Lenguaje, conocimiento y tecnología educativa: nuevos enfoques de aplicación

Mireya Tovar Vidal
Claudia Zepeda Cortés
Darnes Vilariño Ayala
Juan Manuel González Calleros
Josefina Guerrero García

Editores

Dirección General de Publicaciones

Lenguaje, conocimiento y tecnología educativa: nuevos enfoques de aplicación

Lenguaje, conocimiento y tecnología educativa: nuevos enfoques de aplicación

Mireya Tovar Vidal
Claudia Zepeda Cortés
Darnes Vilariño Ayala
Juan Manuel González Calleros
Josefina Guerrero García
Coordinadores

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación
2022

Primera Edición **2022**
ISBN BUAP: 978-607-525-846-1

Hecho en México
Made in Mexico

Prólogo

El presente libro surgió como una iniciativa para establecer un estudio sobre los distintos avances en investigación que se llevan a cabo en la Facultad de Ciencias de la Computación de la Benémerita Universidad Autónoma de Puebla, que permitan al lector conocer sobre los distintos tópicos de investigación que se desarrollan en la misma. Por lo tanto, en este libro titulado "Lenguaje, conocimiento y tecnología educativa: nuevos enfoques de aplicación" se presentan diferentes áreas de investigación en la Ingeniería del Lenguaje y del Conocimiento en las Ciencias de la Computación durante el primer semestre del año 2022.

La obra incluye seis capítulos de investigación, en las áreas de visión computacional, redes neuronales, procesamiento de lenguaje natural, inteligencia artificial aplicada a la educación, entre otras.

Los capítulos que forman parte de esta obra fueron revisados por el sistema de doble par ciego y aprobados para su publicación por expertos en el área de conocimiento, lo que permitió asegurar su calidad científica en las áreas de estudio. A continuación se menciona la aportación de cada uno de ellos.

En el Capítulo 1 se realiza un estudio piloto de construcción de un clasificador y se presenta un análisis de los atributos seleccionados que intentan capturar la complejidad lingüística de un texto. El clasificador permite asignar etiquetas al material disponible y al nuevo. Particularmente, la clasificación por nivel de idioma plantea dos retos: la deseable independencia de dominio y la separación de clases.

En el Capítulo 2 se presentan seis redes neuronales creadas y entrenadas en un software de generación de redes neuronales diseñado en plataforma Matlab, el objetivo de estas redes es la identificación de números en imágenes. Estas redes neuronales permiten poner a prueba el avance en el software, el cual permite cambiar la estructura de la red neuronal y obtener el código para su entrenamiento y posteriormente para su ejecución sin modificar manualmente el código.

En el Capítulo 3 se hace una revisión de cómo se realiza la detección de objetos a partir de imágenes aéreas obtenidas con vehículos aéreos no tripulados. La revisión se centra en encontrar los métodos, bases de datos, métricas y procedencia de los trabajos. Se encuentra que la mayoría de trabajos desarrollan enfoques basados en deep learning.

En el Capítulo 4 se reúne los avances de las técnicas de corrección atmosférica de imágenes satelitales con un enfoque particular en las técnicas basadas en imágenes y las técnicas basadas en áreas geográficas que contienen características pseudo-invariantes.

En el Capítulo 5 se hace una revisión de las técnicas de comparación utilizadas en la metodología del record linkage. Esto, con el fin de reducir errores inherentes en la determinación de las probabilidades de vinculación entre los re-gistros. Reconociendo las consecuencias que implica imputar erróneamente características de un registro a otro, las medidas de comparación son impor-

tantes para lograr una adecuada vinculación. Con este propósito, se analizan dichos métodos, destacando los pros y contras, reportados en la literatura.

Por último, en el Capítulo 6 se presenta una revisión del estado del arte sobre trabajos relacionados con el descubrimiento de tópicos y el uso de aprendizaje profundo.

Finalmente, queremos agradecer a cada uno de los autores por su aportación, a nuestros revisores por su valiosa labor, a la Facultad de Ciencias de la Computación de la Benemérita Universidad Autónoma de Puebla y a todos aquellos cuya participación contribuyó para la publicación de este libro.

Los editores,
Mireya Tovar Vidal
Claudia Zepeda Cortés
Darnes Vilariño Ayala
Juan Manuel González Calleros
Josefina Guerrero García

Índice general

Prólogo IV

Capítulo 1. Hacia la construcción de un Sistema Tutor Inteligente: Estudio piloto de selección de atributos 1
Adelina Escobar Acevedo, Josefina Guerrero García

Capítulo 2. Análisis de seis estructuras diferentes de redes neuronales para identificación de números en imágenes 11
Daniel Marcelo González Arriaga, María Aurora Diozcora Vargas Treviño, Sergio Vergara Limon, Josefina Guerrero García

Capítulo 3. Detección de objetos en imágenes áreas de UAV: Una revisión 21
Edmundo Cortes-Vazquez, Amparo Palomino-Merino

Capítulo 4. Corrección Atmosférica de Imágenes Satelitales para la Teledetección: Descripción general y Técnicas 31
Arturo Jasso Garduño, Ignacio Muñoz Máximo, David Pinto

Capítulo 5. Record Linkage: una revisión de los métodos de comparación 41
Pierre Antoine Delice, María Josefa Somodevilla García

Capítulo 6. Descubrimiento de tópicos con aprendizaje profundo: una revisión preliminar sistemática del estado del arte 54
Ana Laura Lezama Sánchez, Mireya Tovar Vidal, José A. Reyes-Ortiz

Índice de autores 63

Compiladores 64

Revisores 65

Editores 66

Capítulo 1

Hacia la construcción de un Sistema Tutor Inteligente: Estudio piloto de selección de atributos

Adelina Escobar Acevedo, Josefina Guerrero García

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

adelina.escobar@alumno.buap.mx, josefina.guerrero@correo.buap.mx

Resumen. El módulo dominio de un Sistema Tutor Inteligente contiene los materiales de enseñanza que utiliza, por lo que es deseable la construcción de un clasificador para asignar etiquetas al material disponible y al nuevo. Particularmente, la clasificación por nivel de idioma plantea dos retos: la deseable independencia de dominio y la separación de clases. En el presente trabajo se realiza un estudio piloto de construcción de un clasificador y se presenta un análisis de los atributos seleccionados que intentan capturar la complejidad lingüística de los textos.

Palabras Clave: Clasificación de textos, Complejidad lingüística, Sistemas Tutores Inteligentes.

1 Introducción

La clasificación, de forma general, se define como la tarea de asignar a un objeto una de dos o más clases predefinidas (Sierra Araujo, 2006). Ante un problema de clasificación, uno de los pasos clave y fuerte precursor del éxito o fracaso de la tarea, consiste en la selección de atributos. Los atributos son aquellos datos que permiten describir los objetos y nos permitirán diferenciarlos o agruparlos. Es recomendable obtener la mayor cantidad de atributos posible para, posteriormente, identificar los relevantes y descartar los irrelevantes.

En la clasificación de textos se utiliza el vocabulario o subsecuencias (n-gramas) como atributos, esto es pertinente en tareas como clasificación temática donde ciertas palabras clave ayudan a diferenciar los temas (en este caso las clases). La principal desventaja de utilizar palabras es que se construyen clasificadores dependientes del dominio. Otro tipo de atributos son las características textuales.

En un Sistema Tutor Inteligente para apoyo en la lectura, es deseable la construcción de un clasificador de textos por complejidad, a fin de conformar el módulo dominio. El módulo dominio es el que contiene los materiales a utilizar por el tutor y deben

encontrarse clasificados bajo algún nivel de dificultad o guía curricular. Al construir un clasificador, un texto nuevo puede ser asignado a una clase automáticamente. El primer reto es la deseable independencia de dominios relacionada con el hecho de que, en la enseñanza de idiomas, se buscan textos lo más generales posibles. El segundo reto está relacionado con la separación entre niveles (en este caso, las clases), incluso los expertos poseen criterios difusos para indicar cuando una lectura es más compleja que otra.

Los tutores monolingües utilizan los grados escolares para diferenciar los niveles de lectura, sin embargo, en un tutor de segundo idioma lo adecuado es utilizar un marco por nivel de idioma. El utilizado en México es el Marco Común Europeo de Referencia para Lenguas, abreviado MCER o CEFR por sus siglas en inglés. Cabe mencionar que los grados escolares y los niveles de idioma no son directamente alienables. Incluso, el MCER indica que los niveles de idioma no son lineales, un nivel requiere comúnmente el doble del tiempo que el nivel previo para alcanzar las competencias (Instituto Cervantes, 2002).

Este trabajo presenta el análisis de un estudio piloto cuyo objetivo es identificar las características lingüísticas pertinentes para la clasificación de textos por nivel de idioma, para ello se construyó un árbol de clasificación. El resto del trabajo está organizado de la siguiente manera: en la sección 2 se muestran los criterios de complejidad lingüística. La sección 3 presenta una revisión de trabajos relacionados. La sección 4 contiene la descripción del corpus utilizado en este trabajo. La sección 5 experimentos y discusión, y finalmente, las conclusiones del trabajo se presentan en la sección 6.

2 Complejidad lingüística

La complejidad lingüística es un tema vasto que puede observarse desde varios ángulos. Algunos autores se enfocan en criterios sintácticos (cantidad de elementos), semánticos (calidad de la información) y léxicos, aunque otro nivel incluye la pragmática (Ochoa Sierra y Cueva Lobelle, 2020). Esta sección aborda algunas de las características sin ser exhaustiva.

La cantidad de palabras es un indicador común, es la base de la métrica de lecturabilidad más famosa, *Flesch Kincaid Grade Level* abreviada FKGL (McNamara et al., 2014). Sin embargo, determinar el número de palabras de un texto no es trivial sobre todo en el idioma inglés. Si bien los tokenizadores pueden considerar los espacios como un primer delimitador, existen múltiples consideraciones, por ejemplo, verbos compuestos (*get on with*) o palabras separadas por guion (*check-in*). En ese sentido, los *parsers* actuales[1], no sólo toman rasgos léxicos sino también sintácticos y semánticos para asegurar preservar significado durante la fragmentación.

[1] El *parser* Charniak se encuentra como recurso de Stanford https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/parser/charniak/CharniakParser.html

Para identificar estructuras se comparan los árboles sintácticos. Un árbol sintáctico es una representación gráfica que nos permite visualizar la estructura de la oración, inicia dividiendo la oración en la parte nominal y la verbal (sujeto y predicado) y conforme se va desglosando, quedan identificados los elementos individuales: artículos, sustantivos, verbos, preposiciones, entre otros, permitiendo establecer relaciones entre los componentes.

Las etiquetas PoS (*Parts of Speech*), permiten encontrar la frecuencia de ciertas partes de la oración, por ejemplo, conjugaciones de verbos, primera, segunda o tercera persona en sus formas singular o plural, adjetivos, entre otros. En este apartado, es particularmente importante la identificación de conectores. Los conectores no sólo son un indicador de cohesión (relación entre una oración y otra) sino que permiten identificar el propósito del mensaje, siendo aditivos, temporales, causales y de contraste (Graesser et al., 2004). Identificar los conectores y su propósito es una de las estrategias lectoras.

En cuanto a vocabulario, la lecturabilidad por ejemplo, considera que los textos son más fáciles de leer cuando existen palabras concretas y no abstractas, permitiendo a los lectores crear imágenes mentales sobre el texto estudiado, las palabras concretas se obtienen de rankings humanos (Graesser et al., 2017). Bajo el mismo principio, la edad de adquisición es un índice entre 100 y 700 puntos obtenido del vocabulario que aparece en los textos infantiles (Gilhooly y Logie, 1980), los autores determinan aproximadamente 2000 palabras significativas.

Coh-Metrix es una herramienta para análisis de textos que contempla 106 índices, distribuidos en once áreas: descriptivos, de facilidad del texto, de cohesión referencial, de Análisis Semántico Latente (LSA), de diversidad léxica, de conectividad, de modelo situacional, de complejidad sintáctica, de densidad sintáctica de patrones, de información del vocabulario y de lecturabilidad. Para otorgar cada índice, Coh-Metrix utiliza herramientas externas como *parsers*, árboles sintácticos, listas de vocabulario, entre otros, y obtiene índices de más de 50 PoS (Graesser et al., 2017).

3 Revisión del estado del arte

La clasificación de textos por nivel de idioma es una tarea compleja reservada a expertos de enseñanza de segundo idioma. Dado los altos costos en tiempo y esfuerzo, algunos autores han considerado el etiquetado automático ya que uno de los problemas es la falta de corpus libres etiquetados. El único corpus libre conocido por las autoras hasta el momento, OneStopEnglish, no está homologado con el MCER, es decir, los documentos no se encuentran separados por nivel de idioma sino por grado de simplificación (Escobar-Acevedo y Guerrero-García, 2021). Ante estas dificultades Wilkens et al. (2018) utilizaron ensayos de estudiantes de idiomas para entrenar un clasificador y etiquetar automáticamente un gran conjunto de documentos extraídos de Wikipedia, la desventaja es que las producciones de estudiantes no representan un corpus

etiquetado por expertos. El resto de los autores han construido sus propios corpus extrayéndolos de libros de texto, páginas web u otro tipo de material.

Hartmann et al. (2016) realizan clasificación monolingüe en portugués, probaron dividiendo los textos en 5 y 3 niveles de dificultad con una exactitud de 52% y 74 % respectivamente. Utilizaron lecturas extraídas de los libros de texto escolares y consideran componentes tanto lingüísticos como no lingüísticos etiquetados manualmente para entrenar una máquina de vectores de soporte.

Branco et al. (2014) construyen una herramienta de clasificación conforme al CEFR en portugués, su corpus es un conjunto de textos extraídos de exámenes de certificación, altamente desbalanceado (existen más documentos en ciertas clases que en otras), utilizan cuatro atributos para clasificar: densidad de vocabulario, longitud de las palabras, longitud de las oraciones y la métrica de lecturabilidad *Flesch Reading Ease Index*. La herramienta muestra gráficas y estadísticas del texto a fin de apoyar a un experto en la clasificación por nivel.

Dentro de las aplicaciones con Tutores Inteligentes, Balyan et al. (2020) utilizan los dos corpus manualmente etiquetados del tutor iSTART para realizar una clasificación jerárquica. Dichos textos son para lectores monolingües y están divididos por grado escolar, sin embargo, para este experimento los agrupan en tres clases. Crean el clasificador de dos formas, la primera utilizando los datos de Coh-Metrix normalizados y la segunda usando sólo la métrica FKGL, el primer modelo supera al segundo en más de 10%.

La et al., (2015) hacen adaptaciones a un trabajo de clasificación monolingüe para convertirlo en bilingüe, los cambios principales consisten en incluir el largo de las oraciones en la métrica de lecturabilidad y dividir los textos en 7 clases en vez de las 12 acostumbradas cuando son grados escolares. El entrenamiento se realiza con vocabulario porque los autores aseguran que no tiene restricciones de dominio.

Un trabajo más enfocado a utilizar atributos pertinentes es el de Kurdi (2020) quien conscientemente incluye atributos fonológicos, morfológicos, léxicos, sintácticos, discursivos y de complejidad psicológica (seis áreas lingüísticas) y compara con las métricas de lecturabilidad populares, su corpus esta creado con más de 6000 textos extraídos de cinco sitios web de lectura divididos en tres niveles. Concluye que las métricas muestran un resultado inferior que utilizar atributos lingüísticos.

A forma de resumen, se observa primero una falta de estandarización en los corpus, derivada principalmente de la carencia de grandes volúmenes de material etiquetado. Segundo, una tendencia mejorar la clasificación agregando atributos que reflejen la complejidad lingüística. Este trabajo aborda ambas tendencias, utiliza un corpus etiquetado de origen con CEFR y utiliza como atributos un conjunto de características textuales.

4 Corpus

En este trabajo se utilizó el corpus de British Council descargado directamente de su página web[2], los materiales son recursos libres para profesores y alumnos con temas de la vida diaria. Se consideraron 4 de los 5 niveles existentes: A2, B1, B2 y C1, etiquetados conforme al Marco de referencia europeo. No se recuperaron los documentos de nivel A1 debido al formato más visual que textual: horarios, tarjetas de presentación, breves conversaciones de mensajería instantánea que carecen o contienen pocas oraciones. Por la misma razón se retiró un documento de la clase B2, era una conversación por mensajería instantánea. La Tabla 1 muestra la distribución en documentos del corpus recabado y el promedio de palabras nos permite aproximar la longitud de los textos.

Tabla 1. Distribución del corpus

Nivel de idioma	Número de documentos	Promedio de palabras
A2	8	197
B1	12	351
B2	11	468
C1	12	533

Como se puede observar, el corpus no está balanceado, algunas clases contienen menor cantidad de documentos. Con clases desbalanceadas, los clasificadores favorecen a las clases con más ejemplos.

5 Experimentos y discusión

Una vez formado el corpus, se utilizó Coh-Metrix para realizar un análisis de cada documento, obteniendo las 107 métricas proporcionadas por la herramienta. Dado que se busca explicabilidad, se construyó un árbol de clasificación J48 (Figura 2) sin normalización ni reducción de atributos, con validación cruzada a cinco dobleces se obtiene una exactitud de 48.8372% y a diez dobleces 55.8140% de exactitud. De los 107 atributos disponibles, el algoritmo J48 seleccionó sólo 7, la descripción de cada uno se encuentra en la documentación de Coh-Metrix (Graesser et al., 2017) y se extrae en la Tabla 2.

[2] https://learnenglish.britishcouncil.org/skills/reading/pre-intermediate-a2

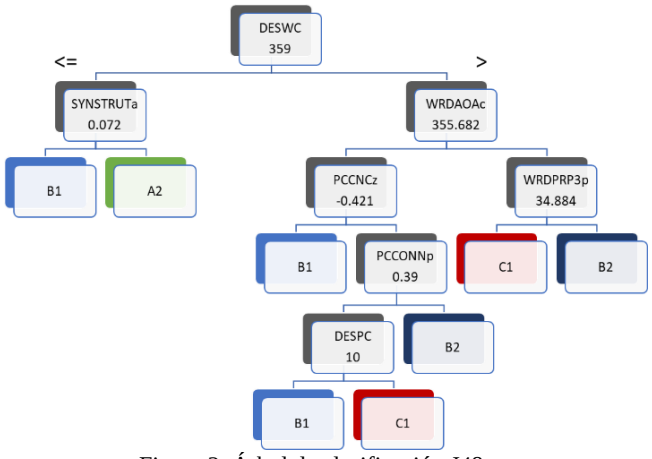


Figura 2. Árbol de clasificación J48

El árbol de clasificación de la Figura 2 indica que el atributo elegido para el nodo raíz fue el número de palabras del documento, por tanto, a pesar de los esfuerzos de varios autores por tener textos con longitud similar, el número de palabras sigue siendo el principal discriminante para la clasificación por nivel de idioma.

