudiante posee sino la capacidad de valoración de otros trabajos. Y es este aspecto el que se va a analizar en este trabajo.

Son variadas las justificaciones planteadas para aplicar la evaluación entre iguales en el ámbito de la docencia y diversos también los modos de calificar este tipo de evaluación. Por ejemplo, en [3] se presenta una experiencia de autoevaluación y evaluación por compañeros en la que no participar en la evaluación hace que se pierda la puntuación que conlleva la realización de una práctica, no contabilizando como nota. En otros, se considera la evaluación realizada por los estudiantes en la evaluación entre iguales sin matices, sin valoración, se tiene en cuenta simplemente por el hecho de que se realice sin grandes errores, con seriedad y coherencia [6]. Y en otros casos, se plantea la evaluación entre iguales como método para incrementar la responsabilidad y el aprendizaje del estudiante [2].

Vemos, por tanto, que la elección de este tipo de evaluación entre iguales surge para reforzar los conocimientos de los estudiantes al verse en la obligación de evaluar a otros compañeros. En otros, como complemento para incrementar la responsabilidad del estudiante en su propio proceso de formación. Sin embargo, en otros casos, surge por la necesidad de aliviar la carga de trabajo del docente aunque al final se incorpora como otro método interesante en el proceso enseñanza-aprendizaje. Pero sea cual sea la justificación para aplicar este método de evaluación, no se tiene en cuenta la validez de la evaluación emitida o tiene muy poco peso. Por tanto, queda pendiente la evaluación del propio proceso de evaluación, ya que como se ha visto, esta parte ha quedado relegada a una pequeña nota que forma parte del conjunto de notas final. En una evaluación global la parte de evaluación entre iguales es una parte más, no el objetivo final.

Algunos docentes sí realizan una comparación entre el resultado emitido por los revisores-estudiantes y el realizado por el propio docente, pero esto incrementa notablemente el trabajo del docente puesto que todos los trabajos deben ser revisados. Para evitar este exceso de trabajo encontramos el otro extremo en el que la valoración emitida por el resto de compañeros se considera apropiada sin entrar en apreciaciones. Ante esta situación, nos planteamos ¿podemos

quedarnos sólo con la evaluación eliminando al experto?

# 3. Evaluación entre iguales y clasificadores automáticos

En la evaluación entre iguales, cada participante debe valorar el trabajo de los demás, asignándole una nota o una categoría. Dicho de otra manera, ante la entrada que proporciona el trabajo a evaluar, el evaluador produce una salida en forma de clasificación. Para realizar esta clasificación, el revisor debe contar con una serie de criterios (por ejemplo, una rúbrica) que le permitan realizar su labor de una manera lo más objetiva posible.

Este proceso de evaluación puede asimilarse al que realizan los clasificadores automáticos. Un clasificador automático es un modelo computacional que asigna a un individuo, caracterizado mediante un conjunto de variables, una etiqueta de entre varias asociadas a diferentes clases. El algoritmo utilizado para la clasificación establece los criterios para poder realizarla. Más allá de las diferencias obvias entre los dos procesos, en ambos tenemos un individuo que debemos clasificar. En el caso de un clasificador automático, el individuo viene caracterizado mediante un conjunto de variables de diferente tipo, susceptibles de ser manejadas automáticamente. En el caso de un evaluador humano, el elemento que caracteriza al individuo es el trabajo que debemos revisar, con lo que la información disponible es mucho más rica pero menos estructurada y difícilmente automatizable. En todo caso, a partir de estas entradas debe aplicarse un algoritmo de clasificación, basado en métodos computacionales en un caso, y basado en una rúbrica y en una labor subjetiva de aplicación de esa rúbrica en el otro. Como resultado del algoritmo, en ambos casos se proporciona como salida una etiqueta que identifica la clase en la que se ha clasificado al individuo.

La pregunta clave es, ¿qué calidad tiene la clasificación obtenida? En el caso de los modelos computacionales los investigadores han empleado mucho esfuerzo en buscar maneras de comparar clasificadores atendiendo a los aciertos y a los fallos que se producen en la clasificación. Puesto que este tipo de métricas se basa en los resultados y no en las características técnicas del algoritmo, ¿sería posible aplicarlas en el caso de una clasificación realizada por personas? Esta es la hipótesis de partida de este trabajo.

