La identificación de enlaces ausentes como competición Kaggle para la enseñanza de teoría de redes

Virginia Ahedo, Ignacio Santos, José Manuel Galán, Luis R. Izquierdo

Recibido: 28 de Enero de 2022
Aceptado: 2 de Febrero de 2022

https://doi.org/10.37610/dyo.v0i79.634

Resumen

En este trabajo describimos una competición en Kaggle en la que los alumnos afrontan el reto de predecir un número de enlaces ausentes de una red social de usuarios que califican películas. Las competiciones "InClass" de Kaggle son una herramienta de gamificación poderosa que puede aprovecharse para mejorar la motivación y el interés de los alumnos mediante el trabajo competitivo en equipo. En esta contribución, mostramos cómo convertir un problema de ciencia de las redes en un problema de clasificación que puede beneficiarse de la infraestructura que ofrece Kaggle.

Palabras clave

Kaggle, gamificación, competición, ciencia de las redes, predicción de enlaces.

1. Introducción

"The purpose of teaching is to inspire the desire for learning in them [students] and make them able to think, understand, and question" (Richard Feynman). Esta frase del físico teórico Richard Feynman, premio nobel y excelente profesor, resume nuestra vocación como docentes. Pensar, comprender y cuestionar son sin duda actitudes que tratamos de inculcar a nuestros estudiantes, promoviendo a su vez que desarrollen el deseo por aprender. Para conseguir este último objetivo, existen múltiples técnicas y metodologías. De entre ellas, en los últimos años se ha popularizado especialmente la "gamificación", que consiste en utilizar elementos propios de las mecánicas de los juegos en contextos diferentes al juego (Deterding, Sicart, Nacke, O'Hara & Dixon 2011), como por ejemplo, en la enseñanza. Cabe destacar que en el ámbito de la docencia el uso del juego como pedagogía no es nuevo, si bien la digitalización de la enseñanza y las posibilidades que ofrecen las Tecnologías de la Información y Comunicación (TICs) están dando nuevos significados al concepto. Por ejemplo, importantes plataformas de e-learning como Khan Academy (https://es.khanacademy.org/) incorporan herramientas de gamificación para mejorar la motivación de los alumnos; algunos ejemplos son la choice board, los bingo boards, o su sistema de recompensa al alumno por objetivos conseguidos (Khan for Educators (US) 2021).

Existe evidencia significativa sobre los resultados de la gamificación en la mejora de la motivación, la atención y la participación de los alumnos (Dicheva, Dichev, Agre & Angelova 2015; Seaborn & Fels 2015). No obstante, la armonización de todos estos aspectos con los objetivos de aprendizaje de cada asignatura sigue siendo más un arte que una ciencia, y, por tanto, muy dependiente del profesor. El trabajo que describimos en este artículo se inscribe claramente en la corriente de gamificación. Hemos diseñado una competición "InClass" en Kaggle para una asignatura del Grado en Ingeniería Informática de la Universidad de Burgos que aborda el estudio de la ciencia de las redes (Barabási 2016; Newman 2018).

La ciencia de redes es un campo científico interdisciplinar capaz de ofrecer un marco teórico y metodológico común para abstraer y modelizar interacciones desde una perspectiva muy general. Su aplicabilidad a diversos contextos ha despertado un importante interés en la comunidad científica, contribuyendo a su desarrollo de forma notable en los últimos años (Havlin et al. 2012).

Los avances científicos en el campo han sido acompañados por un creciente interés docente en hacer la ciencia de redes más accesible a todos los niveles (Cramer, Porter, Sayama & Sheetz 2018; Tanizawa 2018; Cabrejas-Arce, Navarro, Ahedo & Galán 2021). Al tratarse de una disciplina que

✉ Virginia Ahedo (1)
vahedo@ubu.es
ORCID: 0000-0002-9812-388X

Ignacio Santos (1)
jisantos@ubu.es
ORCID: 0000-0002-6653-043X

José Manuel Galán (1)
jmgalan@ubu.es
ORCID: 0000-0003-3360-7602

Luis R. Izquierdo (1)
lrizquierdo@ubu.es
ORCID: 0000-0003-1057-4465

(1) Universidad de Burgos, Departamento de Ingeniería de Organización, Escuela Politécnica Superior, Ed. A1, Avda. Cantabria s/n 09006, Spain.

combina aportaciones de campos tan diversos como las matemáticas, la física, la biología, ciencias de la computación o las ciencias sociales (entre otras), no resulta fácil establecer unos contenidos generales únicos que puedan ser comunes a todos los cursos de redes; por este motivo, cada uno suele tener sus propios matices y/o contenidos específicos (Gera 2018; Sayama 2018).

El objetivo de este artículo es simple: proponer una tarea competitiva en el contexto de la identificación de enlaces ausentes para promover el interés por el estudio de la ciencia de las redes. La motivación por este problema en particular tiene diferentes vertientes. Por un lado, su estructura se puede adaptar bien, como veremos, a un problema de clasificación/regresión (tipología de problema para la que la plataforma Kaggle está especialmente diseñada). A su vez, se trata de un tema relativamente avanzado y con muchas soluciones posibles, que invita al alumno a consultar la literatura científica y le permite profundizar en múltiples aspectos. Por último, la identificación de enlaces ausentes es un problema con numerosas aplicaciones reales en reconstrucción de redes (Guimerà & Sales-Pardo 2009), sistemas de recomendación (Huang, Li & Chen 2005; Lü et al. 2012) y detección de spam (Huang & Zeng 2006), entre otras (Daud, Ab Hamid, Saadoon, Sahran & Anuar 2020; Kumar, Singh, Singh & Biswas 2020).

2. Kaggle "InClass"

Kaggle es una compañía perteneciente a Alphabet (Google) que mantiene la plataforma https://www.kaggle.com/. Esta plataforma ofrece a sus usuarios la posibilidad de publicar conjuntos de datos, hospedar competiciones en las que se pongan a prueba algoritmos de aprendizaje automático (machine learning), ejecutar código Python en ordenadores remotos (cloud computing), y aprender técnicas de aprendizaje automático a través de tutoriales y cursos online. Como consecuencia de todo lo anterior, Kaggle se ha convertido en un punto de encuentro de la comunidad de científicos de datos; en ella, los usuarios no solo discuten y aprenden sobre esta ciencia, sino que señalizan competencias profesionales (a través de sus diferentes insignias) y optan a premios en sus competiciones oficiales.