La rama izquierda usa similitud sintáctica entre oraciones para diferenciar entre los niveles A1 y B1, es decir, gramática constante en el documento. La rama derecha, para los niveles B1, B2 y C1, usa, como segundo criterio, la complejidad del vocabulario (por medio de la edad de adquisición) y el resto de los nodos de decisión consideran la incidencia de ciertas palabras en el texto: número de palabras concretas, 3era persona plural y conectores.

Tabla 2. Atributos Coh-Metrix

Atributo	Descripción
DESWC	Número de palabras calculadas usando el *parser* Charniak.
SYNSTRUTa	Similitud sintáctica entre estructuras adyacentes, proporción de intersección de nodos de los árboles entre oraciones adyacentes.
WRDAOAc	Edad de adquisición, un índice más elevado refleja que dicho vocabulario es adquirido a mayor edad por los niños.
PCCNCz	Índice z de palabras concretas (no abstractas).
WRDPRP3p	Incidencia de pronombres en tercera persona en forma plural.
PCCONNp	Promedio de conectividad, grado en que el texto contiene explícitamente conectores adversos, aditivos y comparativos para expresar las relaciones.
DESPC	Número de párrafos.

En la Figura 3 se observan los detalles de cada índice, para obtenerlas se ordenaron los documentos de cada clase de menor a mayor. Cada gráfica es independiente para su correcta visualización, nos permite conocer el rango de valores de cada clase, así como valores atípicos para un análisis particular.

En la Figura 3(a), se observa que la longitud de los documentos de la clase A2 es más constante con aproximadamente 200 palabras cada uno, mientras el resto de las clases contiene documentos con longitudes más variadas, por ejemplo, en la clase C1 el documento más corto tiene 375 palabras y el más largo 697.

En la Figura 3(b) el documento con mayor similitud sintáctica es el titulado "*English courses prospectus*", un folleto de cursos de idiomas que muestra el título del curso, horarios, precios, objetivos, etc. El formato es por tanto muy homogéneo.

El índice de edad de adquisición, Figura 3(c), calificado por Balyan et al. (2020) como el atributo más discriminante, muestra diferencias entre las clases con algunas excepciones. El documento "*proffesional-profile summaries*", es nivel A2 pero por contener resúmenes curriculares, tiene palabras como universidad, MBA, negocios, entre otras.

La Figura 3(d) se refiere a las palabras concretas, tomando nuevamente como ejemplo los documentos de la clase A2, el archivo con menor índice es "*A message to a new friend*", un mensaje de texto enviado a un amigo relacionado a planes y el de mayor índice es una descripción de un lugar para renta.

Para aparición de tercera persona plural, la Figura 3(e), el nivel A2 es interesante porque tiene valor de cero en todos los documentos excepto dos: "*A message to a new friend*" y el mayor "*study skill tips*" que enlista recomendaciones para un mejor aprendizaje.

En la Figura 3(f) el documento con mayor índice de conectores es publicidad de un gimnasio "*A flyer for a gym*" de la clase B1, resaltando los beneficios de unirse. Los valores bajos en los niveles avanzados sólo demuestran que la proporción de conectores con respecto a la longitud del documento es menor, es decir, las oraciones son más largas.

Finalmente, se observa en la Figura 3(g) que el atributo de número de párrafos no es confiable. Particularmente el archivo A2 muestra tener 44 párrafos y corresponde nuevamente a "*English courses prospectus*", por tanto, el formato del archivo no es adecuado para este análisis.

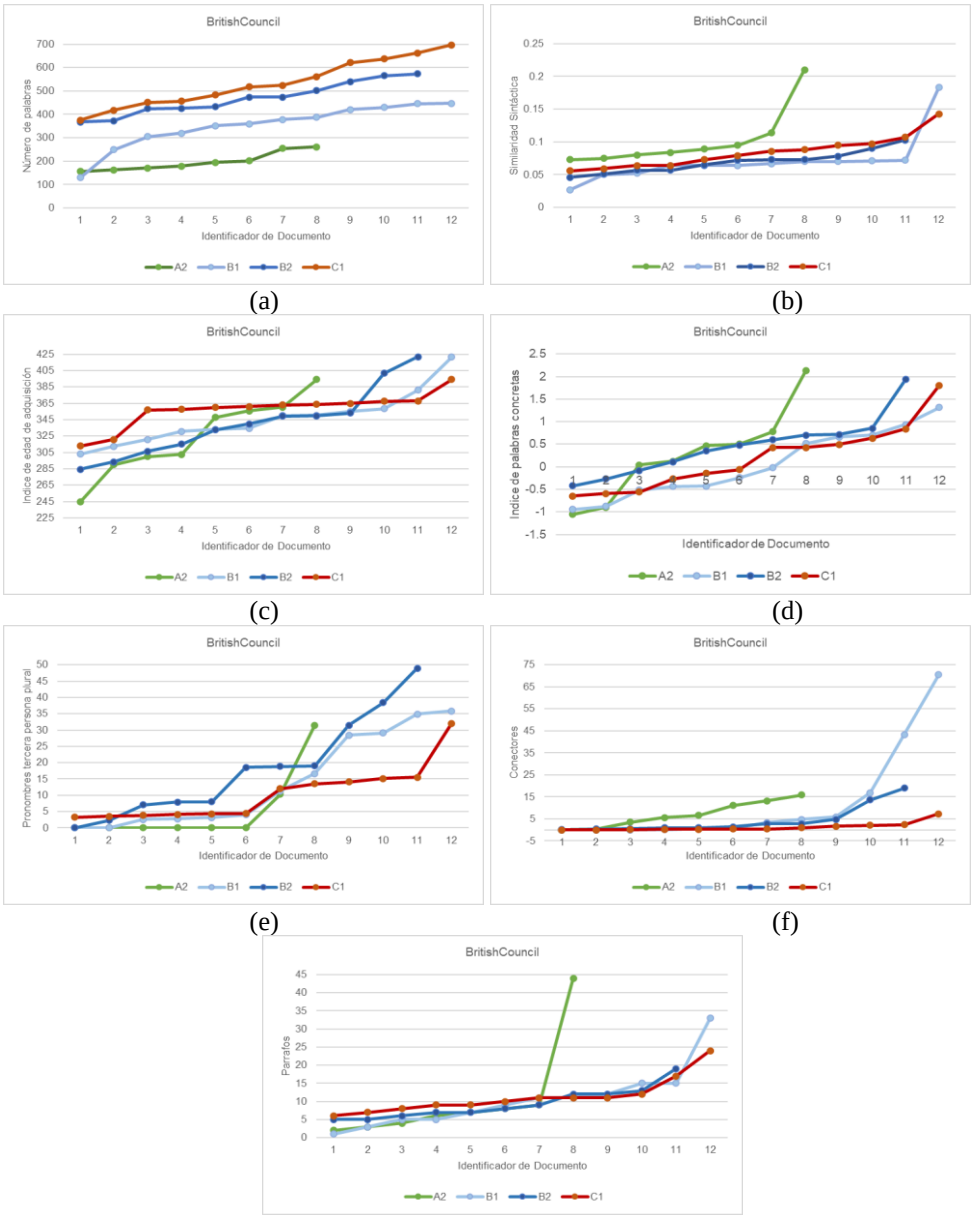


Figura 3. Atributos utilizados por el clasificador y ordenados por clase (a) DESWC, (b) SYNSTRUTa, (c) WRDAOAc, (d) PCCNCz, (e) WRDPRP3p, (f) PCCONNp, (g) DESPC

La matriz de confusión resultante se muestra en la Figura 4.

a	b	c	d	<-- clasificado como
7	1	0	0	a = A2
2	5	3	2	b = B1
0	3	4	4	c = B2
0	1	3	8	d = C1

Figura 4. Matriz de confusión

En la clase A1 la tasa de verdaderos positivos es de 0.875 y de falsos positivos sólo de 0.057. B1 es la más dispersa, se les asignaron todas las etiquetas existentes, si bien la tasa de verdaderos positivos es de 0.417 y falsos positivos 0.161. Por la estructura del árbol, la clase B2 y C1 no reciben etiquetas A2. La clase C1 es más precisa con el 0.571 de precisión, una tasa de verdaderos positivos de 0.667 y falsos positivos de 0.194.

6 Conclusiones

En este trabajo se construyó un árbol de clasificación de textos por nivel de idioma, el corpus utilizado fue etiquetado por expertos y es un recurso disponible para estudiantes y profesores de segundo idioma inglés. A fin de evitar la dependencia de dominio, se utilizaron como atributos los datos extraídos del análisis de Coh-Metrix, otro recurso gratuito y disponible. Los 106 datos son desde descriptivos hasta semánticos incluyendo tres métricas de lecturabilidad: Flesch_Kinkaid grade level, Flesch Reading Ease y RDL2.

Una vez construido el modelo, es posible identificar aquellos atributos relevantes para este corpus en particular, resulta importante indicar que el modelo no incluyó ninguna de las tres métricas de lecturabilidad disponibles en Coh-Metrix. Con lo anterior se refuerzan las conclusiones de los autores, las métricas de lecturabilidad no deben considerarse como un único factor de decisión para asignar en la tarea de clasificación de textos por nivel de idioma, es indispensable agregar atributos que reflejen la complejidad lingüística en tantos niveles como sea posible: datos descriptivos, sintácticos, semántico y de ser viable pragmáticos.

Este estudio es fundamental para la construcción del módulo dominio de un Sistema Tutor Inteligente en el área de lenguaje, ya que, el objetivo principal del tutor es emitir recomendaciones de material acorde al nivel de idioma de los estudiantes. Así, contribuye al área de enseñanza de idiomas.

Como trabajo futuro se desea construir un etiquetador automático por nivel de idioma para documentos no revisados por expertos, pero con alto grado de confiabilidad en la asignación del nivel a fin de expandir el material actual. Se planea experimentar con otros clasificadores y otros esquemas de selección de atributos.

Referencias

Balyan, R., McCarthy, K. S., y McNamara, D. S. (2020). "Applying Natural Language Processing and Hierarchical Machine Learning Approaches to Text Difficulty Classification". *International Journal of Artificial Intelligence in Education*, *30*(3), 337–370. https://doi.org/10.1007/s40593-020-00201-7

Branco, A., Rodrigues, J., Costa, F., y Silva, J. (2014). "Assessing Automatic Text Classification for Interactive Language Learning". *International Conference of Information Society (i-Society 2014)*, 70–78. https://doi.org/10.1109/i-Society.2014.7009014

Escobar-Acevedo, A., y Guerrero-García, J. (2021). "Construyendo el contenido de un Sistema tutor Inteligente: un estudio piloto de métricas populares de lecturabilidad para enseñanza de idiomas". En *Lenguaje, conocimiento y tecnología educativa: avances recientes*. Benemérita Universidad Autónoma de Puebla.

Gilhooly, K. J., y Logie, R. H. (1980). "Age-of-acquisition, imagery , concreteness , familiarity , and ambiguity measures for 1 ,944 words". *Behavior Research Methods & Instrumentation*, *12*(4), 395–427. https://doi.org/https://doi.org/10.3758/BF03201693

Graesser, A. C., McNamara, D. S., y Louwerse, M. M. (2017). *Coh-Metrix*. http://cohmetrix.com/

Graesser, A. C., McNamara, D. S., Louwerse, M. M., y Cai, Z. (2004). "Coh-Metrix: Analysis of text on cohesion and language". *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.

Hartmann, N., Cucatto, L., y Brants, D. (2016). "Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years". *Computational Processing of the Portuguese Language*, *1*, 12–24. https://doi.org/10.1007/978-3-319-41552-9

Instituto Cervantes. (2002). "Marco Común Europeo de Referencia para las Lenguas: aprendizaje, enseñanza, evaluación". En *Instituto Cervantes*. http://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/default.htm

Kurdi, M. Z. (2020). "Text Complexity Classification Based on Linguistic Information: Application to Intelligent Tutoring of ESL". *Journal of Data Mining & Digital Humanities*, *2020*. http://arxiv.org/abs/2001.01863

La, L., Wang, N., y Zhou, D. (2015). "Improving reading comprehension step by step using Online-Boost text readability classification system". *Neural Computing and Applications*, *26*(4), 929–939. https://doi.org/10.1007/s00521-014-1770-2

McNamara, D. S., Graesser, A. C., McCarthy, P. M., y Cai, Z. (2014). "Coh-Metrix Measures of Text Readability and Easability". En *Automated Evaluation of Text and Discourse with Coh-Metrix* (pp. 78–95). Cambridge University Press. https://doi.org/10.1017/CBO9780511894664.007

Ochoa Sierra, L., y Cueva Lobelle, A. (2020). "Complejidad lingüística. Revisión bibliográfica". *Signo y pensamiento*, *39*(77), 2027–2731. https://doi.org/https://doi.org/10.11144/Javeriana.syp39-77.pspi

Sierra Araujo, B. (2006). *Aprendizaje Automático: Conceptos básicos y avanzados* (M. M. Romo (ed.)). Pearson-Prentice Hall.

Wilkens, R., Zilio, L., y Fairon, C. (2018). "SW4ALL: a CEFR-Classified and Aligned Corpus for Language Learning". *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 365–370. http://www.lrec-conf.org/proceedings/lrec2018/pdf/1012.pdf

Capítulo 2

Análisis de seis estructuras diferentes de redes neuronales para identificación de números en imágenes

Daniel Marcelo González Arriaga[1], María Aurora Diozcora Vargas Treviño[2], Sergio Vergara Limon[2], Josefina Guerrero García[1]

[1] Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación
[2] Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Electrónica

dm.gar93@gmail.com, auroravargast@hotmail.com,
svergara2@hotmail.com, josefina.guerrero@correo.buap.mx

Resumen. En el presente trabajo se presentan seis redes neuronales creadas y entrenadas en un software de generación de redes neuronales diseñado en plataforma Matlab, el objetivo de estas redes es la identificación de números en imágenes. Estas redes neuronales permiten poner a prueba el avance en el software, el cual permite cambiar la estructura de la red neuronal y obtener el código para su entrenamiento y posteriormente para su ejecución sin modificar manualmente el código.

Palabras Clave: MNIST, redes neuronales, identificación de imágenes.

1 Introducción

En los últimos años, las técnicas de aprendizaje automático, particularmente cuando se aplican a las redes neuronales, han desempeñado un papel cada vez más importante en el diseño de sistemas de reconocimiento de patrones. De hecho, se podría argumentar que la disponibilidad de técnicas de aprendizaje ha sido un factor crucial en el éxito reciente de aplicaciones de reconocimiento de patrones como el reconocimiento continuo de voz y reconocimiento de escritura.

Actualmente el reconocimiento de patrones, números, letras y otros símbolos siguen siendo tema de investigación (Javier et al., 2022), en (M. Wang et al., 2022) proponen una red de separación de estructura-textura (STSN), que es un marco de aprendizaje de extremo a extremo para el desenredo, la transformación, la adaptación y el reconocimiento conjuntos de escritura en hueso de Oracle, que es el sistema de escritura chino más antiguo conocido de la dinastía Shang y es valioso para la arqueología y la filología. Otro problema ampliamente abordado es el reconocimiento de placas de automóvil, en (Chen & Wang, 2022) proponen un marco que combina una red totalmente convolucional con un amplio

sistema de aprendizaje para el reconocimiento de matrículas. El reconocimiento de movimiento preciso es esencial para dispositivos de asistencia como los exoesqueletos para lograr la comunión entre humanos y robots, en (Zhang & Tao, 2022) tienen como objeto el movimiento de las extremidades inferiores, y las señales de electromiografía de superficie de cinco pasos de subir escaleras sin peso, bajar escaleras sin peso, subir escaleras con peso, bajar escaleras con peso y caminar sobre una se recogieron superficies planas sin peso.

Los problemas de transporte se convierten en un desafío cuando el sistema y el comportamiento de los usuarios es demasiado difícil de modelar y predecir los patrones de viaje. Por lo tanto, la inteligencia artificial se considera una buena opción para el transporte. Un ejemplo de eso incluye transformar los sensores de tráfico en la carretera en un agente inteligente que detecta accidentes automáticamente y predice las condiciones futuras del tráfico (Klügl et al., 2010) Además, hay muchos métodos de IA (inteligencia artificial) que se utilizan en el transporte, como las ANN (Artificial Network Neuronal). Las ANN se pueden utilizar para la planificación de carreteras (Doğan & Akgüngör, 2013), transporte público (Akğngör & Doğan, 2009), detección de incidentes de tráfico (Dia, 2001; Dia & Rose, 1997; R. Wang et al., 2016), y predecir las condiciones del tráfico (Crutchfield, 2017; Huang et al., 2014; Jiang et al., 2016; Król, 2016; Ledoux, 1997; Theofilatos et al., 2016).

Con todas las aplicaciones posibles de las ANN es de utilidad un software que asista en el desarrollo del código de ejecución de redes neuronales. Existen diversas herramientas para la implementación de redes neuronales, en la tabla 1 se muestran algunas de ellas.

Tabla 1 Descripción de herramientas para redes neuronales.

TensorFlow	Empezó en el 2011 como un proyecto interno de Google llamado "Google Brain" y que se hizo público en el 2017 como un sistema de código abierto de aprendizaje profundo, es decir, de una red neuronal, la cual puede correr en múltiples CPUs y GPUs. Se usa para entrenar redes neuronales que puedan detectar y descifrar patrones y correlaciones análogas a las que vemos en el aprendizaje y razonamiento humano.
Caffe	Esta herramienta fue creada por BAIR (Berkeley Artificial Intelligence Research), en el 2014, y se hizo popular en la investigación académica. En un marco de trabajo de aprendizaje profundo usando redes convolutivas.
ONNX	Esta herramienta significa Open Neural Network exchanged y se anunció apenas en septiembre del 2017. Es un esfuerzo conjunto de Microsoft y Facebook. ONNX es un formato pensado para hacer fácil el intercambio de modelos de aprendizaje profundo entre entornos de esta naturaleza. La iniciativa busca hacer más fácil para los desarrolladores usar múltiples entornos de programación de redes neuronales.

Matlab	Cuenta con toolboxes especializadas para trabajar con aprendizaje automático (machine learning), redes neuronales, aprendizaje profundo, visión artificial y conducción autónoma. Puede utilizar MATLAB Coder a fin de generar código C y C++ para su red entrenada, lo cual permite simular una red entrenada en hardware de PCs y, posteriormente, desplegar la red en sistemas embebidos.
Weka	Es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato.