Una de las medidas más sencillas y habituales para evaluar la calidad de un clasificador es su tasa de aciertos o exactitud (*accuracy*) y su tasa de fallos (*error rate*) [7]. En realidad son medidas complementarias puesto que *accuracy* = 1 – *error rate.*

Aunque la tasa de aciertos es un indicador muy popular y tiene la virtud de representar con un único valor una medida de la calidad de un clasificador, tiene el gran problema de asumir que el coste de una mala clasificación es el mismo en cualquier caso. Pongamos un ejemplo para explicar el problema: supongamos un clasificador que emite un diagnóstico sobre una enfermedad, es decir, a partir de un conjunto de valores referidos a pruebas diagnósticas o a síntomas, clasifica al paciente indicando si está afectado o no por la enfermedad. El clasificador tiene una tasa de acierto del 95%, es decir, sólo falla en el 5% de los casos. Debemos preguntarnos ¿estos fallos de clasificación, se refieren a pacientes que tienen la enfermedad pero son clasificados como sanos, o a pacientes sanos que son clasificados como enfermos? Evidentemente, el coste de una mala clasificación no puede ser el mismo, ya que en este caso será preferible un clasificador conservador, que clasifique bien a todos los pacientes enfermos aún a costa de empeorar el ratio de aciertos, a un clasificador que tenga mayor tasa de aciertos pero clasifique como sanos a pacientes enfermos.

Frente a esta medida, se presentan otras métricas más completas que permiten analizar otros aspectos de los clasificadores. Supongamos el caso más sencillo de un clasificador binario. Formalmente, el clasificador, para cada individuo estima una etiqueta *P* (clase positiva) o *N* (clase negativa). Supongamos conocida la clase real a la que pertenece etiquetada como *p* o *n*, para distinguir las clases reales (en minúscula) de las estimadas por el clasificador (en mayúscula). Pueden darse cuatro resultados posibles:

- Si la muestra es positiva (*p*) y se clasifica como positiva (*P*), se contabiliza como un positivo verdadero (PV).
- Si la muestra es positiva (*p*) y se clasifica como negativa (*N*), se contabiliza como un falso negativo (FN).
- Si la muestra es negativa (*n*) y se clasifica como negativa (*N*), se contabiliza como un negativo verdadero (NV).
- Si la muestra es negativa (*n*) y se clasifica como positiva (*P*), se contabiliza como un falso positivo (FP).

Estos resultados se suelen representar en una matriz de confusión o tabla de contingencia (Cuadro 1). Esta matriz es la base de muchas métricas comunes y se puede extender al caso de más de dos clases.

| | p | n |
|---|---|---|
| P | 25 | 1 |
| N | 2 | 19 |

Cuadro 1. Matriz de confusión o tabla de contingencia para un clasificador binario

La diagonal principal muestra las clasificaciones correctas (positivos y negativos verdaderos), y la opuesta los errores o confusiones (falsos positivos y negativos). A partir de la matriz de confusión se pueden calcular varias métricas relacionadas:

- Tasa de aciertos (exactitud) = (PV+NV) / (P+N)
- Precisión = PV/(PV+FP)
- Ratio de positivos verdaderos (sensibilidad) = PV/P
- Ratio de negativos verdaderos (especificidad) = NV/N
- Ratio de falsos positivos (falsa alarma) = FP/N = 1-especificidad
- Ratio de falsos negativos = FN/P
- Medida F = 2/((1/precisión)+(1/sensibilidad))

En el caso de clasificadores multiclase, solo es posible distinguir entre positivos verdaderos, falsos positivos y falsos negativos. La matriz de confusión se construye indicando en las columnas la clasificación real y en las filas la clasificación dada por el clasificador. En el cuadro 2 se presenta una matriz de confusión para un clasificador con tres clases (a, b y c). En la primera columna, por ejemplo, se indica que de los 10 individuos de la clase a, el clasificador etiquetó como 8 como A, 2 como B y 0 como C.

| | a | b | c |
|---|---|---|---|
| A | 8 | 1 | 0 |
| B | 2 | 11 | 1 |
| C | 0 | 1 | 9 |

Cuadro 2. Matriz de confusión o tabla de contingencia para un clasificador multiclase

A partir de la matriz de confusión podemos calcular las siguientes métricas para una determinada clase *x*:

- Los positivos verdaderos (PV) se encuentran en la diagonal, en la posición (*x*,*x*).
- Los falsos positivos (FP) se calculan como la suma de la fila *x* sin la diagonal principal.
- Los falsos negativos (FN) se calculan como la suma de la columna *x* sin la diagonal principal.
- Precisión = PV/(PV+FP)
- Sensibilidad = PV/(PV+FN)
- Medida F = 2/((1/precisión)+(1/sensibilidad))

De estos indicadores nos interesan particularmente dos: la precisión y la sensibilidad. La precisión para una clase indica la proporción de individuos clasificados como de esa clase que realmente pertenecen a ella (aunque no dice nada sobre los individuos de la clase que se clasifican mal). Por su parte, la sensibilidad para una clase indica la proporción de individuos que pertenecen a la clase que se clasifican como de ella (aunque no dice nada sobre los individuos de otras clases que se clasifican como de ésta). Un clasificador ideal debe tener una precisión y una sensibilidad de 1. Esta situación, generalmente no se da, por lo que estos indicadores nos permiten elegir uno u otro clasificador en función de lo que queramos conseguir, privilegiando la precisión sobre la sensibilidad o al revés.

Además, es posible obtener indicadores globales de varias maneras. Una forma sencilla y ampliamente aceptada es utilizar la denominada precisión micro, calculada como la precisión para todos los individuos y todas clases [8].

Para poder evaluar un clasificador y poder tomar todas estas medidas hemos supuesto que la clasificación real de los individuos es conocida, es decir, es necesario disponer de una clasificación canónica de referencia (*gold standard test*) con la que comparar. Por ejemplo, en un sistema de reconocimiento de caracteres se puede utilizar la clasificación hecha por un humano como clasificación de referencia. Esta clasificación canónica debería ser una clasificación perfecta, sin embargo en la práctica muchas veces no está disponible. Por ejemplo, en el caso de la corrección de un trabajo entre iguales, la nota con la que se clasifica al trabajo siempre es subjetiva y no podemos establecer una clasificación perfecta, pero podemos considerar que la nota del profesor, como experto, es la de referencia. En el caso de la revisión de un artículo entre iguales, podríamos tomar la decisión final del editor como clasificación canónica, o incluso incorporar otras métricas de tipo bibliométrico para determinar el impacto de la publicación y suponer que el impacto es una medida de la calidad de la aportación [1].

# 4. Métrica propuesta

Habiendo establecido el paralelismo entre el proceso de revisión y un proceso de clasificación, llega el momento de diseñar la métrica que nos permita valorar la bondad del proceso. Un punto de comienzo puede ser utilizar herramientas que se emplean habitualmente para valorar la precisión de un clasificador.