Kaggle recoge muchos elementos lúdicos (Zichermann & Cunningham 2011), como por ejemplo las competiciones, la formación de equipos, las herramientas de seguimiento del progreso de los marcadores, la reputación en forma de puntos e insignias, y los espacios de interacción social y discusión como son las comunidades y los foros. Por esta razón, Kaggle puede ser una magnífica herramienta de gamificación para utilizar en clase. Conscientes de este potencial uso en educación, desde Kaggle han creado un tipo particular de competiciones llamadas "InClass" (Kaggle Competitions 2021) que están especialmente orientadas a profesores y estudiantes.

Existen alternativas a Kaggle como plataformas para el establecimiento de competiciones y desafíos en ciencia datos, por ejemplo, DrivenData, Crowdanalytix, Signate, Zindi, Alibaba Cloud Tianchi, CodaLab y otras. No obstante, los servicios de Kaggle para diseñar, desplegar, realizar y gestionar competiciones de aprendizaje automático y especialmente su popularidad en la comunidad, la convierten en una opción adecuada para su uso docente y la familiarización de los alumnos con un entorno en el que posteriormente pueden profundizar. Existen tecnologías que pueden ser útiles también en otros tipos de juegos interactivos con fines docentes. Por ejemplo, Netlogo Hubnet (Tisue & Wilensky 2004; Kopf, Scheele, Winschel & Effelsberg 2005) permite que los alumnos interactúen y formen parte de la simulación de sistemas complejos (Wilensky & Stroup 2000; Pereda, Poza, Santos & Galán 2015). Si lo que se busca es la realización de experimentos sociales más sofisticados, como los propuestos en la economía experimental, el popular z-Tree (Fischbacher 2007) o la interesante plataforma open-source oTree (Chen, Schonger & Wickens 2016) utilizada por muchos centros de investigación de comportamiento entre un conjunto amplio de alternativas (Chan, Schilizzi, Iftekhar & Da Silva Rosa 2019).

Una competición en Kaggle propone siempre un problema de aprendizaje supervisado en el que se debe predecir el valor de una variable a partir de un conjunto conocido de variables regresoras o características (problemas de regresión o clasificación). El reto consiste en diseñar modelos de aprendizaje que obtengan los mejores rendimientos (de acuerdo con diferentes métricas de selección de modelos). Este tipo de problemas limita el contexto disciplinar en el que puede utilizarse Kaggle "InClass" a aquellas disciplinas relacionadas con las ciencias de la computación y la ciencia de datos en las que se requiera resolver problemas de predicción e implementar modelos sofisticados para tal fin. En lo que respecta a la ciencia de redes, si bien no suele contener este tipo de problemas, sí existe un problema particular: la identificación de enlaces ausentes, que puede formalizarse mediante un enfoque de aprendizaje supervisado, encajando perfectamente, por tanto, en una competición Kaggle. Los detalles de dicho problema son explicados en la siguiente sección.

3. Identificación de enlaces ausentes

El problema de predicción de enlaces (link prediction) o identificación de enlaces ausentes (missing links) suscita un enorme interés en la ciencia de redes. Generalmente, una red es una representación imperfecta de la topología de las interacciones reales entre sus constituyentes. En redes sociales, la imperfección suele derivarse de la utilización de datos incompletos para formalizar la red, lo que ocasiona que no aparezcan representadas por el correspondiente enlace relaciones entre individuos que realmente sí existen. Estas relaciones que no aparecen en nuestra red pero que sí existen

en la realidad se llaman enlaces ausentes, porque son enlaces que deberían formar parte de la red —como modelo fiel de la realidad— pero están ausentes. Por ejemplo, Berlusconi, Calderoni, Parolini, Verani & Piccardi (2016) estudian el problema de los enlaces ausentes en las redes de criminales, las cuales suelen caracterizarse por una significativa falta de datos.

Para comprender mejor el problema de identificación de enlaces ausentes, vamos a proceder a definirlo formalmente. Supondremos una red G={N,L} formada por un conjunto fijo de nodos N y un conjunto de enlaces (no dirigidos) L de cardinalidad l, para la que queremos predecir m enlaces ausentes de entre los $\binom{N}{2}$-l posibles enlaces que conforman el conjunto complemento L^C. Existen diferentes estrategias para resolver este problema

Una primera aproximación sería utilizar información intrínseca de los nodos (atributos no capturados por la red) para predecir nuevos enlaces. Por ejemplo, en una red en la que los nodos representasen personas, podríamos usar atributos como el sexo, el nivel socioeconómico o el país de residencia de las personas para tratar de inferir enlaces que no aparecen en nuestra red pero que probablemente existan en la realidad. La creación de modelos que relacionan la información intrínseca de los nodos con los enlaces puede abordarse desde el paradigma clásico del aprendizaje supervisado.

Sin embargo, suele ser más habitual aprovechar la información que conocemos de la red G para identificar los enlaces ausentes. En esta segunda aproximación, la topología de la red permite definir diferentes métricas sobre el conjunto de posibles enlaces L^C, lo que nos facilita identificar aquellos más probables. Por ejemplo, asumiendo la hipótesis de que dos nodos que comparten muchos vecinos tienen más probabilidad de tener un enlace entre ellos, una métrica sencilla sería el número de vecinos comunes entre dos nodos. Podemos calcular los vecinos comunes para cada uno de los pares de nodos del conjunto L^C, ordenar esta lista de forma decreciente, y seleccionar los m primeros pares como aquellos entre los que es más probable que existan enlaces ausentes. En los siguientes párrafos nos centramos en esta segunda aproximación, por lo que utilizaremos únicamente la información contenida en la red G.