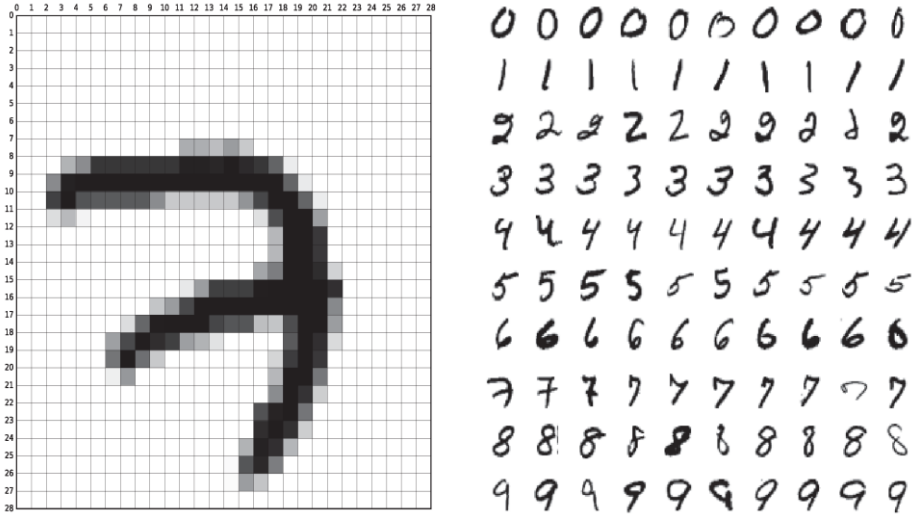
Las herramientas mencionadas sirven para intercambiar modelos de aprendizaje como ONNX, o para realizar el entrenamiento de alguna red neuronal como TensorFlow. Matlab es una herramienta interesante ya que cuenta con diferentes redes neuronales ya implementadas, el principal inconveniente es que no es de código abierto, esto además de generar un costo elevado, no se tiene acceso total a lo que las herramientas del software hacen, y si hay alguna función que no se encuentre implementada no se puede hacer nada. Es por esta razón que se diseño un software en el cual se realicen diseños de diferentes estructuras de redes neuronales de tipo convoluacional y multicapa con la finalidad de facilitar al usuario su implementación y entrenamiento.

2 Base de datos MNIST

La base de datos del MNIST (Mixed National Institute of Standards and Technology) fue presentada por LeCun (*MNIST Handwritten Digit Database, Yann LeCun, Corinna Cortes and Chris Burges*, n.d.) en 1998. Desde entonces, este conjunto de datos ha sido ampliamente utilizado como banco de pruebas para diferentes propuestas de aprendizaje automático y reconocimiento de patrones. La base de datos MNIST contiene un total de 70,000 instancias, de las cuales 60,000 son para entrenamiento y el resto para pruebas.

Las imágenes originales se sometieron a preprocesamiento. Este procedimiento involucró primero la normalización de las imágenes para que cupieran en un cuadro de 20 × 20 píxeles mientras se conservaba la relación de aspecto. Luego, se aplicó un filtro anti-aliasing y, como resultado, las imágenes en blanco y negro se transformaron efectivamente en escala de grises. Finalmente, se introdujo un relleno en blanco para ajustar las imágenes en un cuadro más grande de 28×28píxeles. Un ejemplo de una instancia correspondiente al dígito '7' se puede encontrar en la Figura 1a.

Al observar la figura 1, podemos ver que algunos dígitos pueden confundirse fácilmente. Dependiendo de cómo se escriban, puede ser difícil dilucidar si un determinado dígito es un '4' o un '9', por poner un ejemplo. Algunos casos de dígitos difíciles de reconocer, incluso para humanos, se pueden ver en la Figura 1b (ver el noveno '2' o el noveno '7').



(a) Ejemplo de instancia de digito '7'. (b) 100 muestras del juego de entrenamiento MNIST.

Figura 1 Ejemplo de la base de datos MNIST.

Debido a que MNIST ha sido ampliamente utilizado para probar el comportamiento de muchas implementaciones de clasificadores, ha habido un esfuerzo por publicar algunas clasificaciones en el pasado, utilizando MNIST como punto de referencia. La mayoría de estas clasificaciones, así como la literatura establecida, utilizan la métrica de "tasa de error de prueba" cuando se refieren al rendimiento sobre MNIST. Esta métrica es el porcentaje de instancias clasificadas incorrectamente.

3 Estructura de red neuronal

En este trabajo se desarrollan y comparan 6 estructuras diferentes de redes neuronales, tres redes neuronales convolucionales (red neuronal 1, 2 y 3) y tres redes neuronales multicapa (red neuronal 4, 5 y 6). Las tres redes convolucionales tendrán la misma configuración para la etapa totalmente conectada, dos capas ocultas, de 80 y 60 neuronas respectivamente, la función de activación utilizada es tangente hiperbólica para dichas capas, la capa de salida cuenta con 10 neuronas con sigmoide como función de activación, lo cual nos permitirá identificar los 10 diferentes números posibles de las imágenes. La etapa convolucional ira variando entre cada red neuronal, la primera utilizará un filtro para la convolución de 9×9, la segunda añadirá a la configuración previa otra capa convolucional con un filtro de 5×5, por último, la última añadirá una capa extra con un filtro de 3×3, en la figura 2 se muestra las diferencias de la etapa convolucional de cada

red, ninguna red cuenta con etapa de submuestreo ya que el tamaño de la entada es de 28×28. La operación de convolución esta ilustrada con un asterisco y w es el kernel utilizado en esta operación, b es el valor de sesgo que se suma a cada elemento de la imagen, y_p es el valor de cada elemento en las capas ocultas.

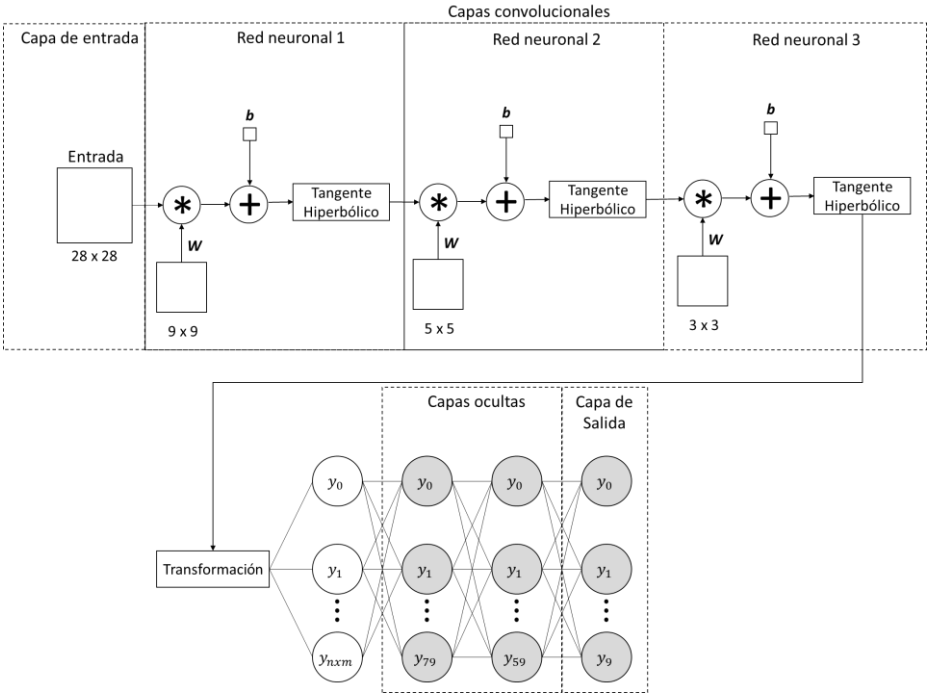


Figura 2 Estructuras redes neuronales convolucionales.

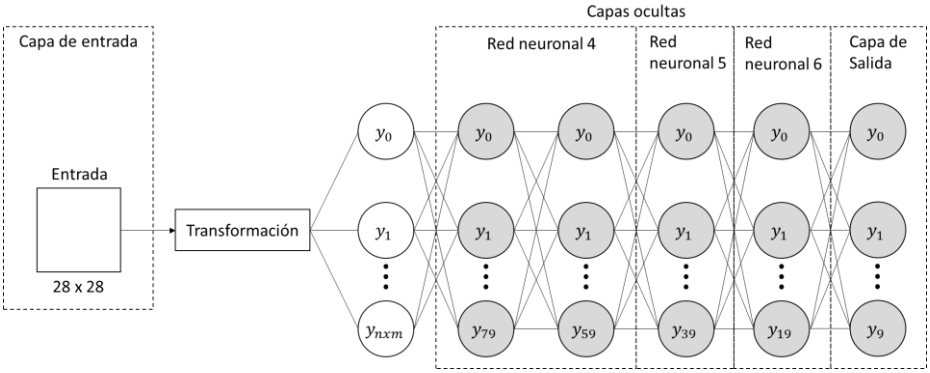


Figura 3 Estructuras redes neuronales multicapa.

En la etapa de transformación se realiza un reacomodo de la información para conectar la etapa convolucional con la completamente conectada, se toma la información en formato de matriz y se convierte en un vector, el número de entradas a las capas ocultas y_n es de $n \times m$ donde n y m son las dimensiones de salida de la última capa convolucional, en el caso de las redes multicapa se hace el reacomodo a la imagen de entrada.

Para las redes neuronales multicapa, en todas se utiliza la tangente hiperbólica como función de activación a excepción de la capa de salida que utiliza función sigmoide, la primera red tiene dos capas de 80 y 60 neuronas respectivamente, la segunda red agrega una capa extra de 40 neuronas y la tercera red agrega otra capa de 20 neuronas, en la figura 3 se pueden observar las diferencias de cada red en la etapa de capas ocultas.

4 Entrenamiento

Para entrenar las redes neuronales aplicamos el algoritmo de retro propagación (D. Rumelhart et al., 1986). El procedimiento de aprendizaje que utilizamos implica la presentación de un conjunto de pares de patrones de entrada y salida. El sistema primero utiliza el vector de entrada para producir su propio vector de salida y luego compara esto con el resultado deseado, o vector de destino. Si no hay diferencia, no hay aprendizaje. De lo contrario, los pesos se cambian para reducir la diferencia. La regla para cambiar pesos después de la comparación del par de entrada/salida (regla delta) δ viene dada por la ecuación 1.

$$\nabla_p w_{ji} = \eta\left(t_{pj} - o_{pj}\right)\ i_{pi} = \eta\delta_{pj} i_{pi} \tag{1}$$

Donde t_{pj} es la entrada de destino para el j componente del patrón de salida para el patrón p, o_{pj} es el j elemento del patrón de salida real producido por la presentación del patrón de entrada p, i_{pi} es el valor del i elemento del patrón de entrada, $\delta_{pj} = t_{pj} - o_{pj}$, y $\nabla_p w_{ji}$ es el cambio que debe hacerse al peso de la i a la j unidad siguiente del patrón p.

Para las seis redes se utiliza una tasa de aprendizaje de $\eta = 0.0001$ y los valores de pesos sinápticos, valores de los kernel y bias se inicializaron de manera aleatoria en un rango de -1 a 1, para todas las estructuras propuestas se realizan tres entrenamientos, de un entrenamiento de 100,000 iteraciones.

En cada iteración de entrenamiento se realiza una corrida de la red neuronal, se obtiene el error,

Para crear al código de ejecución y entrenamiento de las redes neuronales se utilizó el software propio.

Durante el entrenamiento se realiza la validación con una base de datos dedicada para este objetivo diferente a la de entrenamiento como el mostrado en la figura 4. Todas las redes entrenadas tuvieron un comportamiento parecido en si grafica de entrenamiento y de

validación, esto significa que la red si entreno para un caso general y no solo para la base de datos con la que se entrenó.

En la figura 4 se muestra la diferencia de la evolución del entrenamiento de las diferentes redes neuronales. La red 2 y 3 fueron las que tuvieron el peor desempeño, la red 2 tardo mucho en empezar a disminuir el error y tuvo un rebote después de las 20,000 iteraciones, pero logro recuperarse y disminuir el error. La red 3 se estancó y ya no pudo disminuir el error. Las redes 1 y 4 siguieron un comportamiento parecido en el entrenamiento y llegaron a un error parecido que les permitió tener una tasa de error debajo del 15%. La red 6 dejo de entrenar a las 30,000 iteraciones a pesar de que empezó bien el entrenamiento. La red 5 fue la que más rápido redujo el error y mantuvo ese error a partir de las 30,00 iteraciones.

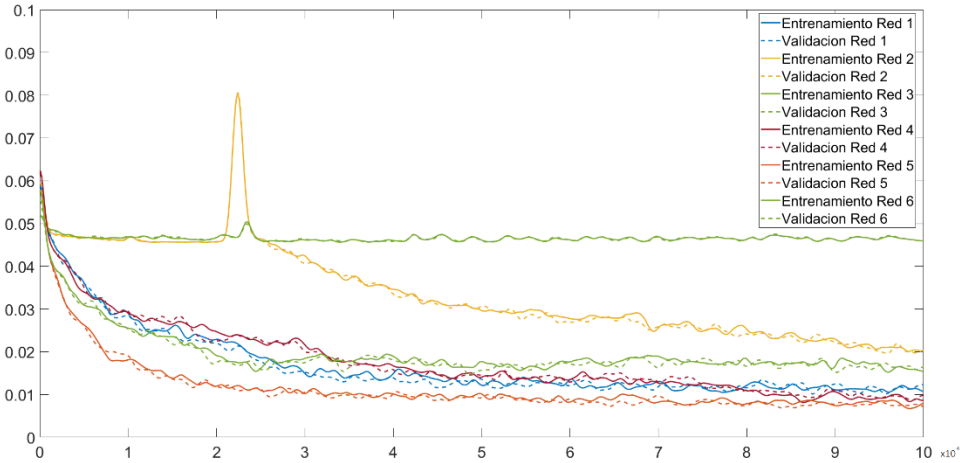


Figura 4 Resultado de entrenamiento de las redes neuronales.

5 Resultados

Se configuro cada estructura en el software y se obtuvo el código de ejecución y de entrenamiento de cada red, en la tabla 2 se muestran los resultados de las redes neuronales convolucionales y en la tabla 3 se muestra los resultados de las redes neuronales multicapa.

En la tabla 3 se muestran los resultados del entrenamiento de las redes multicapa. En esta comparación no se observa una diferencia tan grande en los tiempos de entrenamiento. El mejor desempeño lo tuvo la red 4 con tres capas ocultas.

Estos resultados concuerdan con la gráfica de entrenamiento de la figura 4, en los resultados las peores redes fueron la 2 y la 3, y las mejores fueron 4 y 5.

Tabla 2 Resultados de redes neuronales convolucionales.

Configuración de red neuronal	Numero de iteraciones de entrenamiento	Tasa de aprendizaje (Eta)	Tasa de error (%)	Tiempo de entrenamiento (S)
Red Neuronal 1	100,000	0.0001	12.12	93.23
Red Neuronal 2	100,000	0.0001	16.18	176.74
Red Neuronal 3	100,000	0.0001	16.19	249.25

Tabla 3 Resultados de redes neuronales multicapa.

Configuración de red neuronal	Numero de iteraciones de entrenamiento	Tasa de aprendizaje (Eta)	Tasa de error (%)	Tiempo de entrenamiento (S)
Red Neuronal 4	100,000	0.0001	10.66	24.68
Red Neuronal 5	100,000	0.0001	8.78	25.74
Red Neuronal 6	100,000	0.0001	19.99	26.42

6 Conclusiones

Agregar más capas convolucionales a la red no mejora el rendimiento de la red neuronal como se observa en la tabla 2, el porcentaje más alto de exactitud se obtiene con la red que tiene solo una capa convolucional.

Observando los tiempos de entrenamiento en la tabla 3 y comparándolos con las redes que tienen capas convolucionales, se concluye que las capas convolucionales agregar tiempo de ejecución considerable al entrenamiento. Comparando las 3 redes, la red con 3 capas ocultas es la que mejor rendimiento tuvo.

Comparando las redes con y sin capa convolucional, se obtuvo un mejor rendimiento con las primeras. El tamaño de las imágenes de la base de datos utilizada influye en el bajo rendimiento de las capas convolucionales ya que son de tamaño reducido y con capas ocultas basta para extraer la información.

La facilidad que aporta el software para cambiar el diseño de la red y obtener el archivo de ejecución y de entrenamiento permite hacer comparaciones entre diferentes estructuras de redes neuronales para resolver una problemática específica y elegir la opción optima.

Referencias

Akğngör, A. P., & Doğan, E. (2009). An artificial intelligent approach to traffic accident estimation: Model development and application. *Transport*, *24*(2), 135–142. https://doi.org/10.3846/1648-4142.2009.24.135-142

Chen, C. L. P., & Wang, B. (2022). Random-Positioned License Plate Recognition Using Hybrid Broad Learning System and Convolutional Networks. *IEEE Transactions on Intelligent Transportation Systems*, *23*(1), 444–456. https://doi.org/10.1109/TITS.2020.3011937

Crutchfield, J. P. (2017). The Origins of Computational Mechanics: A Brief Intellectual History and Several Clarifications. *ArXiv*, 1–9.

D. Rumelhart, G. Hinton, & R. Williams. (1986). Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol. Foundations* (1986th ed., pp. 318–362). MIT Press.

Dia, H. (2001). An object-oriented neural network approach to short-term traffic forecasting. *European Journal of Operational Research*, *131*(2), 253–261. https://doi.org/10.1016/S0377-2217(00)00125-9

Dia, H., & Rose, G. (1997). Development and evaluation of neural network freeway incident detection models using field data. *Transportation Research Part C: Emerging Technologies*, *5*(5), 313–331. https://doi.org/10.1016/S0968-090X(97)00016-8

Doğan, E., & Akgüngör, A. P. (2013). Forecasting highway casualties under the effect of railway development policy in Turkey using artificial neural networks. *Neural Computing and Applications*, *22*(5), 869–877. https://doi.org/10.1007/s00521-011-0778-0

Huang, W., Song, G., Hong, H., & Xie, K. (2014). Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, *15*(5), 2191–2201. https://doi.org/10.1109/TITS.2014.2311123

Javier, L., Villalba, G., Divina, F., Yin, L., Hong, P., Zheng, G., Chen, H., & Deng, W. (2022). A Novel Image Recognition Method Based on DenseNet and DPRN. *Applied Sciences 2022, Vol. 12, Page 4232*, *12*(9), 4232. https://doi.org/10.3390/APP12094232

Jiang, H., Zou, Y., Zhang, S., Tang, J., & Wang, Y. (2016). Short-Term Speed Prediction Using Remote Microwave Sensor Data: Machine Learning versus Statistical Model. *Mathematical Problems in Engineering*, *2016*. https://doi.org/10.1155/2016/9236156

Klügl, F., Bazzan, A. L. C., & Ossowski, S. (2010). Agents in traffic and transportation. *Transportation Research Part C: Emerging Technologies*, *18*(1), 69–70. https://doi.org/10.1016/j.trc.2009.08.002

Król, A. (2016). The Application of the Artificial Intelligence Methods for Planning of the Development of the Transportation Network. *Transportation Research Procedia*, *14*, 4532–4541. https://doi.org/10.1016/j.trpro.2016.05.376

Ledoux, C. (1997). An urban traffic flow model integrating neural networks. *Transportation Research Part C: Emerging Technologies*, *5*(5), 287–300. https://doi.org/10.1016/S0968-090X(97)00015-6

MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. (n.d.). Retrieved May 9, 2022, from http://yann.lecun.com/exdb/mnist/

Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2016). Predicting Road Accidents: A Rare-events Modeling Approach. *Transportation Research Procedia*, *14*, 3399–3405. https://doi.org/10.1016/j.trpro.2016.05.293

Wang, M., Deng, W., & Liu, C.-L. (2022). Unsupervised Structure-Texture Separation Network for Oracle Character Recognition. *IEEE Transactions on Image Processing*, *31*, 3137–3150. https://doi.org/10.1109 / TIP.2022.3165989

Wang, R., Fan, S., & Work, D. B. (2016). Efficient multiple model particle filtering for joint traffic state estimation and incident detection. *Transportation Research Part C: Emerging Technologies*, *71*, 521–537. https://doi.org/10.1016/j.trc.2016.08.003

Zhang, X., & Tao, S. (2022). Research on Pattern Recognition of Lower Limb Motion Based on Convolutional Neural Network. *Wireless Communications and Mobile Computing*, *2022*, 1–8. https://doi.org/10.1155/2022/4717413

Capítulo 3

Detección de objetos en imágenes áreas de UAV: Una revisión

Edmundo Cortes-Vazquez[1], Amparo Palomino-Merino[2]

[1] Benemérita Universidad Autónoma de Puebla. Facultad de Ciencias de la Computación.
[2] Benemérita Universidad Autónoma de Puebla. Facultad de Ciencias de la Electrónica.
edmundo.cortesvazquez1@alumno.buap.mx,
amparo.palomino@correo.buap.mx

Resumen. En el presente trabajo se hace una revisión de cómo se realiza la detección de objetos a partir de imágenes aéreas obtenidas con vehículos aéreos no tripulados. La revisión se centra en encontrar los métodos, bases de datos, métricas y procedencia de los trabajos. Se encuentra que la mayoría de trabajos desarrollan enfoques basados en deep learning.