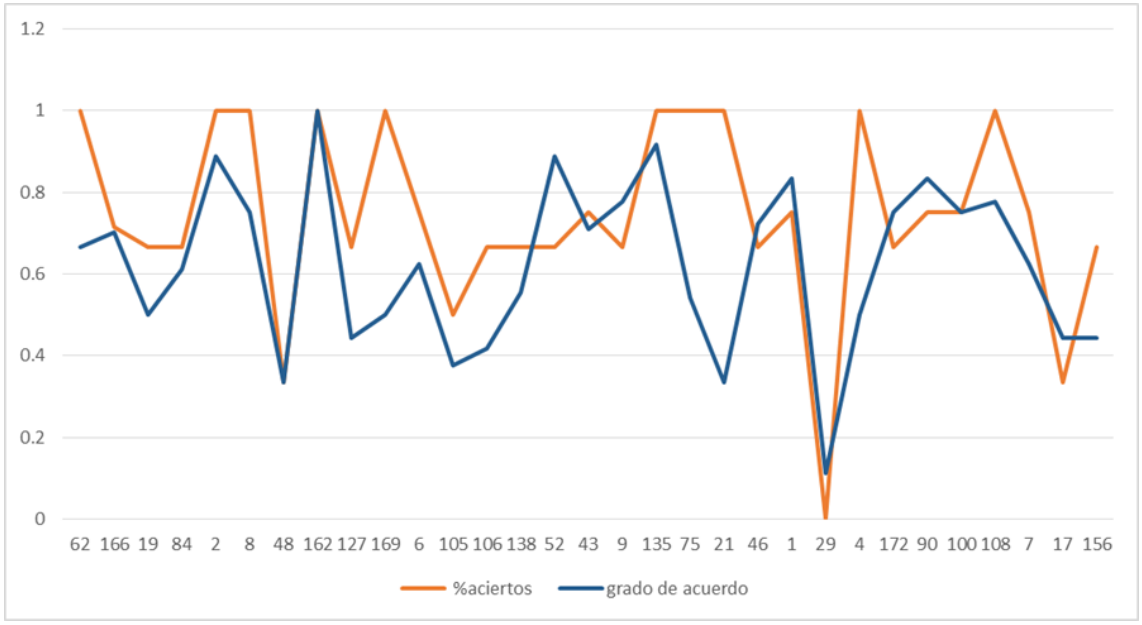


Figura 1: Porcentaje de aciertos y grado de acuerdo entre los revisores de una publicación

- Calculo de la matriz de confusión. Se trata de una matriz cuadrada de n x n, dónde n es el número de clases. Una matriz de confusión permite mostrar para cada revisor los aciertos que lleva. Se entiende por acierto que su valoración de un trabajo coincide con la valoración final que tiene el trabajo (la clasificación canónica de referencia). Se asigna al revisor un porcentaje de aciertos de tal manera que 1 indica acierto total, es decir que no hay falsos positivos ni falsos negativos.
- Grado de acuerdo de un revisor con las valoraciones de otros revisores para un mismo trabajo. La idea es para cada revisor, mirar para cada trabajo que ha revisado las valoraciones que han hecho los otros revisores. Para cada trabajo se obtiene una medida entre 0 y 1 según el grado de acuerdo que ha habido entre todos los revisores, 1 indica acuerdo total, es decir, todos los revisores han valorado dicho trabajo de la misma manera. Para calcular el grado de acuerdo de un revisor se divide la suma de las medidas obtenidas para cada trabajo entre el número total de artículos.

Se ha trabajado con dos ejemplos correspondientes a distintos casos: revisión de trabajos para una publicación (hemos utilizado los datos anonimizados de revisión de las propias JENUI) y evaluación entre estudiantes.

## 4.1. Revisión de trabajos para una publicación

Analizando sólo los factores comentados se pueden obtener interesantes conclusiones. Suponiendo como clasificación de referencia la decisión final (aceptar/rechazar) de los editores tomada a partir de los informes de los revisores, la figura 1 representa para un conjunto de 31 revisores los dos valores comentados en el apartado anterior: el porcentaje de aciertos y el grado de acuerdo. Cada revisor tiene asignado un código numérico, visible en los rótulos del eje X. La línea naranja representa el porcentaje de aciertos de los revisores mientras que la línea azul representa el grado de acuerdo de un revisor con el resto de revisores.

Algunos casos a destacar son:

- Revisor 162: presenta un 1 tanto en porcentaje de aciertos como en grado de acuerdo con el resto de revisores. Los valores de ambos factores están totalmente relacionados y parecen indicar que se trata de un revisor con una gran solidez y perspicacia en sus revisiones.
- Revisor 29: presenta un 0 en aciertos y un 0,11 en grado de acuerdo con el resto de revisores. Puede tratarse de un revisor novel o de alguien que ha tenido que valorar trabajos en los que no es un experto.
- Revisor 21: presenta un 1 en porcentaje de aciertos, sin embargo su grado de acuerdo es de 0,33. Esto puede indicar que los trabajos que ha revisado, han sido revisados por revisores con poca fiabilidad.

El cuadro 3 presenta las matrices de confusión para estos tres casos.