A día de hoy, existen multitud de técnicas para la obtención de métricas y modelos que permiten identificar con éxito enlaces ausentes en una red a partir de su topología y/o información adicional (Gao, Musial, Cooper & Tsoka 2015). La mayoría de estos métodos pueden agruparse en las cuatro categorías representadas en la Fig. 1 (Kumar et al. 2020). Los métodos más básicos son las métricas de similitud, en las cuales nos vamos a centrar más adelante en este trabajo, explicando algunas técnicas con más detalle. Un segundo grupo es el constituido por los modelos probabilísticos. Esta aproximación estima un modelo de formación de la red en base a los datos observados, y a partir de dicho modelo hace predicciones sobre la formación de enlaces. La tercera categoría es la de los métodos basados en la reducción de la dimensionalidad. En esta aproximación se formaliza el problema como de clasificación binaria, lo cual conlleva la conocida maldición de la dimensionalidad en aprendizaje supervisado (Pecli et al. 2015); por este motivo, se emplean diferentes estrategias para reducirla. Por último, en el cuarto grupo se recogen otras aproximaciones basadas en estrategias diversas: aprendizaje, teoría de la información, métodos de perturbación estructural y aproximaciones mixtas, entre otras.

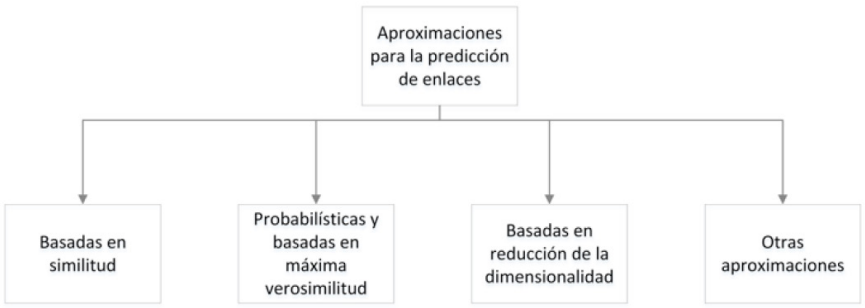


Figura 3 Taxonomía de aproximaciones para la predicción de enlaces (Kumar et al. 2020).

Los métodos más sencillos para la predicción de enlaces son los basados en una medida de similitud topológica entre nodos. Su utilización, de forma única o combinada, es uno de los objetivos docentes de la actividad docente propuesta, ya que muchas de las métricas están basadas en conceptos específicos de teoría de redes, lo cual favorece y motiva el aprendizaje de la asignatura. En todo caso, conviene aclarar que en la competición es posible utilizar cualquier algoritmo o método perteneciente a cualquiera de los grupos listados en la Fig. 1.

Todos los métodos de similitud parten de asignar un valor de similitud S(x,y) a cada par de nodos x e y no conectados, de acuerdo con las propiedades estructurales de los nodos en la red. El tipo de propiedad analizada para calcular ese valor se puede clasificar en local —si toman información basada en los primeros vecinos de cada uno de los nodos x e y— o global —si se requiere información topológica de la red en su conjunto para poder calcular su valor. Una vez que están calculadas las similitudes entre todas las parejas de nodos no conectados, se suele asumir que los enlaces más probables serán aquellos que unen a los nodos desconectados que tienen un mayor valor de similitud entre ellos. Sin ánimo de ser exhaustivos, explicaremos a continuación algunas de las métricas más habituales. Se pueden encontrar revisiones muy completas en Gao et al. (2015), Haghani & Keyvanpour (2019), Kumar et al. (2020) y Yuliansyah, Othman & Bakar (2020)

a. Vecinos comunes: en esta métrica —mencionada previamente— dada una red, el valor de similitud entre dos nodos x e y se define como la cardinalidad del conjunto intersección de Γ(x) y Γ(y), donde Γ representa el conjunto de primeros vecinos de cada nodo en la red. A pesar de su extrema sencillez tanto en términos de cálculo como de interpretación, es una métrica capaz de predecir con éxito nuevos amigos o la colaboración entre científicos (Newman 2001; Kossinets & Watts 2009).

$$S(x,y) = |\Gamma(x) \cap \Gamma(y)| \quad [1]$$

b. Coeficiente de Jaccard. Esta métrica es similar a "vecinos comunes", pero normaliza la cardinalidad de la intersección de primeros vecinos dividiendo por el número total de vecinos que tienen los nodos x e y.

$$S(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad [2]$$

c. Índice de asignación de recursos. (Resource allocation Index (Zhou, Lü & Zhang 2009)). En esta métrica se calcula la fracción de un recurso hipotético que un nodo puede enviar a otro a través de cada uno de sus vecinos comunes. Si z es un vecino común de x e y, y x envía una unidad de recurso a y a través de z, pero z reparte la unidad entre todos sus k_z vecinos, entonces y recibiría solamente la fracción $1/k_z$ del recurso. Al número k_z=#Γ(z) de primeros vecinos de un nodo z, se le denomina el grado de z. En un contexto social, esta métrica se puede interpretar como que un individuo con muchas relaciones no puede dedicar tanto "tiempo" a cada una de ellas como otro con menos. Formalmente:

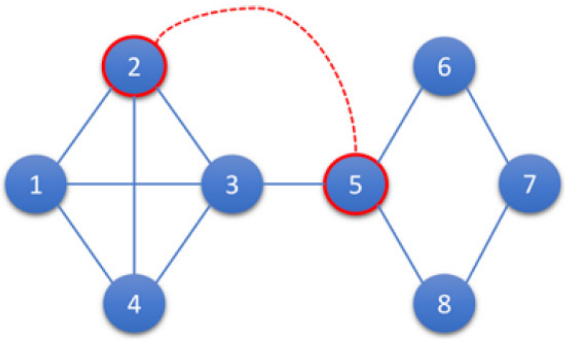
$$S(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \quad [3]$$

d. Índice de Adamic/Adar. Esta métrica recibe su nombre de los autores Adamic & Adar (2003). Es muy similar al resource allocation index, pero tomando logaritmos en el denominador. También implica que los vecinos comunes de menor grado aportan mayor valor a la métrica.

$$S(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z} \quad [4]$$

e. Preferential attachment. Este método se basa en el modelo propuesto por Barabási y Albert (1999) para explicar el mecanismo de formación de redes con distribución de grado tipo ley de potencia. En su modelo, un nuevo nodo se conecta de forma preferente (linealmente) con los nodos que tienen mayor grado en la red. De acuerdo con esta métrica, la similitud entre dos nodos es directamente proporcional a su grado.