Palabras Clave: detección de objetos, deep learning, UAV.

1 Introducción

Los vehículos aéreos no tripulados (UAV), son en la actualidad, una tecnología clave para un gran número de aplicaciones, como: vigilancia, seguridad, gestión de desastres, estacionamiento inteligente, búsqueda y rescate, flujo vehicular, entre otros. Por lo que sistemas de detección de objetos en imágenes obtenidas con estos vehículos son ampliamente estudiados y desarrollados.

Los enfoques de detección de objetos basados en características y clasificadores de aprendizaje automático han sido muy productivos de acuerdo a Felzenszwalb et al., (2009). Sin embargo, de acuerdo a Yebes et al., (2015) cuando se aplican a nuevas tareas, estos requieren de un ajuste de parámetros y una reducción de dimensionalidad para alcanzar un rendimiento razonable.

Por otro lado, las tendencias recientes en Deep Learning han logrado un rendimiento de detección impresionante en varios desafíos, incluida la detección de objetos como menciona Huang et al., (2017). Actualmente se tiene una tendencia creciente a utilizar redes neuronales convolucionales (CNN) para extraer información de imágenes y secuencias de video. Si bien se ha demostrado que las CNN son el mejor enfoque para la clasificación, detección y segmentación semántica de imágenes, las imágenes aéreas tienen muchas peculiaridades. Por ejemplo, los objetos se pueden ver desde diferentes altitudes y puntos de vista. Por lo tanto, una sola clase puede tener muchos patrones y representaciones para aprender.

2 Protocolo de la revisión

El propósito de la revisión de la literatura es analizar el progreso de los métodos de detección de objetos a partir de imágenes obtenidas por un vehículo aéreo no tripulado.

Por lo que se define los siguientes objetivos de la revisión.

a) Identificación de métodos empleados en el reconocimiento de objetos.
b) Identificación de bases de datos.
c) Identificación de métricas de evaluación.
d) Identificación de procedencia de las instituciones.

Siguiendo la metodología propuesta por Okoli y Schambra (2010), utilizando las palabras clave de la Tabla 1 en las bases de datos IEEE Xplore, ACM DL, Springer Link, Science Direct y Scopus, se obtienen miles de coincidencias.

Tabla 1. Palabras clave utilizadas.

Número	Palabras clave
1	objet detection **AND** UAV
2	detection **AND** people **AND** UAV
3	object detection **AND** people **AND** UAV
4	search **AND** rescue **AND** UAV
5	human **AND** detection **AND** UAV **AND** search **AND** rescue **AND** missions

De estos resultados, se filtran los trabajos del 2017 a 2022 (febrero), debido a que por el avance en el área, un trabajo anterior se puede considerar no actual. Se limita a publicaciones en idioma inglés y español, que traten el tema de reconocimiento de objetos mediante imágenes aéreas obtenidas de UAV, después de la aplicación de los criterios de exclusión se tienen aún más de 100 trabajos.

Puntuando los trabajos de acuerdo a los criterios de calidad, que se muestran en la Tabla 2 si el puntaje del trabajo alcanza los 9 puntos se incluye en la revisión, de acuerdo a esto se obtienen 33 trabajos para realizar la revisión.

Tabla 2. Criterios de calidad utilizados.

Ponderación	Criterio de calidad	Puntaje	
		1	**2**
3	Factor de impacto de la revista	≤ 1	> 1
2	Numero de Citas	≤10	>10
1	Disponibilidad de datos	NO	SI

3 Extracción de datos

De los artículos que superaron el paso anterior, se extrae la información de acuerdo a los objetivos de la revisión, en las Tablas 3-6.

Tabla 3. Extracción de la información 1 de 4.

No.	Autores	Contribución	Método de detección	Criterio de evaluación	Base de Datos
1	Tian, G., Liu, J., Yang, W. (2021)	Detección de objetos pequeños	Red neuronal dual	mAP FPS	VisDrone, UAVDT, MS COCO
2	Mittal, P., Sharma, A., Singh, R. (2020)	Revisión de los algoritmos de detección de objetos basados en el aprendizaje profundo	Cascade R-CNN, Faster R-CNN, YOLO, SSD, RetinaNet	mAP	VisDrone Okutama- Action MS-COCO
3	Srivastava, S., Narayan, S., Mittal, S. (2020)	Revisión de técnicas de aprendizaje profundo para detectar vehículos	VGG16, ResNet Faster R-CNN, YOLO, SSD	AP, PR RR, F1 score FPS	DLR 3K, VEDAI, ISPRS, UAVDT, PUCPR+
4	Tian, G., Liu, J., Zhao, H., Yang, W. (2021)	Mejora en detección de objetos pequeños partiendo de detectores de una sola etapa	1: YOLO 2: DSAE y KNN	mAP	VisDrone, UAVDT, MS COCO
5	Martinez-Alpiste, I., Golcarenarenji, G., Wang, Q., Alcaraz-Calero, J. (2021)	Detección de objetos basado en aprendizaje automático en tiempo real integrado en un teléfono.	CNN	mAP FPS	Autores MS COCO
6	Mliki, H., Bouhlel, F., Hammami, M. (2020)	Un enfoque para el reconocimiento de actividades humanas adaptando CNN	CNN	recall precision	UCF-ARG
7	Mandal, M., Kumar, L. K., & Vipparthi, S. K. (2020).	Una base de datos para clasificación de objetos en movimiento		AP FPS	MOR-UAV
8	Shi, Y., Li, H., Ding, W., & Liu, S. (2019)	Reconocimiento de escenas mediante segmentación adaptiva de superpixeles.	Super pixel segmentación Clasificador SMV	recognition rate	Autores

Tabla 4. Extracción de la información 2 de 4.

No.	Autores	Contribución	Método de detección	Criterio de evaluación	Base de Datos
9	Zhan, W., Sun, C., Wang, M., She, J., Zhang, Y., Zhang, Z., & Sun, Y. (2022)	Propone mejoras a YOLO para mejorar la detección de objetos pequeños	Basado en YOLO	mAP FPS	VisDrone
10	Wang, X., Cheng, P., Liu, X., & Uzochukwu, B. (2018)	Realiza una comparación de detectores de una y dos etapas	Faster R-CNN SSD RetinaNet	mAP	Stanford Drone
11	Zeng, Y., Duan, Q., Chen, X., Peng, D., Mao, Y., & Yang, K. (2021)	Una base de datos para la detección de UAV	R-CNN YOLO V3 SDD	mAP FPS	UAVData
12	Liu, M., Wang, X., Zhou, A., Fu, X., Ma, Y., & Piao, C. (2020)	Desarrolla un método de detección de objetos pequeños	Basado en YOLO V3	mAP	Autores UAV123
13	Wu, Q., Zhou, Y., Wu, X., Liang, G., Ou, Y., & Sun, T. (2020)	Propone un método de detección humana de ejecución rápida para UAV, basado en flujo óptico y CNN	CNN Flujo óptico	mAP FPS	Autores
14	Sambolek, S., & Ivasic-Kos, M. (2021)	Una versión mejorada basada en YOLOv4 para detectar personas en misiones SAR	Basado en YOLO V4	mAP	VisDrone
15	Tan, L., Lv, X., Lian, X., & Wang, G. (2021)	Versión basada en YOLOv4 para detectar objetos en un fondo complejo	Basado en YOLO V4	mAP FPS	VisDrone
16	Al-Kaff, A., Gómez-Silva, M. J., Moreno, F. M., De La Escalera, A., & Armingol, J. M. (2019)	Propone detectar a las personas basados en los datos de color y profundidad con un sensor RGB-D	Seguimiento de objetos multiples	MOT	ARMOT
17	Kundid Vasić, M., & Papić, V. (2020)	Combina dos arquitecturas de CNN diferentes en la etapa de propuesta de región y clasificación	CNN	recall precision	HERIDAL

Tabla 5. Extracción de la información 3 de 4.

No.	Autores	Contribución	Método de detección	Criterio de evaluación	Base de Datos
18	Božić-Štulić, D., Marušić, Ž., & Gotovac, S. (2019)	Método para la detección de personas en imágenes aéreas en paisajes mediterráneos.	CNN	recognition rate precision	HERIDAL
19	Vaddi, S. (2019)	Modelo de detección de objetos de extremo a extremo que se ejecuta en una plataforma UAV en tiempo real	ResNet MobileNet	mAP FPS	VisDrone
20	Stojnić, V., Risojević, V., Muštra, M., Jovanović, V., Filipi, J., Kezić, N., & Babić, Z. (2021)	Método para la detección de objetos pequeños en movimiento, mediante segmentación con CNN	CNN	recall precision F1	Autores
21	Hendria, W., Phan, Q., Adzaka, F., & Jeong, C. (2021).	Una fusión de modelos basados en transformadores y basados en redes neuronales convolucionales.	Swin Transformer CNN	AP	VisDrone
22	Dousai, N., & Lončarić, S. (2022)	Un modelo basado en CNN para la detección de personas en imágenes aéreas en ambientes montañosos	CNN EfficientDET	mAP	HERIDAL
23	Boudjit, K., & Ramzan, N. (2021)	Un método para la identificación y detección de personas utiliando CNN, basado en YOLO v2.	CNN YOLO v2	mAP IoU FPS	Autores
24	Liu, C., & Szirányi, T. (2021)	Un método de detección de personas y recocimiento de gestos en tiempo real basados en YOLO v3	Basado en YOLO v3	recall precision	MS COCO
25	Kim, J., & Cho, J. (2021).	Detectar objetos en imágenes aéreas, donde las formas y el entorno dinámico cambia constantemente	RGDiNet Faster R-CNN	AP IoU	SCHF

Tabla 6. Extracción de la información 4 de 4.

No.	Autores	Contribución	Método de detección	Criterio de evaluación	Base de Datos
26	Lee, J., Wang, J., Crandall, D., Šabanović, S., & Fox, G. (2017).	Detección de objetos con el computo en la nube, mientras que la detección de bajo nivel y navegación se realiza a bordo	CNN YOLO SSD	mAP FPS	Pascal VOC
27	Hoai, D. K., & Van Phuong, N. (2017)	Detección de anomalías en diferentes espacios de color transformados a partir de imágenes de UAV	Reed-Xiaoli	AUC: Área bajo la curva ROC	Autores
28	Rohan, A., Rabah, M., & Kim, S. H. (2019)	Un enfoque para detectar y rastrear el objeto objetivo. La CNN detecta objetos en tiempo real	CNN SSD	precisión FPS	Autores
29	Nousi, P., Mademlis, I., Karakostas, I., Tefas, A., & Pitas, I. (2019)	Revisión del estado del arte de la detección y seguimiento de objetos en tiempo real	CNN YOLO SDD Faster R-CNN	recall FPS	Cyclist Detection
30	Kashihara, S., Wicaksono, M. A., Fall, D., & Niswar, M. (2019)	Un algoritmo de detección de objetos	YOLO	AP FPS	Autores
31	Valappil, N., & Memon, Q. (2021)	Utiliza flujo óptico y bibliotecas de deep learning de Matlab para detectar objetos	CNN SVM	precisión	Autores
32	Zhang, H., Sun, M., Li, Q., Liu, L., Liu, M., & Ji, Y. (2021)	Propone una arquitectura CNN capaz de detectar vehículos de imágenes aéreas	CNN	precisión IoU FPS	Autores
33	Kyrkou, C., Plastiras, G., Theocharides, T., Venieris, S., & Bouganis, C. (2018)	Realiza un conjunto de datos para detección de objetos en entornos suburbanos y evalúa el rendimiento de métodos actuales.	CNN YOLO SDD Faster R-CNN R-FCN	AP	MORH

4 Síntesis de la información

La mayoría de los trabajos revisados utilizan deep learning para el reconocimiento de objetos, utilizando algoritmos de detección de objetos de una etapa (YOLO, SDD, RetinaNet), de dos etapas (RCNN, Faster RCNN, Cascade RCNN, R-FCN) y detectores avanzados (CornetNet). Pero también existen trabajos que presentan enfoques clásicos como los basados en visión por computadora, máquinas de vectores de soporte y otros clasificadores.

Las métricas más utilizadas en estos trabajos para comparar el rendimiento en el reconocimiento de objetos es la mAP (precisión media promedio), AP (precisión promedio) y los FPS (Fotogramas por segundo). Aunque, para algunos trabajos también se menciona que es importante contemplar el costo computacional y energético dependiendo el campo de aplicación para alcanzar un equilibrio entre rendimiento, costo y funcionalidad.

En cuanto a las bases de datos utilizadas para el entrenamiento de los algoritmos basados en deep learning, los autores mencionan una falta de colecciones desde la perspectiva del UAV que les sea de interés y funcionalidad para su aplicación, por lo que los autores realizan su propia base de datos, como se observa en la figura 1, son 12 los trabajos que lo realizan. El problema con estas bases de datos es que no son publicadas, lo que dificulta la comparación del desempeño de sus algoritmos de detección de objetos.

Después de las colecciones propias de los autores las bases de datos más utilizadas para comparar el rendimiento de los métodos de detección de objetos son VisDrone y MS COCO.

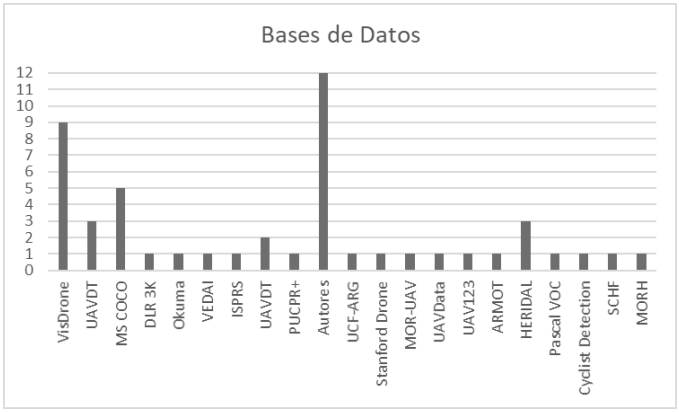


Fig. 1. Bases de datos de imágenes de UAV.

La procedencia de la mayoría de los trabajos revisados es de instituciones en China, seguido por Croacia, Estados Unidos y Corea, como se puede observar en la figura 2.

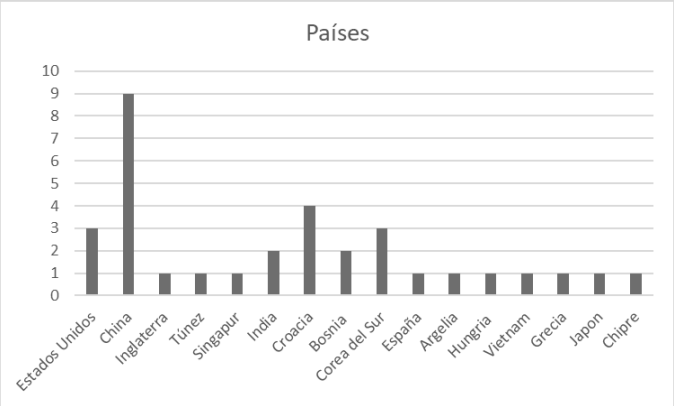


Fig. 2. Países de procedencia de las instituciones del autor principal.

5 Conclusiones

En esta revisión se encuentra que existen diferentes métodos para la detección de objetos desarrollados por inteligencia artificial, de los cuales la mayoría están basados en deep learning. Esta revisión se centra en los trabajos publicados entre 2017 y 2022, ya que se consideran los resultados más recientes, sin demeritar publicaciones anteriores, en las cuales se establecen las bases de la detección.

Para la selección e implementación de estos métodos es importante tener en cuenta las limitaciones de recursos disponibles (computacionales, arquitectura, etc.). Se identifica que los detectores de una etapa basados en YOLO son muy recurridos cuando se quiere utilizar dispositivos de bajo consumo y de respuesta en tiempo real.

También se encuentra un área de oportunidad para la creación de una base de datos para el reconocimiento de personas en imágenes aéreas obtenidas por UAV en entornos montañosos, que sirva para aplicaciones de búsqueda y rescate.

Referencias

Al-Kaff, A., Gómez-Silva, M. J., Moreno, F. M., De La Escalera, A., y Armingol, J. M. (2019). "An appearance-based tracking algorithm for aerial search and rescue purposes". *Sensors*, 19(3), pp. 652-681.

Boudjit, K., y Ramzan, N. (2021). "Human detection based on deep learning YOLO-v2 for real-time UAV applications". *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1-18.

Božić-Štulić, D., Marušić, Ž., y Gotovac, S. (2019). “Deep learning approach in aerial imagery for supporting land search and rescue missions”. *International Journal of Computer Vision*, 127(9), pp. 1256-1278.

Dousai, N., y Lončarić, S. (2022). “Detecting Humans in Search and Rescue Operations Based on Ensemble Learning”. *IEEE Access*, 10, pp. 26481-26492.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., y Ramanan, D. (2009). “Object detection with discriminatively trained part-based models”. *IEEE transactions on pattern analysis and machine intelligence*, 32(9) pp.1627-1645.