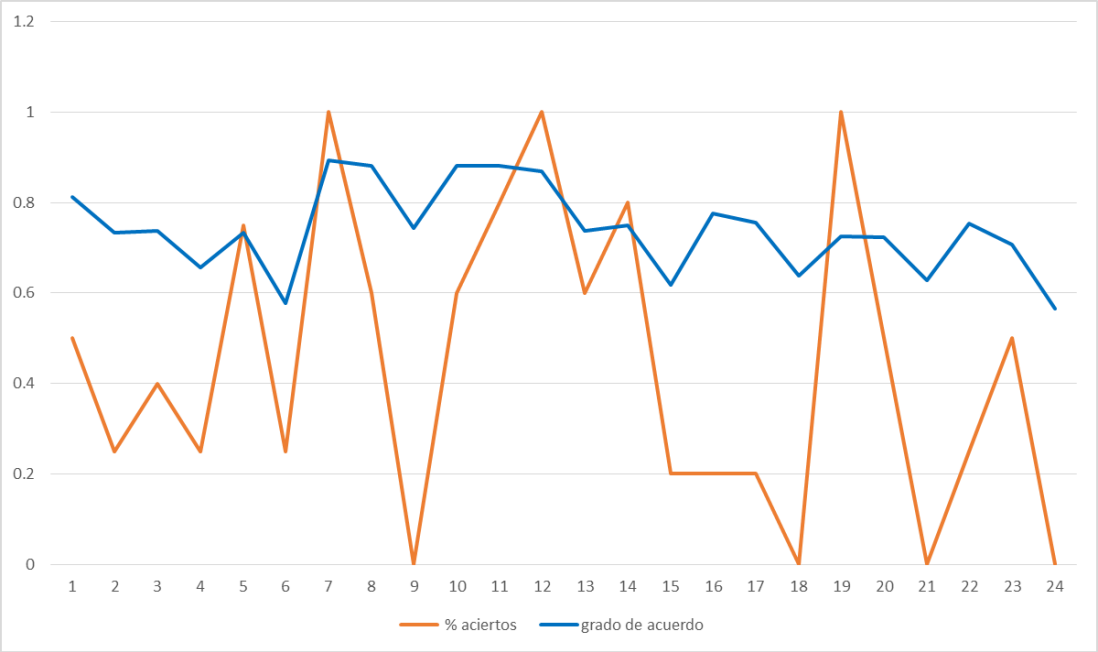


Figura 2: Porcentaje de aciertos y grado de acuerdo entre los revisores

| | p | n |
|---|---|---|
| P | 3 | 0 |
| N | 0 | 0 |

| | p | n |
|---|---|---|
| P | 0 | 2 |
| N | 1 | 0 |

| | p | n |
|---|---|---|
| P | 1 | 0 |
| N | 0 | 2 |

Cuadro 3: Matrices de confusión para los revisores 162, 29 y 21.

El modelo debería incorporar esta información así como otros factores como por ejemplo el grado de confianza de un revisor. El grado de confianza debería servir para ajustar la valoración ya que si un revisor no tiene mucha confianza sobre su dominio acerca del trabajo que está valorando, es comprensible que su evaluación no sea óptima mientras que si tiene una gran confianza, es de esperar que su evaluación sea lo más precisa posible.

Otro factor que sería interesante considerar es el factor tiempo. La experiencia de un revisor debe considerarse un valor añadido y por tanto debería de alguna forma incidir en el resultado final.

Una vez incorporados éstos y otros factores, la métrica sirve para valorar la bondad de un revisor, de tal forma que en el proceso de valoración final de un trabajo, se le puede dar más peso a la opinión de un revisor con un valor alto de la métrica mientras que la evaluación de revisores con bajo valor de la métrica se puede considerar con menor peso.

## 4.2. Evaluación entre estudiantes

La asignatura Innovación Tecnológica Aplicada del Master en Ingeniería Informática utiliza un sistema de evaluación entre iguales. Cada estudiante evalúa el trabajo de sus compañeros asignando una nota y el resultado final de la evaluación es la media de ellas. Este valor final se toma como clasificación de referencia. Se ha aplicado el modelo comentado previamente para tratar de valorar el proceso de evaluación seguido.

A diferencia del ejemplo anterior en el que las evaluaciones eran categóricas (aceptar/rechazar), en este caso se dispone de evaluaciones numéricas (notas entre 0 y 10). Se ha calculado el grado de acierto de cada evaluador así como su grado de acuerdo con el resto de evaluadores. Para obtener el grado de acuerdo de un revisor, se ha calculado el sumatorio de las diferencias en valor absoluto entre la nota puesta por él mismo y las notas puestas por los distintos revisores. Este cálculo se ha realizado para cada trabajo valorado y se ha realizado finalmente un proceso de normalización de tal forma que el valor 1 indica acuerdo total y el 0 desacuerdo total.