$$S(x,y) = |\Gamma(x)| \cdot |\Gamma(y)| \quad [5]$$



$$S_{cn}(2,5) = |\{3\}| = 1$$

$$S_{jc}(2,5) = \frac{|\{3\}|}{|\{1,3,4,6,8\}|} = \frac{2}{5} = 0.2$$

$$S_{ra}(2,5) = \sum_{z\in\{3\}} \frac{1}{k_z} = \frac{1}{4} = 0.25$$

$$S_{aa}(2,5) = \sum_{z\in\{3\}} \frac{1}{\log(k_z)} = \frac{1}{\log(4)} = 0.72$$

$$S_{pa}(2,5) = |\{1,3,4\}||\{3,6,8\}| = 9$$

Figura 2 Ejemplo de cálculo de las métricas de similitud entre los nodos 2 y 5 de acuerdo con los métodos de: (a) vecinos comunes (S_{cn}), (b) coeficiente de Jaccard (S_{jc}), (c) resource allocation Index (S_{ra}), (d) índice de Adamic/ Adar (S_{aa}) y € preferential attachment (S_{pa}).

Otras métricas algo más sofisticadas utilizan la estructura de comunidad de la red para la predicción de los enlaces. En teoría de redes una comunidad es un conjunto de nodos densamente conectados entre sí, pero menos densamente conectados con el resto de la red. Existen multitud de algoritmos diferentes para encontrar comunidades en una red (Fortunato & Hric 2016). La hipótesis de las métricas de similitud basadas en comunidades es que los pares de nodos que pertenecen a una misma comunidad y tienen muchos vecinos comunes, son similares.

f. Common Neighbor Soundarajan-Hopcroft score. Esta métrica generaliza de forma sencilla la métrica de vecinos comunes (1), bonificando a aquellos pares de nodos que pertenecen a la misma comunidad y tienen vecinos comunes en la misma comunidad (Soundarajan & Hopcroft 2012)

$$S(x,y) = |\Gamma(x) \cap \Gamma(y)| + \sum_{z\in\Gamma(x)\cap\Gamma(y)} f(z) \quad [6]$$

Donde

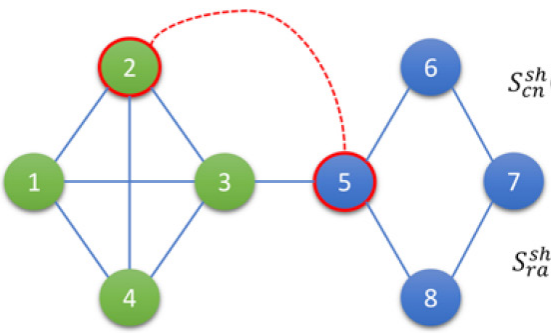
$$f(z) = \begin{cases} 1, \text{si } z \text{ está en la misma comunidad que } x \text{ e } y \\ 0, \text{en caso contrario} \end{cases} \quad [7]$$

g. Resource Allocation Soundarajan-Hopcroft score. Esta métrica generaliza en este caso el índice de asignación de recursos (3) considerando únicamente nodos que pertenecen a la misma comunidad (Soundarajan & Hopcroft 2012).

$$S(x,y) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{f(z)}{k_z} \quad [8]$$

Donde

$$f(z) = \begin{cases} 1, \text{si } z \text{ está en la misma comunidad que } x \text{ e } y \\ 0, \text{en caso contrario} \end{cases} \quad [9]$$



$$S_{cn}^{sh}(2,5) = S_{cn}(2,5) + \sum_{z\in\{3\}} f(z) = 1 + 0 = 1$$

$$S_{ra}^{sh}(2,5) = \sum_{z\in\{3\}} \frac{f(z)}{k_z} = \frac{0}{4} = 0$$

Figura 3 Ejemplo de cálculo de las métricas de similitud entre los nodos 2 y 5 de acuerdo con los métodos (f) common Neighbor Soundarajan-Hopcroft score (S_{cn}^{sh}) y (g) resource Allocation Soundarajan-Hopcroft score (S_{ra}^{sh}). Los colores de los nodos representan una partición de la red en dos comunidades.

En las Fig. 2 y Fig. 3 se representan los valores de las diferentes métricas de similitud calculados para dos nodos (2 y 5) de una red sencilla, con objeto de facilitar la comprensión de cada una de ellas. En el material suplementario S1 se ofrece un notebook en Python con los valores de las métricas para todos los pares de nodos (enlaces posibles) y cómo calcularlas utilizando la librería NetworkX (Hagberg, Swart & Chult 2008).

Como ejemplo de métrica global podríamos considerar el índice de Katz, aunque hay muchos otros. En general, el problema de muchas de estas medidas globales es el coste computacional en redes grandes, que impide en muchas ocasiones su uso.

h. Índice de Katz. Esta medida de similitud considera todo el conjunto de caminos (paths) de longitud l que unen los nodos x e y, pero penaliza exponencialmente los caminos de mayor longitud mediante el parámetro β. Formalmente se puede expresar a partir de la matriz de adyacencia A de la red.

$$S(x,y) = \sum_{}^{\infty} \beta^l (A^l) \quad [8]$$

Existen multitud de algoritmos para la predicción de enlaces, dependiendo su bondad e idoneidad del contexto de aplicación, y normalmente no pudiéndose conocer esta a priori. Una técnica habitual para elegir uno (o una combinación de ellos) es seguir el procedimiento común en aprendizaje automático para la selección de mode-los: Inicialmente se dividen los datos en entrenamiento y test (una o varias veces), se ejecuta entonces cada algoritmo en base a la información de los datos de entrenamiento y se predice sobre el conjunto de test. A continuación, se recoge la precisión obtenida con cada algoritmo y se comparan los resultados entre sí, seleccionándose el algoritmo de mayor precisión sobre el conjunto de test. Una vez elegido el algoritmo, se utilizan todos los datos disponibles para hacer predicciones futuras.