Hendria, W., Phan, Q., Adzaka, F., y Jeong, C. (2021). “Combining transformer and CNN for object detection in UAV imagery”. *ICT Express*.

Hoai, D. K., y Van Phuong, N. (2017). “Anomaly color detection on uav images for search and rescue works”. *In 2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 287-291.

Kashihara, S., Wicaksono, M. A., Fall, D., y Niswar, M. (2019). “Supportive Information to Find Victims from Aerial Video in Search and Rescue Operation”*, 2019 IEEE International Conference on Internet of Things and Intelligence System.* pp. 56-61.

Kim, J., y Cho, J. (2021). “RGDiNet: Efficient Onboard Object Detection with Faster R-CNN for Air-to-Ground Surveillance”. *Sensors*, 21(5), pp. 1677-1693.

Kundid, M., y Papić, V. (2020). “Multimodel deep learning for person detection in aerial images”. *Electronics*, 9(9), pp. 1459-1473.

Kyrkou, C., Plastiras, G., Theocharides, T., Venieris, S. I., y Bouganis, C. S. (2018). “DroNet: Efficient convolutional neural network detector for real-time UAV applications”. *In 2018 Design, Automation & Test in Europe Conference & Exhibition* (DATE), pp. 967-972.

Lee, J., Wang, J., Crandall, D., Šabanović, S., y Fox, G. (2017). “Real-time, cloud-based object detection for unmanned aerial vehicles”. *In 2017 First IEEE International Conference on Robotic Computing (IRC)*, pp. 36-43.

Liu, C., & Szirányi, T. (2021). “Real-time human detection and gesture recognition for on-board UAV rescue”. *Sensors*, 21(6), pp. 2180-2200.

Mandal, M., Kumar, L. K., y Vipparthi, S. K. (2020). “MOR-UAV: A benchmark dataset and baselines for moving object recognition in UAV videos”. *In Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2626-2635.

Martinez-Alpiste, I., Golcarenarenji, G., Wang, Q., y Alcaraz-Calero, J. M. (2021). “Search and rescue operation using UAVs: a case study”. *Expert Systems with Applications*, 178.

Mittal, P., Singh, R., y Sharma, A. (2020). “Deep learning-based object detection in low-altitude UAV datasets: A survey”. *Image and Vision computing*, 104.

Mliki, H., Bouhlel, F., y Hammami, M. (2020). “Human activity recognition from UAV-captured video sequences”. *Pattern Recognition*, 100.

Nousi, P., Mademlis, I., Karakostas, I., Tefas, A., y Pitas, I. (2019). “Embedded UAV real-time visual object detection and tracking”. *In 2019 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pp. 708-713.

Okoli, C., Schabram, K. (2010). “A guide to conducting a systematic literature review of information systems research”.

Rohan, A., Rabah, M., y Kim, S. H. (2019). "Convolutional neural network-based real-time object detection and tracking for parrot AR drone 2". *IEEE access*, 7, pp. 69575-69584.
Sambolek, S., y Ivasic-Kos, M. (2021). "Automatic person detection in search and rescue operations using deep CNN detectors". *IEEE Access*, 9.
Shi, Y., Li, H., Ding, W., y Liu, S. (2017). "Scene recognition for complicated UAV images based on land surface classification of superpixel". *In Proceedings of the 2017 International Conference on Artificial Intelligence, Automation and Control Technologies*, pp. 1-7.
Srivastava, S., Narayan, S., y Mittal, S. (2021). "A survey of deep learning techniques for vehicle detection from UAV images". *Journal of Systems Architecture*, 117.
Stojnić, V., Risojević, V., Muštra, M., Jovanović, V., Filipi, J., Kezić, N., y Babić, Z. (2021). "A method for detection of small moving objects in UAV videos". *Remote Sensing*, 13(4), pp. 653-672.
Tan, L., Lv, X., Lian, X., y Wang, G. (2021). "YOLOv4_Drone: UAV image target detection based on an improved YOLOv4 algorithm". *Computers & Electrical Engineering*, 93.
Tian, G., Liu, J., y Yang, W. (2021a). "A dual neural network for object detection in UAV images". *Neurocomputing*, 443, pp. 292-301.
Tian, G., Liu, J., Zhao, H., y Yang, W. (2021b). "Small object detection via dual inspection mechanism for UAV visual images". *Applied Intelligence*, 52(4), pp. 4244-4257.
Vaddi, S. (2019). "Efficient object detection model for real-time UAV applications". *Doctoral dissertation*, Iowa State University.
Valappil, N. K., y Memon, Q. A. (2021). "CNN-SVM based vehicle detection for UAV platform". *International Journal of Hybrid Intelligent Systems*, pp.1-12.
Wang, X., Cheng, P., Liu, X., y Uzochukwu, B. (2018). "Fast and accurate, convolutional neural network based approach for object detection from UAV". *In IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, IEEE, pp. 3171-3175.
Wu, Q., Zhou, Y., Wu, X., Liang, G., Ou, Y., y Sun, T. (2020). "Real-time running detection system for UAV imagery based on optical flow and deep convolutional networks". *IET Intelligent Transport Systems*, 14(5), pp. 278-287.
Yebes, J. J., Bergasa, L. M., y García-Garrido, M. (2015). "Visual object recognition with 3D-aware features in KITTI urban scenes". *Sensors*, 15(4), pp. 9228-9250.
Zeng, Y., Duan, Q., Chen, X., Peng, D., Mao, Y., y Yang, K. (2021). "UAVData: A dataset for unmanned aerial vehicle detection". *Soft Computing*, 25(7), pp. 5385-5393.
Zhan, W., Sun, C., Wang, M., She, J., Zhang, Y., Zhang, Z., y Sun, Y. (2022). "An improved Yolov5 real-time detection method for small objects captured by UAV". *Soft Computing*, 26, pp. 361-373.
Zhang, H., Sun, M., Li, Q., Liu, L., Liu, M., y Ji, Y. (2021). "An empirical study of multi-scale object detection in high resolution UAV images". *Neurocomputing*, 421, pp. 173-182.

Capítulo 4

Corrección Atmosférica de Imágenes Satelitales para la Teledetección: Descripción general y Técnicas

Arturo Jasso Garduño, Ignacio Muñoz Máximo, David Pinto

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación
arturo.jasogarduno@viep.com.mx, ignacio.munozmax@correo.buap.mx, dpinto@cs.buap.mx

Resumen. La teledetección proporciona información valiosa sobre objetos o áreas a distancia, ya sea en modo activo (p. ej., RADAR y LiDAR) o pasivo (p. ej., multiespectral e hiperespectral) mediante un satélite. La calidad de los datos adquiridos por los sensores de imágenes de detección remota (tanto activos como pasivos) a menudo se ve degradada por una variedad de elementos en especial la atmósfera terrestre. La corrección atmosférica de imágenes satelitales, que es un campo de investigación activo en la comunidad de sensores remotos, es la tarea de recuperar la verdadera imagen desconocida de la imagen degradada observada debido a los efectos de la atmósfera. Cada imagen capturada por los sensores del satélite es afectada por la atmósfera de manera que es necesario compensar o eliminar los efectos de ésta para poder hacer posteriormente un análisis de la imagen en cuestión. Este artículo de revisión reúne los avances de las técnicas de corrección atmosférica de imágenes satelitales con un enfoque particular en las técnicas basadas en imágenes (Image Based) y las técnicas basadas en áreas geográficas que contienen Características Pseudo-Invariantes. Por lo tanto, se intenta proporcionar un punto de partida integral y específico de la disciplina para investigadores de diferentes niveles (es decir, estudiantes, investigadores e investigadores senior) que deseen investigar el importante tema de la corrección atmosférica de imágenes satelitales.

Palabras Clave: Corrección atmosférica, Transferencia Radiativa, Objetos Obscuros, Áreas y Características Pseudo-Invariantes.

1. Introducción

Un sistema de teledetección óptica se puede dividir en cinco subsistemas (Liang 2004, Fig. 1): el modelo de transferencia radiativa de la escena, el modelo de transferencia radiativa atmosférica, el sistema de navegación, el sistema de sensores y el sistema de cartografía y agrupación. El modelo de transferencia radiativa de la escena describe la relación entre las señales radiativas superficiales y los parámetros característicos de la superficie. El modelo de transferencia radiativa atmosférica caracteriza los impactos atmosféricos en las señales radiativas superficiales que reciben los sensores remotos. El sistema de navegación involucra principalmente los sistemas de imágenes de superficie de satélites transportados, aeronaves y otras plataformas de sensores. El sis-

tema de sensores incluye principalmente la respuesta espectral del sensor, la respuesta espacial, la división de bandas, el tratamiento del ruido y el procesamiento digital, y el sistema de mapeo y agrupación implica principalmente la transformación de proyección y el muestreo de imágenes.

El modelo de transferencia de radiación atmosférica es un enlace crítico que conecta los parámetros característicos de la superficie con las señales recibidas por los sensores remotos. Este modelo también se considera un factor importante que influye en la recuperación cuantitativa de los parámetros característicos de la superficie utilizando imágenes de teledetección.

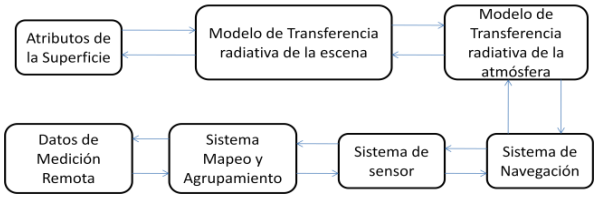


Fig. 1: Modelos de sistemas de teledetección óptica

La cantidad medida por un espectrómetro de imágenes, que es típicamente radiación solar, está sujeta a efectos atmosféricos como la absorción y la dispersión. Por lo tanto, para estudiar con precisión las propiedades de la superficie, estos efectos atmosféricos deben tenerse en cuenta y corregirse, en un proceso llamado corrección atmosférica. El proceso en el que se hace esto ha evolucionado con el tiempo desde métodos empíricos basados en líneas hasta enfoques más recientes, como el modelado utilizando la ecuación de transferencia radiativa. Los métodos para corregir los efectos de la atmósfera en la percepción remota se pueden dividir en dos categorías (John R. Jensen 2015):

- Correcciones Absolutas
- Correcciones Relativas

1.1. Contribución

En este documento, proporcionamos una descripción general de las técnicas de corrección atmosférica de vanguardia propuestas para los dos tipos más importantes de corrección en los sistemas detección remota. Los enfoques de corrección atmosférica considerados en este documento incluyen los principales paradigmas en la eliminación de los efectos atmosféricos, es decir, técnicas absolutas basadas en la ecuación de transferencia radiativa y las basadas en la propia imagen así como las técnicas relativas basadas principalmente en áreas geográficas con características pseudo-invariantes conocidas como PIF o alternativamente conocidas con el nombre de áreas pseudo-invariantes PIA.

2. Técnicas de Corrección Atmosférica Absolutas

Esta categoría incluye correcciones que requieren condiciones ópticas atmosféricas (por ejemplo, profundidad óptica de aerosoles) como parámetros de entrada y conducen a la reflectancia de la superficie y correcciones que dependen exclusivamente de la imagen obtenida por el sensor satelital, sin embargo al obtenerse la reflectancia aproximada de la superficie terrestre hablamos de igual manera de correcciones absolutas. Esta categoría de Correcciones Absolutas se puede subdividir en dos subcategorías: correcciones atmosféricas basadas en imágenes (por ejemplo, el píxel más oscuro, método de matriz de covarianza) y correcciones que utilizan datos independientes para las condiciones ópticas atmosféricas (incluidas las mediciones in situ o el registro histórico) de los que existen varios algoritmos que hacen uso de software que resuelve la ecuación de Transferencia Radiativa RTE (Hadjimitsis 2004).

2.1.1. Software de Modelado de la RTE

Los modelos de transferencia de radiación atmosférica (RTM) son herramientas de software que ayudan a los investigadores a comprender los procesos de radiación que ocurren en la atmósfera de la Tierra (Liang 2012, Jensen 2014). Dada su importancia en las aplicaciones de teledetección, la comparación de RTM atmosféricos es, por lo tanto, una de las principales tareas para evaluar el rendimiento del modelo e identificar las características que difieren entre los modelos. Algunos de los modelos más utilizados son 6S, MODTRAN y libRadtran. Adicionalmente se ha utilizado o se puede utilizar la plataforma AERONET que también provee información que se puede utilizar para hacer corrección atmosférica.

2.1.2. MODTRAN

MODTRAN (Moderate-Resolution Atmospheric Transmittance and Radiance Code) es el software utilizado para calcular la transmitancia atmosférica y el algoritmo de transferencia radiativa con una resolución espectral moderada. MODTRAN se puede utilizar para calcular la transmitancia atmosférica, la radiación atmosférica de fondo (radiación atmosférica ascendente y descendente), la radiancia de la dispersión solar o lunar única, la irradiancia solar directa, etc. MODTRAN se basó en LOWTRAN, con una resolución espectral de 1 cm^{-1}. Se han actualizado varias bases de datos importantes en MODTRAN 5.0, lo que mejora en gran medida la precisión del cálculo. También se mejoran las precisiones de la dispersión de Rayleigh y el índice de refracción complejo, y se agregan en DISORT las opciones relacionadas para calcular el azimut de la contribución de la dispersión solar.

2.1.3. 6S

El Modelo 6S (Second Simulation of the Satellite Signal in the Solar Spectrum) se compila con el lenguaje de programación FORTRAN y se aplica al cálculo de simulación del modelo de transferencia radiativa atmosférica en la banda de reflexión solar (0.25-4 μm). En este modelo se utilizan los últimos algoritmos de aproximación y aproximación sucesiva para calcular la dispersión y la absorción, en los que se mejora la entrada de parámetros para proporcionar una simulación más precisa. Basado en la

suposición de un cielo sin nubes, en este modelo se consideran muchos problemas, incluida la absorción por vapor de agua, dióxido de carbono, ozono y oxígeno; dispersión molecular y de aerosoles; y reflectancia bidireccional superficial no uniforme. Entre estas consideraciones, la absorción de gas se calcula con un intervalo espectral de 10 cm^{-1} y un paso integral espectral de 2,5 nm; por lo tanto, este método se usa principalmente para procesar sensores remotos multiangulares.

2.1.4. Algoritmos que hacen uso del Modelado de la Transferencia Radiativa (RTM)

A continuación se describen brevemente los algoritmos de corrección atmosférica encontrados en la literatura (Hadjimitsis 2004, Vicent 2019, López Serrano 2016).

2.1.5. ATCOR

ATCOR es un módulo de ERDAS IMAGINE para la corrección atmosférica y la eliminación de la neblina, que se aplica a la corrección de los impactos atmosféricos en la reflectancia espectral de las superficies y la eliminación de los impactos de las nubes delgadas y la neblina. Este módulo se puede utilizar para corregir imágenes planas en el área de la imagen e imágenes con cambios de altura mayores, lo que requiere un modelo de elevación digital (DEM) del área de la imagen.

2.1.6. ACORN

Atmospheric CORRection Now se basa en el código de transferencia radiativa MODTRAN 4. Realiza la corrección atmosférica de imágenes multiespectrales e hiperespectrales en la región desde 250 ± 2500 nm. Está diseñado para funcionar con todos los sistemas de teledetección calibrados aerotransportados (drones) y espaciales (satélites), como Hyperion, ASTER, Landsat ETM+, AV IRIS, SPO T, GeoEye-1, etc. Define la relación a partir de las contribuciones de la fuente solar exoatmosférica, una atmósfera paralela en un plano homogéneo, y la superficie con respecto a la radiación medida por un sensor remoto con vista a la Tierra (ImSpec ACO RN, 2014).

2.1.7. FLAASH

FLAASH (Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes) es un programa de corrección atmosférica que corrige datos de sensores remotos en la región desde 400-3000 nm. FLAASH se puede utilizar para corregir atmosféricamente datos de sensores hiperespectrales como HyMAP, AVIRIS, HYDICE, HYPERION, Probe-1, CASI y AISA y sensores multiespectrales como ASTER, IRS, Landsat, RapidEye y PUNTO. FLAASH utiliza simulaciones MODTRAN de radiación espectral calculadas para varias condiciones atmosféricas, de vapor de agua y de visualización (ángulos solares) en un rango de reflectancias de superficie para establecer tablas de búsqueda para los parámetros atmosféricos de columna de vapor de agua, tipo de aerosol y visibilidad para su uso posterior.

2.1.8. Corrección atmosférica absoluta usando Calibración de línea empírica

Para usar la calibración de línea empírica, el analista de imágenes generalmente selecciona dos o más áreas en la escena con diferentes albedos (por ejemplo, un objetivo brillante, como arena, y un objetivo oscuro, como un cuerpo de agua profundo y no turbio). Las áreas seleccionadas deben ser lo más homogéneas posible. Las mediciones de espectro radiómetro in situ de estos objetivos se realizan en el suelo. Los espectros in situ y derivados de la detección remota se someten a regresión y se calculan los valores de ganancia y compensación. Luego, los valores de ganancia y compensación se aplican a los datos del sensor remoto banda por banda, eliminando la atenuación atmosférica. Hay que tener en cuenta que la corrección se aplica banda por banda y no píxel por píxel, como ocurre con ACORN, FLAASH y ATCOR (Hadjimitsis 2009, Jensen 2014).

2.1.9. ACOLITE

El procesador ACOLITE es una aplicación independiente para la corrección atmosférica de Landsat (5, 7 y 8) y Sentinel-2 (A/B). Este método se utiliza principalmente para la corrección atmosférica sobre cuerpos de agua. Sin embargo, también se puede utilizar satisfactoriamente para la corrección atmosférica sobre tierra. El proceso de corrección consta de dos pasos: primero, la corrección de Rayleigh se realiza utilizando una tabla de búsqueda generada por la corrección atmosférica 6SV. En segundo lugar, la corrección de aerosoles se realiza en base a la suposición de señal cero en las bandas SWIR sobre cuerpos de agua (Valdivieso-Ros 2021, Moravec 2021).

2.1.10. SEN2COR

Sen2Cor es un algoritmo desarrollado por la ESA específicamente para los satélites Sentinel 2, y el código está disponible gratuitamente como complemento SNAP (Sentinel Application Platform). Los usuarios también pueden obtener productos L2A corregidos con el algoritmo Sen2Cor descargándolos directamente desde Copernicus Open Access Hub. El algoritmo realiza una detección de nubes y una clasificación de escenas, seguida de una recuperación de aerosoles y vapor de agua de las imágenes L1C. Los valores de reflectividad BOA (Bottom Of Atmosphere) se obtienen luego usando esos valores (Valdivieso-Ros 2021, Moravec 2021, Rumora 2020).

2.1.11. Otros Algoritmos

Debido a razones de espacio, solo se mencionan otros algoritmos encontrados en la literatura entre los cuales tenemos S2AC, iCOR, SREF, STDSREF. Se sugiere consultar la bibliografía mostrada para más detalles (Moravec 2021).