La figura 2 representa el porcentaje de aciertos y el grado de acuerdo para los 24 estudiantes de la asignatura. Se puede observar que hay bastante grado de acuerdo entre todos los estudiantes mientras que el porcentaje de aciertos es mucho más variable. La gráfica nos puede ayudar a detectar evaluadores con una alta discrepancia con los demás, por ejemplo los evaluadores 9, 18, 21 y 24 tienen una baja tasa de aciertos y podría no tenerse en cuenta sus evaluaciones para proporcionar más robustez al sistema.

Los cuadros 4 y 5 presentan matrices de confusión para dos revisores correspondientes a casos muy distintos. Aunque la matriz es de tamaño 10 x 10, sólo se muestra un extracto de la misma, ya que en el resto

de celdas el valor es 0. La columna 11 (FP) indica el número de falsos positivos por clase, la fila 11 (FN) presenta el falsos negativos por clase. Por último la columna 12 (Prec) y la fila 12 (Sens) indican la precisión y la sensibilidad por clase respectivamente. Tal y como se indicó previamente, la suma de las celdas de la diagonal indica el número de trabajos en los que coincide la evaluación del revisor con la evaluación final.

El cuadro 4 presenta la matriz de confusión para un revisor con una elevada solidez en sus evaluaciones: ha evaluado 5 trabajos y tiene una precisión por clase de 1 y una sensibilidad por clase de 1 para aquellas clases de las que se disponen ejemplos. Su grado de acuerdo con el resto de revisores es de 0.89.

| | **7** | **8** | **9** | **10** | **FP** | **Prec** |
|---|---|---|---|---|---|---|
| **7** | 0 | 0 | 0 | 0 | 0 | … |
| **8** | 0 | 1 | 0 | 0 | 0 | 1 |
| **9** | 0 | 0 | 4 | 0 | 0 | 1 |
| **10** | 0 | 0 | 0 | 0 | 0 | … |
| **FN** | 0 | 0 | 0 | 0 | … | … |
| **Sens** | … | 1 | 1 | … | … | … |

Cuadro 4. Matriz de confusión para el revisor 7

El cuadro 5 presenta la matriz de confusión para un revisor con bajo porcentaje de aciertos en sus evaluaciones: ha evaluado 5 trabajos y tiene una precisión por clase de 0 y una sensibilidad por clase de 0 para aquellas clases de las que se disponen ejemplos. Su grado de acuerdo con el resto de revisores es de 0.62.

| | **7** | **8** | **9** | **10** | **FP** | **Prec** |
|---|---|---|---|---|---|---|
| **7** | 0 | 0 | 0 | 0 | 0 | … |
| **8** | 0 | 0 | 0 | 0 | 0 | … |
| **9** | 0 | 0 | 0 | 0 | 0 | … |
| **10** | 2 | 0 | 2 | 0 | 4 | 0 |
| **FN** | 2 | 0 | 2 | 0 | … | … |
| **Sens** | 0 | … | 0 | … | … | … |

Cuadro 5. Matriz de confusión para el revisor 21

# 5. Conclusiones

La evaluación entre iguales se ha convertido en un elemento muy importante en los sistemas de evaluación. En algunos casos complementa a otros métodos de medida, como en el caso de la evaluación entre estudiantes que normalmente complementa a la valoración del profesor. En otros casos, sin embargo, se convierte en el elemento único o, al menos, primordial del proceso de evaluación. Tal es el caso de las revisiones entre iguales para las publicaciones o conferencias, o el proceso de revisión de proyectos de investigación para la obtención de ayudas. Los beneficios de las revisiones entre iguales se han destacado en muchos ámbitos, pero en este tipo de evaluación permanece una componente subjetiva inherente a los procesos en los que hay intervención humana. Esta componente puede ser interesante desde diversos puntos de vista pero debe controlarse convenientemente. En definitiva, es importante evaluar el propio proceso de evaluación. Esto nos ha llevado a plantearnos la pregunta clave que propusimos al principio del artículo: ¿Es posible establecer algún criterio de evaluación de la labor de los revisores en un sistema de evaluación entre iguales?

En este artículo hemos tratado de responder a esta pregunta estableciendo un paralelismo entre la labor de un revisor en un proceso de evaluación entre iguales y el funcionamiento de un clasificador automático. Esto nos ha permitido aprovechar las métricas habituales en la evaluación de la calidad de los clasificadores automáticos para establecer la calidad de la evaluación entre iguales. De esta manera podemos aprovechar el importante trabajo hecho en este ámbito para abrir una nueva línea de estudio sobre las revisiones entre pares.