4. Competición "InClass" de predicción de enlaces

En esta sección describimos el proceso de diseño y puesta en marcha de nuestra competición "InClass", la cual ha sido titulada "Link prediction for social networks", y se encuentra publicada en *https://www.kaggle.com/c/link-prediction-for-social-networks*. Kaggle dispone de un asistente para crear competiciones "InClass" y también ofrece una guía de configuración (Kaggle Competitions 2021).

Figura 4 Ejemplo de la tabla de calificaciones de MovieLens. Los identificadores de usuario y de película, así como las calificaciones, están escogidos arbitrariamente con la única intención de explicar el proceso de formalización de la red.

user_id	movie_id	rating
U1	M1	1
U2	M1	3
U2	M2	4
U2	M3	3
U3	M1	5
U3	M2	5
U4	M3	2
U5	M3	1

El primer paso es definir el conjunto de datos del problema. En nuestro caso, proponemos una red social de usuarios que evalúan películas y que hemos construido a partir de los datos de la web de recomendaciones https://movielens.org, la cual es administrada por el grupo de investigación GroupLens del departamento de Ciencias de la Computación e Ingeniería de la Universidad de Minnesota. GroupLens ha recogido los datos de valoración de un conjunto de películas, además de información demográfica de los usuarios (Harper & Konstan 2016). La tabla de calificaciones (ver Fig.4) puede representarse como una red pesada bipartita en la que los usuarios tienen un enlace con cada una de las películas que han visto y valorado, siendo el peso del enlace la calificación (1-5) de la película (ver Fig. 5 izquierda). Posteriormente, nosotros hemos proyectado la red bipartita sobre los usuarios, obteniendo una red unimodal pesada de usuarios, en la que existe un enlace entre dos usuarios si ambos han calificado al menos una misma película, siendo el peso de dicho enlace el número de películas evaluadas por ambos (ver Fig. 5. derecha).

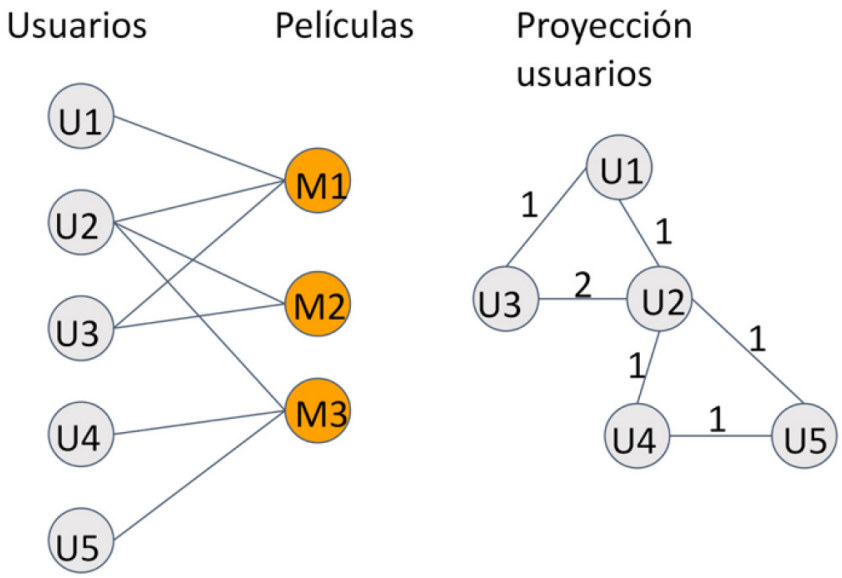


Figura 5 Representación de la tabla de calificaciones de la Fig. 1 como red bipartita (izquierda), y su proyección sobre usuarios como red unimodal pesada (derecha).

Para reducir la densidad de la red hemos considerado únicamente como enlace aquellas películas para las que la calificación de los dos usuarios es similar. La Fig. 6 muestra la red de usuarios final.

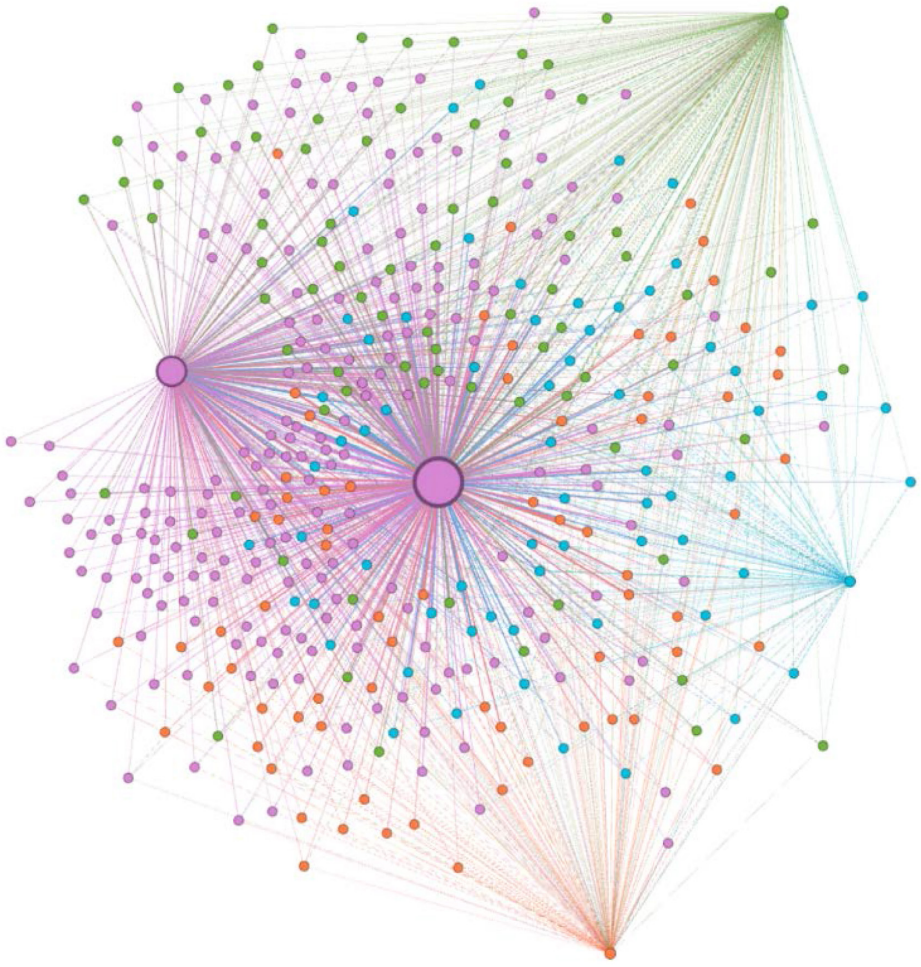


Figura 6 Visualización de la red unimodal de usuarios de MovieLens utilizada en la competición de predicción de enlaces. El tamaño de los nodos es proporcional a su grado y los colores muestran la partición de la red en comunidades de acuerdo al algoritmo de Louvain (Blondel, Guillaume, Lambiotte & Lefebvre 2008). Sólo han sido considerados como enlace aquellos derivados de películas en las que la calificación de ambos usuarios era similar.