2.1.12. Deep Learning

La solución de aprendizaje profundo propuesta consta de dos pasos principales. El primero es el entrenamiento de la red, que generalmente es una tarea costosa desde el punto de vista computacional y que requiere mucho tiempo, y la creación de instancias de la solución (inferencia), que es un proceso muy rápido. Las redes de aprendizaje profundo requieren grandes cantidades de datos para construir internamente una

representación conceptual de las relaciones entre la entrada y la salida y, por lo tanto, es más práctico generalmente usar datos simulados versus observados para generar el gran conjunto de datos requerido. Los datos sintéticos se generan, por ejemplo, usando el software MODTRAN para simular la radiación total, el afloramiento y la transmisión atmosférica para diferentes ángulos geométricos, modelos atmosféricos, hora del día y año, y objetivos específicos con reflectividades variables (Xu 2020). Se supone que los datos se recopilan en diferentes ángulos utilizando un sensor aerotransportado un satélite usualmente.

2.2. Técnicas de Corrección Atmosférica Basado en Imágenes (Image Based)

Los métodos basados en imágenes utilizan datos derivados de análisis estadísticos de los datos de píxeles sin procesar, en el método más simple (píxel más oscuro), simplemente restando el DN (Digital Number) mínimo en cada banda de todos los demás en esa banda.

2.2.1. QUAC

QUAC es una corrección atmosférica que requiere solo una especificación aproximada de las ubicaciones de la banda del sensor. Utiliza un enfoque de escena y, por lo tanto, es más rápido que las correcciones con modelos radiativos. El principio QUAC asume que la curva espectral promedio de varios (>10, típicamente 50) materiales diversos de una imagen debe tener la misma firma espectral que la firma "universal" pre-calculada derivada del promedio de diversas colecciones de espectros de reflectancia de la biblioteca espectral. Si hay una diferencia entre el espectro promedio de la biblioteca y el promedio del espectro observado de los miembros finales, representa un efecto de la atmósfera.

2.2.2. DOS (Dark Object Substraction)

La sustracción de objetos oscuros (DOS) es un método ampliamente utilizado para reducir la neblina (haze) dentro de una imagen y se realiza para cada banda individualmente. Se supone que hay píxeles dentro de cada banda de una imagen multiespectral que tienen una reflectancia en el suelo muy baja o nula, y que la diferencia entre el valor de brillo de estos píxeles y cero se debe a la neblina. Esta diferencia estimada por banda se resta de cada banda de la imagen. La mayoría de las técnicas de sustracción de objetos oscuros asumen que existe una alta probabilidad de que haya al menos unos pocos píxeles dentro de una imagen que deberían ser negros (0% de reflectancia).

2.2.3. COST (Cosine of the Sun Zenith Angle)

Se trata de un método de calibración radiométrica que considera el efecto atmosférico y se basa íntegramente en las características de la imagen de satélite, a diferencia de otros métodos de corrección atmosférica, como ATCOR2, FLAASH o 6S, que requieren algunos parámetros extra, como perfiles atmosféricos, los modelos de los aerosoles o visibilidad (López-Serrano 2016).

2.2.4. Eliminación de ruido por Umbralización de Wavelet

Los efectos de la atmósfera se pueden modelar como ruido aditivo (scattering) mientras que la absorción de la atmósfera tiene un afecto multiplicativo. Debido a esto, algunos autores han propuesto utilizar los algoritmos para "denoising" por ejemplo usando umbralización de Wavelet (Bhosle 2011). Puede ser adecuado si el objetivo es la clasificación de áreas urbanas pero no para estudios de análisis de vegetación. La regresión Wavelet da mejores resultados en términos de corrección atmosférica y análisis de vegetación. Sin embargo, este método solo corrige el efecto de dispersión atmosférica. Es simple y fácil de implementar. Este método se puede evaluar aún más mediante la clasificación de imágenes y los algoritmos de detección de cambios.

3. Técnicas de Corrección Atmosférica Relativas

En los últimos años se han desarrollado otras técnicas de corrección radiométrica relativa. Relativa se refiere a que se realiza una normalización respecto de una imagen de referencia. La corrección radiométrica relativa se puede usar para normalizar las intensidades entre las diferentes bandas dentro de una imagen de detección remota de una sola fecha, y para normalizar las intensidades de las bandas de datos de sensores remotos en múltiples fechas de imágenes a una escena estándar seleccionada como la referencia.

3.1. Normalización de una sola imagen usando el ajuste de histograma "Histogram Matching"

La normalización del histograma también conocido como "Histogram Matching" es una técnica para distribuir las frecuencias del histograma en un rango más amplio que el rango actual. La calidad de la imagen normalizada usando la normalización del histograma es cercana a la calidad de la imagen de referencia. El método no lineal más utilizado para la normalización radiométrica es la coincidencia de histogramas o "Histogram Matching". Este método simple se basa principalmente en el hecho de que los datos infrarrojos (> 0.7 µm) están en gran parte libres de los efectos de dispersión atmosférica, mientras que la región visible (0.4 ± 0.7 µm) está fuertemente influenciada por ellos. El método implica evaluar los histogramas de las diversas bandas de datos de detección remota de la escena deseada. Normalmente, los datos recopilados en las longitudes de onda visibles (por ejemplo, las bandas 1 a 3 de TM) tienen un valor mínimo más alto debido a la mayor dispersión atmosférica que tiene lugar en estas longitudes de onda. Por el contrario, la absorción atmosférica resta brillo de los datos registrados en los intervalos de longitud de onda más largos (p. ej. TM bandas 4, 5 y 7). Este efecto comúnmente hace que los datos de las bandas infrarrojas tengan mínimos cercanos a cero, incluso cuando pocos objetos en la escena realmente tienen una reflectancia de cero.

3.2. Normalización de imágenes de fechas múltiples mediante regresión

La normalización de imágenes de fecha múltiple implica la selección de características pseudoinvariantes (PIF), a menudo denominadas puntos de control terrestre ra-

diométrico (Helmer 2010, Jensen 2014, Hadjimitsis 2009). Para ser de valor en el proceso de normalización de imágenes de fechas múltiples, las características pseudo-invariantes deben tener ciertas características.

La regresión se utiliza para relacionar las características espectrales PIF de la imagen base con las características espectrales PIF de otras fechas de imágenes. El algoritmo asume que los píxeles muestreados en el tiempo b+1 ó b-1 están linealmente relacionados con los píxeles para las mismas ubicaciones en la imagen base "b". Esto implica que las propiedades de reflectancia espectral de los píxeles muestreados no han cambiado durante el intervalo de tiempo. Por lo tanto, la clave del método de regresión de imágenes es la selección de características pseudoinvariantes de calidad.

Numerosos científicos han investigado la utilidad de las características pseudoinvariantes para normalizar imágenes de varias fechas. Algunos de ellos desarrollaron una técnica de rectificación radiométrica que corregía imágenes de las mismas áreas mediante el uso de elementos del paisaje cuyas reflectancias eran casi constantes en el tiempo, este tipo de áreas geográficas que contienen elementos o características pseudo-invariantes se conocen como áreas pseudo-invariantes o PIA por sus siglas en inglés (G. Moré, X. Pons 2014).

4. Discusión

Se ha encontrado en algunos trabajos que el código 6S rara vez proporcionaba buenas correcciones, probablemente debido al uso de atmósferas estándar (Hadjimitsis 2004). Estos resultados sugieren que los modelos genéricos estándar de las atmósferas incluidos en los códigos de corrección atmosférica, como ATCOR-2 y el código 6S, no son lo suficientemente precisos cuando se trata de objetivos oscuros como cuerpos de agua. De todas las técnicas de corrección atmosférica basadas en imágenes, se ha encontrado que el método Darkest Pixel funciona mejor en las bandas 1, 2 y 3 de Landsat TM, y para imágenes donde la cobertura de nubes era leve. Métodos como el método de matriz de covarianza, el método de regresión y el método de intersección de regresión produjeron resultados poco confiables en la mayoría de los casos. La aplicación del método MTN (Multitemporal Normalization) ha mostrado que esta corrección atmosférica era relativamente ineficaz (Hadjimitsis 2004). Solo se eliminó una pequeña cantidad del brillo agregado de las bandas 1, 2 y 3 de Landsat TM, lo que sugiere que la imagen de referencia retuvo una contribución atmosférica significativa a la radiación medida. La selección de cualquier imagen de referencia como una "imagen clara", libre de efectos atmosféricos, no se puede realizar con mucha confianza (Hadjimitsis 2004), de que será la elección óptima al aplicar este método de Normalización Multitemporal. Trabajos más recientes como el de Moré y Pons 2012, proponen metodologías que permiten hacer una selección automática de las áreas pseudo-invariantes (PIA) para lo cual usan serie de 60 imágenes MODIS sobre la escena completa, distribuidas a lo largo de los doce meses del año y entre los años 2002 y 2008 y con el cual logran obtener mejores resultados. Mencionan que la diferencia entre la reflectividad estimada con el modelo y la reflectividad de referencia para los 3000 polígonos de prueba varía entre el ±2% en reflectividad y sin detectarse patrones temporales en las diferencias. Estos valores deseados son cercanos a los valores de ruido del sensor propio. A continuación se presenta una tabla comparativa

de los diferentes tipos de métodos de corrección atmosférica. Los métodos relativos se clasifican como métodos basados en normalización (respecto a una imagen de referencia) ((Kyo 2019, Jensen 2014):

	Facilidad de uso (Conveniencia)	Precisión en la corrección	Tipo de Corrección	Tipo de imágenes a corregir	Tipo de Procesamiento
Basados en RTM	Compleja	Depende de datos	Absoluta	Cualquiera	Pixel por Pixel y banda por banda
Basados en Imágenes	Sencilla	Consistente	Absoluta	Preferentemente con cuerpos de agua	Pixel por Pixel
Basados en normalización	Intermedia	Buena o Muy buena	Relativa	Preferentemente con PIF's	Pixel por Pixel

Tabla 1: Comparativa entre los métodos de Corrección Atmosférica

5. Conclusión

Existen diferentes técnicas disponibles para hacer corrección atmosférica, los métodos absolutos, llamado así debido a que se obtienen valores de reflectancia equivalente a la reflectancia terrestre, tienen la ventaja de que permiten obtener reflectancias equivalente a la reflectancia de la superficie terrestre, pero puede ser difícil obtener todos los parámetros necesarios para modelar la transferencia radiativa. Por otra parte los métodos relativos requieren imagen de referencia de área geográfica que contenga características pseudo-invariantes, es decir, características que varían poco en el tiempo de manera que se les puede considerar quasi-estáticas lo cual puede no ser aplicable en todas las imágenes que se desee analizar. Sobre que método elegir se puede decir lo siguiente:

- Cuando se cuente con información suficiente de calidad de las condiciones atmosféricas, día, hora del momento de la captura de la imagen, se puede decir que los métodos basados en la Ecuación de Transferencia Radiativa deberían preferirse ya que estos métodos pueden modelar con precisión el efecto de la atmósfera en la radiancia captada por el sensor del satélite.
- Cuando no se conocen con precisión o se ignoran las condiciones atmosféricas presentes al momento de capturar la imagen y se desea hacer una corrección atmosférica razonable en poco tiempo, el método DOS es la mejor opción ya que únicamente se requiere la propia imagen para obtener resultados aceptables.
- Cuando se cuenta con una gran cantidad de imágenes y se cuenta con el tiempo para hacer un estudio sobre áreas pseudo-invariantes, este método puede obtener excelentes resultados, por lo que dadas estas dos condiciones, debería optarse por este método.

El tema de corrección atmosférica es un área de investigación abierta, su importancia está en que al ser un método de pre-procesamiento, si se hace de manera correcta se van a obtener mejores resultados al momento de clasificar o segmentar las imágenes.

Referencias

D. G. Hadjimitsis, C. R. I. Clayton & V. S. Hope. (04 Jun 2010). An assessment of the effectiveness of atmospheric correction algorithms through the remote sensing of some reservoirs. International Journal of Remote Sensing, 25:18, 3651-3674.

G. Moré, X. Pons, J. Cristóbal, L. Pesquer y O. Gonzalez. (29-05-12). Authomatic radiometric correction of Landsat TM imagery through pseudoinvariant areas and modtran modelling. Revista de Teledetección, 37, 67-73.

Zhengwei Yang, Rick Mueller. (2008). UNBIASED HISTOGRAM MATCHING QUALITY MEASURE FOR OPTIMAL RADIOMETRIC NORMALIZATION. ASPRS 2008 Annual Conference, NA, 1-12.

Carmen Valdivieso-Ros, Francisco Alonso-Sarria, and Francisco Gomariz-Castillo. (1 May 2021). Effect of Different Atmospheric Correction Algorithms on Sentinel-2 Imagery Classification Accuracy in a Semiarid Mediterranean Area. Chiman Kwan, 13, 1-123.

G. Hong a & Y. Zhang. (06 October 2014). A comparative study on radiometric normalization using high resolution satellite images. Taylor & Francis, 29:2, 425-438.

Fangcao Xu, Guido Cervone, Gabriele Franch, Mark Salvador. (May 22, 2020). Multiple geometry atmospheric correction for image spectroscopy using deep learning. Journal of Applied Remote Sensing, 14(2), 024518-1 to 024518-16.

Yong Hu, Liangyun Liu, Lingling Liu, Quanjun Jiao. (2011). Comparison of absolute and relative radiometric normalization use landsat time series images. SPIE, 8006, 800616-2 al 8.

Eileen H. Helmer . (4 Oct. 2010). Radiometric Normalizationradiometric normalization. 14/Abril/2022, de Sage Sitio web: https://www.fs.fed.us/global/iitf/pubs/ja_iitf_2010_helmer002.pdf

Pablito M. López-Serrano, José J. Corral-Rivas 2, Ramón A. Díaz-Varela, Juan G. Álvarez-González and Carlos A. López-Sánchez. (29 April 2016). Evaluation of Radiometric and Atmospheric Correction Algorithms for Aboveground Forest Biomass Estimation Using Landsat 5 TM Data. Remote Sensing, 8, 1-19.

John R . Jensen. (2015). INTRODUCTORY DIGITAL IMAGE PROCESSING A Remote Sensing Perspective. University of South Carolina: Pearson Education.

Marcela Pereira-Sandoval, Ana Ruescas, Patricia Urrego, Antonio Ruiz-Verdú, Jesús Delegido, Carolina Tenjo, Xavier Soria-Perpinyà, Eduardo Vicente and Juan Soria and José Moreno. (21 June 2019). Evaluation of Atmospheric Correction Algorithms over Spanish InlandWaters for Sentinel-2 Multi Spectral Imagery Data. Remote Sensing MDPI, 11, 1-23.

JOHN R. SCHOTT, CARL SALVAGGIO, AND WILLIAM J. VOLCHOK. (1988). Radiometric Scene Normalization Using Pseudoinvariant Features. REMOTE SENSING OF ENVIRONMENT, 26, 1-16.

Priti Tyagi, Dr. Udhav Bhosle. (2011). Atmospheric Correction of Remotely Sensed Images in Spatial and Transform Domain. International Journal of Image Processing, 5, 564-579.

Seo, Dae Kyo, Eo, Yang Dam. (24/10/2019). Local-Based Iterative Histogram Matching for Relative Radiometric Normalization. Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, 37, 323-330.

Capítulo 5

Record Linkage: una revisión de los métodos de comparación

Pierre Antoine Delice, María Josefa Somodevilla García

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación
padelice@mail.com, mariajsomodevilla@gmail.com

Resumen. El presente trabajo hace una revisión de las técnicas de comparación utilizadas en la metodología del *record linkage*. Esto, con el fin de reducir errores inherentes en la determinación de las probabilidades de vinculación entre los registros. Reconociendo las consecuencias que implica imputar erróneamente características de un registro a otro, las medidas de comparación son importantes para lograr una adecuada vinculación. Con este propósito, se analizan dichos métodos, destacando los pros y contras, reportados en la literatura.

Palabras Clave: *record linkage*, algoritmo de comparación, medidas de similitud.

1 Introducción

La integración y vinculación de datos conocidas bajo el término técnico *record linkage* han sido de una amplia aplicación en las diversas áreas científicas como, por ejemplo: salud, demografía, negocios, finanzas entre otras. Señalado desde 1946 por primera vez por Halbert L. Dunn, el concepto es definido como una serie de procedimientos concebidos para vincular dos o más fuentes de información respecto a una misma entidad (individuo o evento) (Dunn, 1946). A más de 76 años, dicha técnica ganó su popularidad a partir de la década 80, alcanzando más de 5,000 publicaciones en 2018 (Asher et al., 2020).

Su uso ha sido ampliamente documentado en la literatura. Para algunos autores, es una necesidad llevar a cabo estudios que incluyen dos o más fuentes de información. Mientras que, para otros, está necesidad se resume a la de corregir las fuentes de información, buscando la completez de los registros. La conciliación demográfica, durante muchos años, ha reconocido la importancia de esta metodología para explicar las transiciones y otros fenómenos demográficos.

Además de lo anterior, es importante considerar factores como la digitalización de registros y el compromiso de los distintos gobiernos en promover políticas de datos abiertos, que incrementan la oferta de la información. Finalmente, los avances computacionales en diseño de algoritmo, difusión de librerías y software junto con la mejora en la capacidad de

procesamiento de las computadoras convencionales han hecho de la aplicación de esta metodología aún más al alcance de muchos académicos y/o organizaciones gubernamentales.

En este sentido, México requiere impulsar proyectos de vinculación de registros, en salud, por ejemplo, la aplicación de esta metodología ayudaría a generar nuevos conocimientos con base en los datos existentes, lo que tomaría tiempo y recursos si tuvieran que levantar la información. En el ámbito de la justicia, personas buscadas o desplazadas pueden ser rastreadas. Para los programas sociales, se pueden buscar beneficiarios duplicados para eficientar la asignación de recursos. En este sentido, son muchos los beneficios que aportan la vinculación de registros para la toma de decisiones.

Algunos proyectos de integración de información impulsados por el Consejo Nacional de Ciencia y Tecnología (CONACyT) y otras instituciones académicas carecen de un enfoque basado en la trazabilidad del individuo. Consisten principalmente en una recopilación de atributos para efecto de visualización y análisis donde los eventos no logran comunicarse entre sí, lo que limita comprobar ciertas hipótesis relacionadas con la trayectoria del individuo.

En miras a una aplicación de los métodos de vinculación en México, el actual trabajo busca presentar las técnicas de comparación más usadas para establecer la vinculación de una misma entidad. Sin embargo, este proceso no es libre de errores por lo que es importante conocer estas mediciones y aplicarlas adecuadamente para lograr una vinculación adecuada. Esta revisión contribuirá en el debate sobre la aplicación de una reducción considerable de los errores en la automatización y la integración de los registros.