Para ilustrar esta propuesta, hemos analizado dos casos distintos: el caso del proceso de revisión de una conferencia (en este caso las propias JENUI), y el caso de la evaluación entre iguales en una actividad correspondiente a una asignatura de un máster. Aunque ambos procesos parten del mismo principio, hay diferencias que indican la necesidad de adaptar las métricas a cada caso: en el primer caso se trata de una clasificación binaria (aceptar/rechazar) mientras que en el segundo la evaluación es numérica en el intervalo [0,10] aunque finalmente se ha tratado como una clasificación multiclase (con 10 clases, correspondientes a la división del intervalo en 10 rangos de notas). En ambos casos, además de las matrices de confusión y de la tasa de aciertos se ha construido un nuevo indicador que mide el grado de acuerdo de cada revisor con los demás. Este grado de acuerdo tiene diferente configuración según el tipo de clasificador del que hablemos (binario o multiclase) pero en cualquier caso permite detectar aquellos revisores que discrepan abiertamente de los demás. Una vez detectados estos revisores corresponde a los responsables

del sistema (los editores de la publicación o el profesor, según el caso) determinar qué acciones realizar. Por ejemplo, se podría determinar eliminar las evaluaciones de esos revisores por considerar que introducen valores atípicos, o se podría analizar el histórico de sus evaluaciones por si se trata de revisores con un punto de vista excéntrico pero interesante. En cualquier caso, el método permite establecer indicadores de la labor del revisor y detectar diferentes perfiles en los revisores.

Esta experiencia es muy preliminar y quedan muchas vías por estudiar, pero permite aventurar que puede ser una línea importante de trabajo en el futuro. A partir de este trabajo nos planteamos aplicar otras métricas habituales en el área de los clasificadores automáticos, definir nuestras propias métricas, elaborar un estudio de qué indica exactamente cada indicador y comprender y mejorar el proceso de evaluación entre iguales.

## Agradecimientos

Los autores queremos hacer constar nuestro agradecimiento a los responsables de JENUI por poner a nuestra disposición los datos anonimizados de revisión de estas jornadas en los últimos años.

## Referencias

[1] María Bordons y Mª Ángeles Zulueta. Evaluación de la actividad científica a través de indicadores bibliométricos. *Revista española de cardiología.* Vol. 52, núm. 10, octubre 1999.

[2] Reyes Grangel, Cristina Campos. Contratos de aprendizaje y evaluación entre iguales para responsabilizar al alumno de su aprendizaje. *Actas de las XIX Jornadas de Enseñanza Universitaria de Informática, Jenui 2013*, pp. 45-52. Castellón, 2013.

[3] Mercedes Marqués, José M. Badía y Ester Marínez-Martín, E., Una experiencia de autoevaluación y evaluación por compañeros. *Actas de las XIX Jornadas de Enseñanza Universitaria de Informática, Jenui 2013*, pp. 93-100. Castellón, 2013.

[4] Adrian Mulligan, Louise Hall and Ellen Raphael Peer. Review in a changing world: an international study measuring the attitudes of researchers. J. Am. Soc. Inf. Sci. Technol. 2013, vol. 64, nº 1, pp. 132-161. DOI: 10.1002/asi.22798.

[5] Miguel Riesco y Marián Díaz. La revisión entre iguales como herramienta de aprendizaje y evaluación en la asignatura de SS.OO. *Actas de las XIII Jornadas de Enseñanza Universitaria de Informática, Jenui 2007*, pp. 277-284. Teruel, 2007.

[6] Pablo Sánchez y Carlos Blanco. Una metodología para fomentar el aprendizaje mediante sistemas de evaluación entre pares. *Actas de las XIX Jornadas de Enseñanza Universitaria de Informática, Jenui 2013*, pp. 37-44. Castellón, 2013.

[7] Marina Sokolova, Guy Lapalme. A systematic analysis of performance measures for classification tasks. Journal Information Processing and Management. Volume 45, issue 4, July 2009. Pages 427–437. ISSN:0306-4573.

[8] Vicent Van Asch. Macro- and micro-averagedn evaluation measures [[BASIC DRAFT]]. September 9, 2013. Disponible en: `http://www.clips.uantwerpen.be/~vincent/pdf/microaverage.pdf`.