La métrica utilizada para medir la similitud entre usuarios la mantenemos en secreto para evitar que los participantes puedan reconstruir la red a partir de los datos originales de MovieLens y distorsionar la competición de predicción.

Además, proporcionamos información intrínseca de cada usuario de naturaleza demográfica (edad, género, ocupación y código postal) y sobre sus preferencias. Puesto que cada película está clasificada en uno o más géneros (acción, aventuras, animación, infantil, comedia, crimen, etc.) resulta sencillo calcular para cada usuario el número de películas vistas de cada género. Estos metadatos de los nodos ofrecen la posibilidad de explorar modelos de predicción sofisticados que permitan mezclar técnicas de aprendizaje automático junto con técnicas de teoría de redes.

Una vez establecido el conjunto de datos, hay que especificar el criterio de evaluación de las predicciones. Kaggle ofrece diferentes métricas de tanteo para problemas de clasificación y de regresión. Nuestra competición de predicción de enlaces corresponde a un problema de clasificación. Los participantes dispondrán de una lista de enlaces en la que se asignará a cada elemento un 1 ó 0 en función de si se predice o no la presencia del enlace. Como medida de la bondad de una predicción hemos escogido el "F beta score" (Sokolova & Lapalme 2009), que computa una media armónica ponderada de la sensibilidad (recall, R) y la precisión (P)

$$F_\beta = (1+\beta^2)\frac{P*R}{\beta^2 P+R} \quad [9]$$

La sensibilidad (R) mide la ratio de enlaces verdaderos predichos con exactitud por un clasificador (Eq. 10), mientras que la precisión (P) mide la ratio de enlaces verdaderos de entre los enlaces predichos por el clasificador (Eq. 11). En las ecuaciones TP es el acrónimo en inglés de verdadero positivo (True Positive), FN de falso negativo (False Negative) y FP de falso positivo (False Positive).

$$R = \frac{TP}{TP+FN} \quad [10]$$

$$P = \frac{TP}{TP+FP} \quad [11]$$

Kaggle requiere subir a la plataforma los datos de entrenamiento y los datos de test. En nuestro caso, los datos de entrenamiento son la red de usuarios de MovieLens (ver Fig. 4), de la que se han eliminado 100 enlaces escogidos aleatoriamente. También proporcionamos información demográfica y de preferencias de los usuarios. La lista de parejas de usuarios candidatas a ser identificadas como enlaces ausentes comprende las $\binom{N}{2}$-l parejas desconectadas en la red original, más las 100 parejas correspondientes a los 100 enlaces eliminados. El objetivo es clasificar cada pareja de esta lista como un enlace ausente o no, proponiendo un total de 100 enlaces ausentes. Además, hay que proporcionar a Kaggle la solución para que pueda evaluar la entrega de un alumno.

Teniendo en cuenta que para esta competición se ha fijado en 100 el número de enlaces ausentes a predecir, es decir, TP+FN=TP+FP=100, la sensibilidad (R), la precisión (P), y el "F beta score" coinciden todos con el tanto por ciento de acierto. En cualquier otro caso en el que no se establezca esta limitación, proponemos fijar un β igual a 2 para sobreponderar la sensibilidad sobre la precisión, puesto que generalmente el número de enlaces verdaderos suele ser mucho más pequeño que el conjunto de enlaces posible L^C.

El asistente de Kaggle permite configurar los parámetros de la competición, subir los datos de entrenamiento, test y solución, subir ejemplos de entrega (para comprobar cómo funciona la métrica de tanteo), y escribir las diferentes secciones que conforman la página de la competición y orientan a los participantes (la Tabla 1 recoge los datos a cumplimentar de la página de una competición en Kaggle)

Tabla 1 Pestañas e información a cumplimentar de la página de una competición en Kaggle.).

Pestaña	Información a cumplimentar
Overview	Descripción de la competición y la evaluación.
Data	Ficheros con los datos de entrenamiento, test, y un ejemplo de entrega.
Rules	Descripción de las normas de la competición y fechas importantes.
Host	Título, imagen de la competición y fecha límite. Número máximo de entregas por día. Privacidad de la competición y enlace para invitar a los participantes. Fichero de solución y fichero de entrega de pruebas.

Un parámetro importante de la competición es el que define su privacidad. En nuestro caso, y puesto que solo queremos que participen los alumnos de una clase de la Universidad de Burgos, la competición se fija como privada, pudiéndose participar solamente previa invitación mediante el envío de un enlace. Una vez en marcha la competición, Kaggle gestiona la tabla de marcadores donde los alumnos pueden ver sus entregas ordenadas por el valor de tanteo obtenido. Además, ofrece un foro para la discusión entre participantes, así como un repositorio público de Notebooks donde puede publicarse código que pueda ser útil a los alumnos.

5. Conclusiones

La tarea de "gamificar" las actividades formativas para mejorar el interés y la motivación de los alumnos no suele ser sencilla. En este trabajo hemos puesto de manifiesto algunas de las características que ofrece Kaggle como herramienta de gamificación. Kaggle "InClass" brinda una poderosa infraestructura tecnológica que permite diseñar, desarrollar y evaluar competiciones. El proceso de creación es sencillo, resulta muy fácil escalar el juego, y, además, ofrece una inmejorable visibilidad a las instituciones, los profesores y los alumnos. La principal dificultad se encuentra en el tipo de problemas que pueden modelizarse como competiciones en Kaggle. Estos deben enmarcarse necesariamente como problemas de aprendizaje supervisado de regresión o clasificación, lo que limita significativamente su aplicación en muchas disciplinas.