Para eso, en la siguiente sección, se procederá a un breve análisis de los fundamentos teóricos del *record linkage*. Posteriormente, se presenta un bosquejo de la metodología de vinculación, que comprende los enfoques utilizados en la literatura, el pre-procesamiento, la indexación, así como las técnicas de comparación. Finalmente, se concluye el trabajo analizando los pros y contras de las métricas de comparación.

2 Preliminares

En la literatura, existen varios sinónimos para referirse al concepto de *record linkage*, encontramos desde: *data matching*, *data linkage*, *record matching*, *entity resolution* hasta *join* o *merge*. Sin embargo, en estricto sentido no suelen ser considerados como sinónimos propios sino como métodos que se basan en integración de bases de datos.

El campo de aplicación de este concepto es muy amplio y su utilización va más allá de la vinculación de registros con el fin de sumar atributos relativos a una misma entidad. También, se suele usar para estudiar el comportamiento de una entidad en el tiempo, así como, vincular diferentes compras realizadas por una persona en el tiempo, establecer conexión de vuelos para pasajeros, proceder a la evaluación de impacto de una política pública en el tiempo. En epidemiología, se suele usar para corregir problemas de cobertura asociados a las estadísticas vitales: nacimiento y mortalidad (Rao & Kellya, 2017).

Dr. Halbert Dunn (1946), quien por primera vez mencionó este concepto tenía el objetivo de crear un libro de vida para cada persona, empezando por los datos de nacimientos y

terminando por los de mortalidad (Dunn, 1946). Antes de abundar en la evolución de dicha metodología, es importante entender ¿por qué se busca vincular los registros? Dicho de otra manera ¿cuáles son los motivos que buscan los investigadores a través de este procedimiento?

Una primera respuesta surge a partir del trabajo de William Farr (1803-1887), uno de los epidemiólogos ingleses del siglo XIX; quien empíricamente mostró por primera vez la importancia de organizar y combinar datos de distintas fuentes sobre los individuos. Enfatizó en que este procedimiento permitiría estudiar problemas de salud desde un enfoque longitudinal y de un alcance sin precedente (Eyler, 1973; Lunde, 1975).

Sin embargo, a pesar de contar con la idea, nada o poco pudieron hacer para automatizar el ejercicio durante el siglo XIX. Es a partir de la mitad del siglo XX, con la disponibilidad de computadoras más poderosas y el desarrollo de técnicas de manejo de datos que se empieza a contar con la operacionalización de la vinculación de los registros.

En Canadá, por ejemplo, su uso inició con la necesidad de controlar la eficiencia del programa de asignación de prestación a menores de edad por medio de sus padres (Schwartz, 1946). En Estados Unidos, autores del departamento de sociología de la Universidad de Purdue usaron *record linkage* para mostrar que ciertos grupos de la población falsificaron su edad al momento de contraer matrimonio (Christensen et al., 1953). Otra aplicación consiste en la exploración, por la Oficina del adulto mayor y del seguro de supervivencia en Estados Unidos, para crear estimaciones intercensales por grupos de edad para entender la dinámica demográfica (Shryock, 1957).

Así, en 1959 Harold Newcombe y colegas, con el fin de seguir en el tiempo a un grupo de personas que fue expuesto a niveles de radiación en la provincia de British Columbia en Canadá y, determinar la asociación con la causa de muertes, sentaron la base del ejercicio de vinculación de registros desde un enfoque probabilístico usando un programa computarizado (Newcombe et al., 1959).

Este trabajo fue retomado por Fellegi y Sunter en 1969 cuyo objetivo se orienta explícitamente al problema de fusionar el contenido de la información de grandes archivos administrativos para crear una nueva fuente de información útil (Fellegi & Sunter, 1969). Con eso, crearon formalmente el primer modelo matemático para determinar cuan similar son dos campos y la combinación para su vinculación. Desde entonces, su modelo es considerado como la base teórica para la mayoría de los trabajos de vinculación de registros.

Bajo esta perspectiva, la apropiación de esta metodología se ha ido diversificando en la literatura, donde algunos interesados en entender la relación entre dos o más atributos de una persona (p.ej. educación, ingreso y mortalidad) reconozcan la importancia de la vinculación de diversas fuentes de información (Case & Deaton, 2020; Deaton, 2013; Kawachi et al., 2014). Otros, se han dedicado a usar esta metodología para estimar la sensibilidad y especificidad de los registros de mortalidad para la identificación de muertes (Lozano-Esparza et al., 2022). Respecto a los procesos de vinculación, autores como William Trick y colegas (2019) probaron diversos algoritmos para optimizar la vinculación de registros de salud en una comunidad rezagada (Trick et al., 2019). Mientras que Harron y colegas (2017) propusieron una metodología para identificar errores en los procesos de vinculación (Harron et al., 2017).

En resumen, se presentaron las motivaciones de diferentes científicos para utilizar la metodología de *record linkag*e. En lo que respecta a la presente investigación, se describirá de manera integral la metodología basada en el trabajo de Fellegi y Sunter (1969), para luego analizar las medidas de comparación.

3 Metodología

La metodología del *record linkage* consiste en una serie de pasos que inicia con el preprocesamiento y la extracción de las características para la identificación, luego la indexación, la comparación y clasificación de los registros. En la figura 1, se puede apreciar visualmente el proceso integrado.

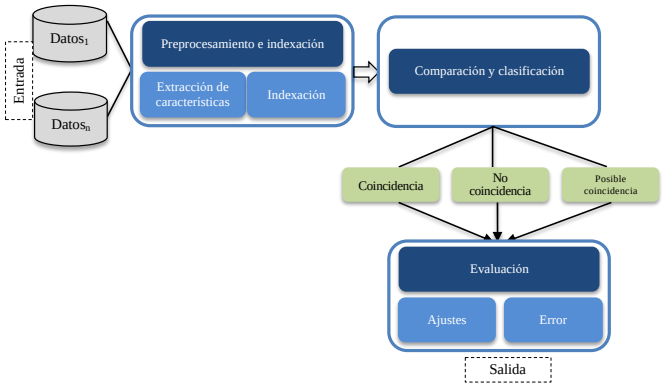


Figura 1. Proceso de vinculación de registros (Holmlund, 2019)

Antes de proceder al análisis de las etapas de la metodología, es importante entender los enfoques inherentes al proceso de vinculación.

3.1 Enfoque de vinculación

Cuando existe un identificador único en todos los conjuntos de datos, entonces el problema de vinculación es trivial. Una simple operación de unión de datos es todo lo que se requiere. Sin embargo, si los conjuntos de datos no comparten llave única, se deben aplicar varias técnicas de vinculación de registros. Estas técnicas pueden ser clasificadas en deterministas o probabilistas (Christen & Churches, 2002).

3.1.1 Técnicas deterministas

De acuerdo al enfoque determinista, las variables seleccionadas para el proceso de vinculación tienen el mismo grado de importancia ya que estamos en presencia de identificadores o llave única (p.ej. id oficial, pasaporte). Cuando concuerdan significa que los registros corresponden a la misma persona o dicho de otra manera están vinculados. Sin embargo, si por error no existe concordancia entre los registros, el proceso resulta en una no-vinculación (Ariel et al., 2014).

Una forma de generalizar este enfoque consiste en suponer que existe acuerdo o desacuerdo en un conjunto de variables $k = 1, 2, \dots, K$, (ecuación 1).

$$y_{kij} = \begin{cases} 1, si\ los\ pares(i,j)\ concuerden\ con\ las\ variables\ k \\ 0, de\ otra\ manera \end{cases} \quad (1)$$

La comparación de todas las variables para para los registros (i,j) se escribe en la ecuación 2:

$$f_{ij} = \sum_{k} \gamma_{kij} \quad (2)$$

La regla de decisión para seleccionar o no un par de registros vinculados (i,j) se presenta en la ecuación 3.

$$x_{ij} = \begin{cases} 1, f_{ij} \geq \beta \\ 0, de\ otra\ manera \end{cases} \quad (3)$$

Donde $\beta \in \{k - n, \dots, k - 1, k\}$ y n representa la cantidad de variables en las que haya desacuerdos, $0 \leq n < k$. Este modelo considera que la vinculación entre un par de registros (i,j) ocurre si los valores concuerdan al menos en $k - n$ variables. Cuando el vínculo resulta con todas las variables k entonces $n = 0$.

3.1.2 Técnicas probabilísticas

El enfoque probabilista generalmente se usa cuando no se tiene un identificador único en las bases de datos a vincular. La idea consiste en que para dos archivos (I,J), los posibles pares pueden ser divididos entre dos conjuntos distintos M (vinculado) y U (no-vinculado). Un par de registros (i,j) es un sub conjunto de M si ello se relaciona con la misma persona, de otra manera (i,j) pertenece a U. A priori, los conjuntos de M y U se desconocen. Así, el método procede a clasificar cada par de registros en M o en U (Fellegi & Sunter, 1969).

Tomando los antecedentes de la ecuación (1), se divide la probabilidad de correspondencia entre dos clases: enlaces verdaderos y enlaces falsos. La correspondencia entre enlaces verdaderos está también relacionada con errores que surjan del conjunto de variables de enlace k; por lo que, si el error es mínimo se considera, la probabilidad asociada como un enlace verdadero (cercano a 1). Así mismo, la concordancia en el subconjunto de

enlaces falsos dependerá del poder discriminatorio de la variable k. Intuitivamente, una variable discriminatoria tiende a resultar en una probabilidad baja entre los enlaces falsos.

Supone que para cada variable de enlace k, se tiene la probabilidad de correspondencia m_k entre el conjunto de los enlaces verdaderos y u_k la probabilidad de los enlaces falsos, así estas probabilidades pueden escribirse como en la ecuación 4.

$$\begin{aligned} m_k &= P\{\gamma_{kij} = 1 | (i,j) \in M\} \\ u_k &= P\{\gamma_{kij} = 1 | (i,j) \in U\} \end{aligned} \tag{4}$$

Donde γ_{kij} es el resultado binario de la comparación entre pares de registros (i,j) de la variable de enlace k. Entonces, asumiendo independencia entre las variables de enlaces, se definen estas probabilidades como en la ecuación 5.

$$\begin{aligned} \mathrm{P}\{\gamma_{\mathrm{kij}} | \mathrm{M}\} &= \prod_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{m}_{\mathrm{k}}^{\gamma_{\mathrm{kij}}} (1 - \mathrm{m}_{\mathrm{k}})^{1-\gamma_{\mathrm{kij}}} \\ \mathrm{P}\{\gamma_{\mathrm{kij}} | \mathrm{U}\} &= \prod_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{u}_{\mathrm{k}}^{\gamma_{\mathrm{kij}}} (1 - \mathrm{u}_{\mathrm{k}})^{1-\gamma_{\mathrm{kij}}} \end{aligned} \tag{5}$$

La razón de momios entre las probabilidades puede ser usada como una forma para verificar si dos registros (i,j) se enlacen o no (ver ecuación 6)

$$\frac{\mathrm{P}\{\gamma_{\mathrm{kij}} | \mathrm{M}\}}{\mathrm{P}\{\gamma_{\mathrm{kij}} | \mathrm{U}\}} \tag{6}$$

Dado que M y U se desconocen, m y u tienen que ser estimados por lo que los pares de registros se definen en (7)

$$g_{ij} = \begin{cases} (1,0)\ si\ el\ par\ (i,j) \in M \\ (0,1)\ si\ el\ par\ (i,j) \in U \end{cases} \tag{7}$$

Suponiendo una generalización de los datos en (8)

$$G = \langle \gamma_{ij}, g_{ij} \rangle \tag{8}$$

La función de probabilidad de máxima verosimilitud de todos los registros (i,j) se escribe en (9)

$$f(G|m,u,p) = \prod_{(i,j)} p g_{ij} P(\gamma_{ij}|M) + (1-p) g_{ij} P(\gamma_{ij}|U) \tag{9}$$

Donde p representa la proporción de pares de registros (i,j) que pertenece a M, y $(1-p)$ los que se encuentran en U. Resolviendo la ecuación (9) para obtener m,u,p es imposible porque los valores de concordancia indicado en la ecuación (1) son observados, mientras los valores de la variable en la ecuación (7) son desconocidos. Para resolver este problema, se puede aplicar el algoritmo Maximización Esperada conocida como EM propuesto por Dempster y colegas en 1977 (Dempster et al., 1977).

El algoritmo en cuestión consiste en dos pasos en cada iteración, la primera es la estimación de la esperanza y la segunda la maximización. Por lo que en el caso de la ecuación (9), se empieza con la estimación de m, u y p. En esta etapa se estiman los valores de la variable faltante para completar g. Una vez obtenido el valor de g, será usado como entrada para la ecuación (9), donde los nuevos valores de m, u y p se obtendrán por maximización.

Así regresamos en la ecuación (6) donde los valores de m y u serán usados para calcular las razones de momios de la correspondencia. Este resultado se conoce como el peso, dicho de otra manera:

$$\begin{cases} w_k^a = log_2\left(\frac{m_k}{u_k}\right) \; si\ las\ variables\ de\ enlace\ k\ corresponden \\ w_k^d = log_2\left(\frac{1 - m_k}{1 - u_k}\right) \; si\ las\ variables\ de\ enlace\ k\ no\ correspond \end{cases} \tag{10}$$

Donde w_k^a se refiere al peso de la correspondencia y w_k^d el peso cuando no ocurre. Asumiendo la independencia entre las variables, Porter y Winkler (1997) sugieren una generalización de los pesos de acuerdo a la fórmula:

$$T_{ij} = (w_k^a - w_k^a)\delta_{kij} + w_k^d, 0 \leq \delta_{kij} \leq 1$$

La fórmula anterior se suele usar para clasificar a los pares entre enlazados, no enlazados o posibles enlaces.

3.2 Preprocesamiento

Este paso se reconoce como la preparación de los datos, donde se procede a la estandarización de las bases. Ahí, la idea es tratar de homologar los campos bajo un mismo criterio, p. ej. ordenar el campo del nombre completo en el mismo orden, donde aparece primero los apellidos (paterno y materno) y después el nombre o separar estos campos en distintas variables para mayor comparación.

3.3 Métodos de indexación

Aunque no es una necesidad, la indexación suele ayudar en la preparación de los datos para la comparación. Constituye un filtro para reducir el costo de operación del algoritmo. Sin indexación, el espacio de comparación S para una base de datos D se estiman en $S = |D| * \frac{|D|-1}{2}$, lo que aproxima a un costo cuadrático. Por lo que, para grandes volúmenes de datos, el algoritmo es muy costo computacionalmente. Para tratar estos casos, diferentes tipos de indexación se usan, donde básicamente se reduce el número de combinaciones de

pares. Uno de los más usado es bloqueo (*blocking*), consistiendo en usar algunos campos de la base de datos cuyas características se aproximan o son iguales.

3.4 Tipos de métodos de comparación

La comparación juega un papel preponderante en el proceso de vinculación, consiste en establecer la existencia de concordancia entre dos registros. Hay diferentes formas de llevar a cabo una comparación, sin embargo, se distinguen dos grandes perspectivas: algoritmo basado en edición de caracteres o en token.

3.4.1 Basado en edición de caracteres

Los métodos basados en caracteres son especiales para los casos con errores de tipografía, generalmente introducidos por humano. En esta sección, se presentan algunos de estos algoritmos: Levenshtein (1966), Jaro (1989), Jaro-Winkler (1990), Q-Grama, Cosine, y Smith-Waterman (1981).

3.4.1.1 Levenshtein (1966)

La distancia de Levenshtein, conocida como distancia de edición mide la cantidad mínima de operación necesaria para convertir una cadena de texto A en otra B. El algoritmo de Levenshtein toma sus valores en un rango de $[0, \max(|A|, |B|)]$, donde 0 significa que las cadenas sean iguales (Levenshtein, 1965). Matemáticamente, se escribe como:

$$d_l(A, B) = 1.0 - \frac{Levenshtein(A, B)}{max(|A|, |B|)}$$

3.4.1.2 Jaro (1989)

Este algoritmo tiene como objetivo medir la cantidad de transposición requerida de un carácter para transformar (t) una cadena de texto a otra (Jaro, 1989). La similitud de Jaro entre 2 cadenas de textos está dada por la siguiente formula:

$$d_j(A, B) = \frac{1}{3}\left(\frac{m}{|A|} + \frac{m}{|B|} + \frac{m - t/2}{m}\right)$$

3.4.1.3 Jaro-Winkler (1990)

Winkler y Thibaudeau (W. Winkler, 1990; W. E. Winkler & Thibaudeau, 1987) presentaron una versión mejorada del algoritmo de Jaro, asignando mayor probabilidad a las cadenas de textos cuyas primeras letras coinciden. El algoritmo de similitud Jaro-Winkler es dado por:

$$d_{jw}(A,B) = d_j(A,B) + \frac{c}{10}\Big(1 - d_j(A,B)\Big)$$

c representa el número de caracteres concordantes al principio de cada instancia.

3.4.1.4 Q-Gram

Este algoritmo conocido como n-grama divide las cadenas de texto entre subsecuencias de longitud q, luego se hace una comparación de los subconjuntos para determinar el grado de similitud. La ecuación está dada por:

$$d_{qg}(A,B,q) = \frac{|Qgram(A,q) \cap Qgram(B,q)|}{max(|Qgram(A,q), Qgram(B,q)|)}$$

3.4.1.5 Cosine

La medida de similitud del Coseno prioriza el contenido entre las cadenas de texto en ves del orden. La medición contempla el número de caracteres únicas y espacios entre las cadenas. La fórmula del coseno del ángulo entre dos vectores A y B es dada por:

$$d_{cs}(A,B) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \; \sqrt{\sum_{i=1}^{n} B_i^2}}$$

3.4.1.6 Smith-Waterman (1981)

Smith y Waterman propusieron un algoritmo parecido al de Levenshtein con el fin de de medir el número de eventos requerido para convertir unas secuencias de proteínas en otra. A diferencia del algoritmo de Levenshtein, este asigna un puntaje para las diferencias y los caracteres específicos. Esto permite un sistema de puntajes donde caracteres similares tendrán puntajes bajos que una total coincidencia, y un puntaje alto que uno negativo o cero.

Smith-Waterman también considera los espacios al momento de comparar dos cadenas lo que arroja distintas puntuaciones de espacios y de longitud (Smith & Waterman, 1981).

3.4.2 Basado en Token

A diferencia de las medidas de similitud basadas en caracteres, las de token se suelen usar cuando las cadenas de textos presentan errores de estructura como p. ej. "María José" y "José María". Al igual que los algoritmos basados en caracteres, las medidas de similitud asociadas con este enfoque arrojan valores entre 0 y 1, con total disimilitud y 1 total similitud. En esta sección, se analizarán las medidas de similitud de Jaccard y Dice.