Sin embargo, hemos mostrado cómo aprovechar Kaggle en la enseñanza de la teoría de redes, a través del problema de identificación de enlaces ausentes, que puede implementarse de forma sencilla como una competición. Las soluciones a este problema son muy abiertas y permiten variadas aproximaciones, muchas de las cuales exigen el estudio en profundidad de los conceptos de la teoría de redes. Nuestro objetivo principal es el aprendizaje y la enseñanza de la ciencia de las redes, siendo la competición accesoria. Sin embargo, en este caso, nos ha resultado sencillo integrar muchos de los elementos de un juego con el contenido y el objetivo de la asignatura. Pensamos que, al menos para este ejemplo, Kaggle "InClass" ayuda a promover la curiosidad intelectual en los contenidos e incentiva mediante la competición la iniciativa del alumno, que puede explorar diversas aproximaciones para resolver el problema planteado.

Agradecimientos Los autores agradecen la financiación del Ministerio de Ciencia e Innovación español (RED2018-102518-T), del Ministerio de Economía, Industria y Competitividad (HAR2017-90883-REDC), de la Agencia Española de Investigación (PID2020-118906GB-I00/AEI/10.13039/501100011033) de la Junta de Castilla y León – Consejería de Educación (BDNS 425389) y de la Fundación la Caixa (2020/00062/001).

Material suplementario (S1)

Notebook en Python y NetworkX con las métricas de similitud explicadas en este trabajo calculadas para una red sencilla de ejemplo

https://anaconda.org/jismartin/link_prediction_ algorithms/notebook

Referencias

ADAMIC, L. A. & ADAR, E. (2003). «Friends and neighbors on the Web». Social Networks, 25(3), pp. 211–230, doi:10.1016/S0378-8733(03)00009-1.

BARABÁSI, A.-L. (2016). Network Science. Cambridge, UK: Cambridge University Press.

BARABÁSI, A.-L. & ALBERT, R. (1999). «Emergence of Scaling in Random Networks». Science, 286(5439), pp. 509–512, doi:10.1126/science.286.5439.509.

BERLUSCONI, G., CALDERONI, F., PAROLINI, N., VERANI, M. & PICCARDI, C. (2016). «Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis». PLOS ONE, 11(4), p. e0154244, doi:10.1371/journal.pone.0154244.

BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. & LEFEBVRE, E. (2008). «Fast unfolding of communities in large networks». Journal of Statistical Mechanics: Theory and Experiment, 2008(10), p. P10008, doi:10.1088/1742-5468/2008/10/P10008.

CABREJAS-ARCE, L. M., NAVARRO, J., AHEDO, V. & GALÁN, J. M. (2021). «NetExtractor. A Semi-automatic Educational Tool for Network Extraction Conceived to Differentiate by Student Interest». In: Herrero, A., Cambra, C., Urda, D., Sedano, J., Quintián, H., & Corchado, E. (eds.) The 11th International Conference on EUropean Transnational Educational (ICEUTE 2020). ICEUTE 2020. Cham: Springer, pp. 205–214.

CHAN, S. W., SCHILIZZI, S., IFTEKHAR, M. S. & DA SILVA ROSA, R. (2019). «Web-based experimental economics software: How do they compare to desirable features?». Journal of Behavioral and Experimental Finance, 23, pp. 138–160, doi:10.1016/j.jbef.2019.04.007.

CHEN, D. L., SCHONGER, M. & WICKENS, C. (2016). «oTree—An open-source platform for laboratory, online, and field experiments». Journal of Behavioral and Experimental Finance, 9, pp. 88–97, doi:10.1016/j.jbef.2015.12.001.

CRAMER, C. B., PORTER, M. A., SAYAMA, H. & SHEETZ, L. (2018). Network Science In Education. Network Science In Education. Cham, Switzerland: Springer.

DAUD, N. N., AB HAMID, S. H., SAADOON, M., SAHRAN, F. & ANUAR, N. B. (2020). «Applications of link prediction in social networks: A review». Journal of Network and Computer Applications, 166, p. 102716, doi:10.1016/j.jnca.2020.102716.

DETERDING, S., SICART, M., NACKE, L., O'HARA, K. & DIXON, D. (2011). «Gamification. using game-design elements in non-gaming contexts». In: Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11. New York, New York, USA: ACM Press, p. 2425.

DICHEVA, D., DICHEV, C., AGRE, G. & ANGELOVA, G. (2015). «Gamification in education: A systematic mapping study». Journal of Educational Technology & Society, 18(3), pp. 75–88.

FISCHBACHER, U. (2007). «z-Tree: Zurich toolbox for ready-made economic experiments». Experimental Economics, 10(2), pp. 171–178, doi:10.1007/s10683-006-9159-4.

FORTUNATO, S. & HRIC, D. (2016). «Community detection in networks: A user guide». Physics Reports, 659, pp. 1–44, doi:10.1016/j.physrep.2016.09.002.

GAO, F., MUSIAL, K., COOPER, C. & TSOKA, S. (2015). «Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics». Scientific Programming, 2015, pp. 1–13, doi:10.1155/2015/172879.

GERA, R. (2018). «Leading Edge Learning in Network Science». In: Network Science In Education. Cham, Switzerland: Springer, pp. 23–44.

GUIMERÀ, R. & SALES-PARDO, M. (2009). «Missing and spurious interactions and the reconstruction of complex networks». Proceedings of the National Academy of Sciences, 106(52), pp. 22073–22078, doi:10.1073/pnas.0908366106.

HAGBERG, A., SWART, P. & CHULT, D. S. (2008). Exploring network structure, dynamics, and function using NetworkX (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos, NM: Los Alamos National Lab.

HAGHANI, S. & KEYVANPOUR, M. R. (2019). «A systemic analysis of link prediction in social network». Artificial Intelligence Review, 52(3), pp. 1961–1995, doi:10.1007/s10462-017-9590-2.

HARPER, F. M. & KONSTAN, J. A. (2016). «The MovieLens Datasets». ACM Transactions on Interactive Intelligent Systems, 5(4), pp. 1–19, doi:10.1145/2827872.

HAVLIN, S., KENETT, D. Y., BEN-JACOB, E., BUNDE, A., COHEN, R., HERMANN, H., KANTELHARDT, J. W., KERTÉSZ, J., KIRKPATRICK, S., KURTHS, J., PORTUGALI, J. & SOLOMON, S. (2012). «Challenges in network science: Applications to infrastructures, climate, social systems and economics». The European Physical Journal Special Topics, 214(1), pp. 273–293, doi:10.1140/epjst/e2012-01695-x.

HUANG, Z., LI, X. & CHEN, H. (2005). «Link prediction approach to collaborative filtering». In: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05. New York, New York, USA: ACM Press, p. 141.

HUANG, Z. & ZENG, D. D. (2006). «A Link Prediction Approach to Anomalous Email Detection». In: 2006 IEEE International Conference on Systems, Man and Cybernetics. IEEE, pp. 1131–1136.

KAGGLE COMPETITIONS (2021). Kaggle Competitions. https://www.kaggle.com/c/about/inclass [2021-03-2].

KHAN FOR EDUCATORS (US) (2021). Khan for Educators (US). https://www.khanacademy.org/khan-for-educators/k4e-us-demo [2021-03-2].

KOPF, S., SCHEELE, N., WINSCHEL, L. & EFFELSBERG, W. (2005). «Improving Activity and Motivation of Students with Innovative Teaching and Learning Technologiesitle». In: Proc. of Methods and Technologies for Learning. Palermo: WIT Press, pp. 551–556.

KOSSINETS, G. & WATTS, D. J. (2009). «Origins of Homophily in an Evolving Social Network». American Journal of Sociology, 115(2), pp. 405–450, doi:10.1086/599247.

KUMAR, A., SINGH, S. S., SINGH, K. & BISWAS, B. (2020). «Link prediction techniques, applications, and performance: A survey». Physica A: Statistical Mechanics and its Applications, 553, p. 124289, doi:10.1016/j.physa.2020.124289.

LÜ, L., MEDO, M., YEUNG, C. H., ZHANG, Y.-C., ZHANG, Z.-K. & ZHOU, T. (2012). «Recommender systems». Physics Reports, 519(1), pp. 1–49, doi:10.1016/j.physrep.2012.02.006.

NEWMAN, M. E. J. (2001). «Clustering and preferential attachment in growing networks». Physical Review E, 64(2), p. 025102, doi:10.1103/PhysRevE.64.025102.

NEWMAN, M. E. J. (2018). Networks. 2nd ed. New York, NY, USA: Oxford University Press.

PECLI, A., GIOVANINI, B., C. PACHECO, C., MOREIRA, C., FERREIRA, F., TOSTA, F., TESOLIN, J., VINICIUS DIAS, M., FILHO, S., CLAUDIA CAVALCANTI, M. & GOLDSCHMIDT, R. (2015). «Dimensionality Reduction for Supervised Learning in Link Prediction Problems». In: Proceedings of the 17th International Conference on Enterprise Information Systems. SCITEPRESS - Science and and Technology Publications, pp. 295–302.

PEREDA, M., POZA, D., SANTOS, J. I. & GALÁN, J. M. (2015). «Quality Uncertainty and Market Failure: An Interactive Model to Conduct Classroom Experiments». In: Herrero, Á., Baruque, B., Sedano, J., Quintián, H., & Corchado, E. (eds.) International Joint Conference CISIS'15 and ICEUTE'15. Cham: Springer, pp. 549–557.

SAYAMA, H. (2018). «Mapping the Curricular Structure and Contents of Network Science Courses». In: Cramer, C., Porter, M. A., Sayama, H., Sheetz, L., & Uzzo, S. (eds.) Network Science In Education. Cham, Switzerland: Springer, pp. 101–116.

SEABORN, K. & FELS, D. I. (2015). «Gamification in theory and action: A survey». International Journal of Human-Computer Studies, 74, pp. 14–31, doi:10.1016/j.ijhcs.2014.09.006.

SOKOLOVA, M. & LAPALME, G. (2009). «A systematic analysis of performance measures for classification tasks». Information Processing & Management, 45(4), pp. 427–437, doi:10.1016/j.ipm.2009.03.002.

SOUNDARAJAN, S. & HOPCROFT, J. (2012). «Using community information to improve the precision of link prediction methods». In: Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion. New York, New York, USA: ACM Press, p. 607.

TANIZAWA, T. (2018). «Network Science in Your Pocket». In: Cramer, C. B., Porter, M. A., Sayama, H., Sheetz, L., & Uzzo, S. M. (eds.) Network Science In Education. Cham, Switzerland: Springer, pp. 189–199.

TISUE, S. & WILENSKY, U. (2004). «NetLogo: Design and Implementation of a Multi-Agent Modeling Environment». In: Macal, C. M., Sallach, D., & North, M. J. (eds.) Proceedings of the Agent 2004 Conference on Social Dynamics: Interaction, Reflexivity and Emergence. Chicago: Argonne National Laboratory, pp. 161–181.

WILENSKY, U. & STROUP, W. M. (2000). «Networked Gridlock: Students Enacting Complex Dynamic Phenomena with the HubNet Architecturee». In: Fishman, B. & O'Connor-Divelbiss, S. (eds.) Proceedings of The Fourth Annual International Conference of the Learning Sciences. Mahwah: Erlbaum, pp. 282–289.

YULIANSYAH, H., OTHMAN, Z. A. & BAKAR, A. A. (2020). «Taxonomy of link prediction for social network analysis: A review». IEEE Access, 8(1), pp. 183470–183487, doi:10.1109/ACCESS.2020.3029122.

ZHOU, T., LÜ, L. & ZHANG, Y.-C. (2009). «Predicting missing links via local information». The European Physical Journal B, 71(4), pp. 623–630, doi:10.1140/epjb/e2009-00335-8.

ZICHERMANN, G. & CUNNINGHAM, C. (2011). Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps. O'Reilly Media, Inc.