3.4.2.1 Jaccard (1912)

La medida de similitud de Jaccard es definida como la división entre el tamaño de la intersección y el de la unión de un conjunto de dos cadenas de texto (Jaccard, 1912). Matemáticamente, es representada por la siguiente formula:

$$d_{jc}(S_a, S_b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}$$

3.4.2.2 Coeficiente de Dice (1945)

El coeficiente de Dice es una medida de asociación entre dos conjuntos, que fue diseñado para calcular el grado de similitud entre dos especies (Dice, 1945). Matemáticamente, se escribe de esta manera:

$$d_{dc}(S_a, S_b) = 2 * \frac{|S_a \cap S_b|}{|S_a| + |S_b|}$$

4 Conclusiones

En este trabajo, se hizo una breve revisión de los principales métodos de comparación existentes en la literatura para evaluar el grado de similitud entre dos cadenas de texto y con eso proceder a la vinculación de los registros. Como destacado en el texto, la vinculación de los registros plantea diversos retos a lo largo de su ejecución. Tratarlos implica analizar

las diferentes etapas de la metodología, haciendo una revisión de las principales técnicas utilizadas en cada paso (Wang & Dong, 2020). En efecto, las técnicas de comparación juegan un papel preponderante, ya que definen las probabilidades de correspondencia entre dos o más cadenas de texto. Dichas probabilidades pueden resultar en tres clases cuyas categorías son: vinculado, no vinculado o posible vínculo (*Vinculado* cuando se encuentra un par exacto, *No vinculado* cuando no existe este par y *posible vínculo* cuando existen dudas sobre la correspondencia). De esta manera, dichas técnicas son consideradas como las métricas que definen cuando dos o más registros se corresponden. Para esto, se ha analizado algunas de las más citadas en la literatura, enfatizando en los pros y contras para hacer más eficiente su uso y de esta manera disminuir los errores durante el proceso de vinculación.

Actualmente, se está trabajando en el diseño de un experimento para comparar empíricamente las métricas analizadas, con el objetivo de determinar bajo qué condiciones es recomendable utilizar cada una. Se usarán los resultados de este experimento, para proceder a la vinculación de la información pública para completar el proceso de *record linkage*. En particular, se aplicará a un proyecto de vinculación de las bases de datos de mortalidad, nacimientos y variables socio-económicas, provenientes de encuestas y censos realizado en México en el periodo 1985-2020.

Referencias

Ariel, A., Bakker, B. F. M., Groot, M. de, Grootheest, G. van, Laan, J. van der, Smit, J. H., & Verkerk, B. (2014). Record linkage in health data: A simulation study. *Statistics Netherlands*, 64.

Asher, J., Resnick, D., Brite, J., Brackbill, R., & Cone, J. (2020). An Introduction to Probabilistic Record Linkage with a Focus on Linkage Processing for WTC Registries. *International Journal of Environmental Research and Public Health*, *17*(18), 6937. https://doi.org/10.3390/ijerph17186937

Case, A., & Deaton, A. (2020). *Deaths of Despair and the Future of Capitalism*. Princeton University Press; JSTOR. https://doi.org/10.2307/j.ctvpr7rb2

Christen, P., & Churches, T. (2002). *Febrl—Freely extensible biomedical record linkage*. http://hdl.handle.net/1885/40723

Christensen, H. T., Andrews, R., & Freiser, S. (1953). Falsification of Age at Marriage. *Marriage and Family Living*, *15*(4), 301–304. JSTOR. https://doi.org/10.2307/347835

Deaton, A. (2013). *The Great Escape: Health, wealth, and the origins of inequality*. Princeton University Press. http://www.amazon.com/The-Great-Escape-Origins-Inequality/dp/069115354X

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. JSTOR. http://www.jstor.org/stable/2984875

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. JSTOR. https://doi.org/10.2307/1932409

Dunn, H. L. (1946). Record Linkage. *American Journal of Public Health and the Nation's Health*, *36*(12), 1412–1416. https://doi.org/10.2105/AJPH.36.12.1412

Eyler, J. M. (1973). William Farr on the Cholera: The Sanitarian's Disease Theory and. The Statistician's Method. *Journal of the History of Medicine*, 79–100. http://jhmas.oxfordjournals.org/

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *69*(328), 1183–1210.

Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., & van der Meulen, J. H. (2017). A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology*, *46*(5), 1699–1710. https://doi.org/10.1093/ije/dyx177

Holmlund, O. (2019). Evaluating record linkage methods for manifold identity detection [Student thesis]. En *UMNAD: Vol. Independent thesis Advanced level (degree of Master (Two Years))*. DiVA. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-164720

Jaccard, P. (1912). The distribution of the Flora in the Alpine zone 1. *New Phytologist*, *11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414–420. JSTOR. https://doi.org/10.2307/2289924

Kawachi, I., Glymour, M. M., Berkman, L. F., & Kawachi, I. (2014). Social epidemiology. En *Social Epidemiology* (2 ed.). Oxford University Press.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, *10*, 707–710.

Lozano-Esparza, S., Zazueta, O. E., Hernández-Ávila, J. E., & Lajous, M. (2022). Comparing the usefulness of two mortality registries for data-linkage for prospective cohorts in Mexico. *Salud Pública de México*, *64*(1), 96–99. https://doi.org/10.21149/13384

Lunde, A. S. (1975). The Birth Number Concept and Record Linkage. *American Journal of Public Health*, *65*(11), 1165–1169.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic Linkage of Vital Records. *Science*, *130*(3381), 954–959. https://doi.org/10.1126/science.130.3381.954

Rao, C., & Kellya, M. (2017). *Overview of the principles and international experiences in implementing record-linkage mechanisms to assess completeness of death registration* [Technical paper]. United Nations Department of Economic and Social Affairs.

Schwartz, E. E. (1946). Some Observations on the Canadian Family Allowances Program. *Social Service Review*, *20*(4), 451–473. https://doi.org/10.1086/636025

Shryock, J. H. S. (1957). Development of Postcensal Population Estimates for Local Areas. En *Regional Income*. NBER. http://www.nber.org/books/unkn57-3

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Trick, W. E., Doshi, K., Ray, M. J., & Angulo, F. (2019). Development and Evaluation of Record Linkage Rules in a Safety-Net Health System Serving Disadvantaged Communities. *ACI Open*, *03*(02), e63–e70.

Wang, J., & Dong, Y. (2020). Measurement of Text Similarity: A Survey. *Information*, *11*(9). https://doi.org/10.3390/info11090421

Winkler, W. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*.

Winkler, W. E., & Thibaudeau, Y. (1987). An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census. En *Technical report, US Bueraur of the Census*.

Capítulo 6

Descubrimiento de tópicos con aprendizaje profundo: una revisión preliminar sistemática del estado del arte

Ana Laura Lezama Sánchez[1], Mireya Tovar Vidal[1], and José A. Reyes-Ortiz[2]

[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, 72590, Puebla, México
ana.lezama@alumno.buap.mx, mireya.tovar@correo.buap.mx
[2]Departamento de Sistemas, Universidad Autónoma Metropolitana, Azcapotzalco, 02200, Ciudad de México
jaro@azc.uam.mx

Resumen. En el presente trabajo se presenta una revisión del estado del arte relacionado con enfoques para el descubrimiento de tópicos con aprendizaje profundo. El objetivo es dar a conocer algunas metodologías existentes. Cada autor incorpora en su metodología algún modelo para descubrimiento de tópicos existente o propone una nueva variante, además incorpora el uso de aprendizaje profundo con la finalidad de obtener una representación más ordenada. El propósito es obtener una evaluación de los tópicos a través de la métrica por coherencia del tópico con mayor puntaje y este pueda ser una tarea que se incorpore en alguna aplicación más detallada.

Palabras Clave: Descubrimiento de tópicos, aprendizaje profundo, coherencia del tópico.

1 Introducción

El descubrimiento de tópicos (*topic discovery*) es una tarea para el reconocimiento de tópicos en un documento, informando de manera general su contenido y proporcionando resultados apropiados para el desarrollo de tareas más complejas como establecer relaciones entre conceptos (Vilca, 2017).

La tarea de descubrir los tópicos presentes en un documento se puede llevar a cabo de forma manual, pero es costosa y tediosa si la cantidad de documentos a revisar es elevada. Por esta razón ha sido necesario la existencia de modelos computacionales capaces de llevar a cabo esta tarea de manera similar en como el humano lo realiza, pero en un menor tiempo (Duan et al., 2021).

El objetivo del descubrimiento de tópicos es extraer información a partir de textos para encontrar el tópico del cual tratan. Es decir, el propósito es que ciertos temas aparezcan en documentos relevantes, siendo útil cuando se usan para asociar con un documento como palabras clave para una búsqueda más intuitiva o para generar un resumen.

El descubrimiento de tópicos ha sido un problema ampliamente estudiado durante varios años, por medio de diferentes técnicas o métodos computacionales. Algunas de las técnicas más usadas son Análisis Semántico Latente (LSA por sus siglas en inglés, *Latente Semantic Analysis*), Análisis de Dirichlet Latente (LDA por sus siglas en inglés, *Latent Dirichlet Allocation*) y Análisis Semántico Latente Probabilístico (PLSA por sus siglas en inglés, *Probabilistic Latent Semantic Analysis*) (Alfaro-Flores, 2014). Los resultados proporcionados por las técnicas antes mencionadas son evaluados con la métrica de coherencia del tópico que mide el nivel de coherencia de los tópicos recuperados. Cabe recalcar que esta métrica de evaluación utiliza un corpus externo para su operación.

Algunos autores reportan el descubrimiento de tópicos con algoritmos adicionales tales como similitud coseno, modelos de agrupamiento y/o clasificación, entre otros; reportando resultados prometedores. Sin embargo, todos concluyen en la importancia de aplicar otros algoritmos para la mejora de los resultados obtenidos. Los algoritmos propuestos por varios autores para mejorar los resultados obtenidos del descubrimiento de tópicos son principalmente modelos de aprendizaje profundo e incrustación de palabras.

Este documento esta estructurado de la siguiente manera, en la sección 2 se presentan algunos trabajos encontrados en la literatura sobre el descubrimiento de tópicos con aprendizaje profundo agrupados de acuerdo a la evaluación realizada por cada autor, es decir, extrínseca e intrínseca. La característica que presentan estos trabajos es que incorporan en sus procesos algún modelo neuronal como *autoencoder, LSTM, HAN,* entre otros. La sección 3 presenta un resumen del aprendizaje profundo y su incursión en el descubrimiento de tópicos. Por otro lado, la sección 4 expone la relación entre el aprendizaje profundo y el descubrimiento de tópicos detectado durante el análisis. En la sección 5 se proporciona nuestro punto de vista sobre los trabajos expuestos, se presenta el trabajo a futuro y finalmente se exponen las referencias consultadas hasta el momento.

2 Enfoques para el descubrimiento de tópicos con aprendizaje profundo

En la literatura existen diferentes enfoques para el descubrimiento de tópicos con aprendizaje profundo. El propósito es obtener una representación ordenada, con la probabilidad de que la tarea de descubrimiento de tópicos obtenga un mayor resultado de evaluación a través de la métrica de coherencia (Mohanapriya and Beena, 2021).

A continuación, se presentan trabajos relacionados que reportan el uso de aprendizaje profundo y un método tradicional para el descubrimiento de tópicos o uno nuevo generado por los autores basados en los existentes o con otras herramientas como similitud coseno, agrupamiento, entre otros.

Para (Londoño, 2016) todo tema de investigación cuenta con información sobre resultados previos obtenidos por otros autores en la misma área en cuestión. Por lo que para dedicarse a investigar sobre un tema es necesario realizar la lectura de artículos científicos publicados en revistas especializadas,

bibliotecas, entre otros. Por lo que este es el único camino para la construcción de un estado del arte y consecuentemente comprender los procesos y resultados proporcionados de cualquier investigación. Por lo tanto, los trabajos reportados en este documento se encuentran en diversas bases de datos, por mencionar algunas: Web of Science y Springer.

Evaluación de tópicos de manera extrínseca

Los trabajos expuestos en este apartado llevan a cabo la tarea de descubrir los tópicos en documentos orientado a alguna aplicación en particular.

En (Pandey, 2021) se propone un método para el descubrimiento de sentimientos en torno a COVID-19 en las redes sociales. Los autores utilizan *BERT* para el descubrimiento de tópicos y la clasificación de sentimientos sobre los comentarios existentes sobre COVID-19 en la red social *Reddit*. El método emplea *LDA* y el mapeo de *Gibbs*, además de la búsqueda de palabras clave relacionadas con la salud mental y la creación de bigramas y una red de palabras relevantes para formar una opinión general sobre varios temas. El modelo *BERT* incorpora los tópicos latentes y usa la relación entre diferentes palabras para mejorar la polaridad de los sentimientos. Para determinar la polaridad de las palabras utilizaron el lexicón *VADER*, etiquetando cada comentario en positivo, negativo o neutral.

Por otro lado, en (Huang et al., 2020) aplicaron el uso de factorización matricial no negativa y un *autoencoder* con medición de divergencia de información para monitorear la evolutiva de difusión del tópico para comprender cómo los tópicos de investigación cambian con el tiempo. Los resultados obtenidos por los autores mostraron que el enfoque propuesto es capaz de identificar la evolución de los tópicos y su difusión. Los autores adoptaron el concepto de factorización matricial no negativa (*NMF*) y aunado al *autoencoder* construyen un modelo de tópicos que descubren una estructura de término-tópico multicapa que nombraron *DNAE*. El modelo propuesto construye un codificador y decodificador comprimiendo y descomprimiendo una matriz de términos. Además, entrenaron modelos de tópicos que procesaron una serie de corpus y utiliza la divergencia de información para evaluar la magnitud de las difusiones de tópicos a lo largo del tiempo. Los autores extrajeron 31,904 documentos relacionados con aprendizaje automático para llevar a cabo pruebas con el modelo propuesto.

Un método que analiza la asociación entre el sentimiento de los comentarios de COVID-19 y un método de PLN basado en el modelado de tópicos para descubrir problemas relacionados con COVID-19 a partir de las opiniones públicas es propuesto por (Jelodar et al., 2020). El método hace uso de una red neuronal recurrente *LSTM* para la clasificación de sentimientos de los comentarios sobre COVID-19. El objetivo de los autores fue demostrar la importancia de utilizar opiniones públicas y técnicas computacionales adecuadas para comprender los problemas que rodean la enfermedad y guiar la toma de decisiones.

Además (Ma et al., 2016) expone un método para el análisis de la opinión pública en la red social *weibo.com*. Como modelo de aprendizaje profundo los autores emplean *word2vec* para calcular el vector de cada término del corpus, y luego obtener palabras emocionales y su correspondiente intensidad emocional mediante el cálculo de su distancia de coseno; a continuación, mediante el uso

de análisis de series temporales, se rastrea el cambio de la intensidad emocional del público junto con la evolución del evento. *LDA* es usado para explorar los diferentes puntos de vista y opiniones del público sobre un evento que sea tendencia en las redes sociales.

Evaluación de tópicos de manera intrínseca

Los trabajos expuestos en este apartado llevan a cabo la tarea de descubrir los tópicos en documentos y cada modelo de descubrimiento de tópicos es evaluado por medio de alguna métrica tradicional del procesamiento del lenguaje natural (PLN).

En (Eslami et al., 2020) exponen un método para encontrar la relación entre un fármaco y sus efectos secundarios según lo informado por los usuarios habituales de un sitio web llamado pregunte a un paciente. Los autores desarrollaron un método basado en aprendizaje profundo donde encontraron que los comentarios de los usuarios sobre los efectos secundarios de los medicamentos se encuentran sesgados. Cada comentario lo clasificaron con el método *HAN* (*Hierarchical Attention Network*) y *FastText* y *word2vec* para la representación de los documentos. Posteriormente por medio de la factorización matricial no negativa (*NMF*) descubren los tópicos para identificar eventos farmacológicos que conducen a una reacción inmediata y al descubrimiento de estos eventos. El modelo fue evaluado por medio de las métricas precisión, medida-F_1, exactitud y *Kappa*. Los fármacos descubiertos fueron aquellos utilizados en antidepresivos, anticonceptivos y para la digestión.

En (Shahbazi and Byun, 2020) expusieron un método para la combinación de aprendizaje por refuerzo (*autoencoder y decoder*) y un modelo de factorización de matriz no negativa asistido por semántica para extraer tópicos significativos y subyacentes de contenidos de documentos breves. El método propuesto generó resultados aceptables para la tarea de descubrimiento de tópicos. El objetivo de los autores fue reducir el problema de la información repetitiva y la escasez de datos en textos cortos. La evaluación la realizaron con la métrica de coherencia del tópico, precisión, exhaustividad, medida-F_1 y exactitud. Los conjuntos de datos utilizados por los autores fueron yahoo.Ans y DBLP. Los autores implementaron factorización de matrices no negativas, *LDA, GPUDMM* y modelo de tema basado en pseudo-documento (*PTM*). Los resultados del modelo propuesto demostraron que los tópicos extraídos de los datos son significativos y los términos son coherentes con los tópicos.

Un modelo para el descubrimiento de tópicos con una red neuronal recurrente *(LSTM* memoria a largo plazo, en inglés *Long Short-Term Memory)* fue propuesto por (Shafqat et al., 2020). El objetivo principal de los autores fue implementar un sistema de recomendación que utilice el modelo de descubrimiento de tópicos. El sistema es capaz de aprender sobre las preferencias del usuario y posteriormente le recomienda un producto. El modelo utilizó datos no textuales y de reseñas. Los datos no textuales rastrearon la información estática proporcionada. Los datos textuales los usaron en *LDA* para extraer vectores latentes de tópicos y adoptan la arquitectura *LSTM* para generar vectores latentes de documentos. Los autores proponen una recopilación de datos (comentarios

hechos por usuarios relacionados con un producto). El método fue evaluado por medio de la métrica de exactitud.

En (Srivastava and Sutton, 2017) proponen el uso de una red neuronal para lograr un ajuste de un método de inferencia basado en codificación automática de *Bayes variacional (AEVB)* para *LDA*. Los autores realizaron evaluaciones por medio de la métrica de coherencia del tópico en un millón de documentos. Los córpora utilizados durante los experimentos fueron *20-Newsgroup* y *Reuters*.

Por otro lado, (Bougteb et al., 2019) propone el uso de un autoencoder para la extracción de los tópicos presentes en un conjunto de tweets sobre noticias y empresas como *Apple, Google* y *Microsoft*. Los autores hacen pruebas con el modelo de ventana obteniendo resultados significativos con 5 y 10. Posteriormente ocuparon el algoritmo *k-means++* para obtener el número de grupos necesarios para la clasificación de los tópicos. El segundo conjunto de datos lo probaron con los algoritmos de agrupamiento *CluStream, DenStream y Dstream* mejorando los resultados al incorporar el *autoencoder*. El conjunto de datos es representado por medio de *Word2Vec* y la similitud coseno. Los autores evaluaron con las métricas de precisión, exhaustividad y medida-F_1.

Un método de minería de textos llamado neural *topic embedding* capaz de extraer representaciones útiles e interpretables de textos a través de autocodificador variacional es propuesto por (Chai and Li, 2019). El método es capaz de resolver tareas de aprendizaje supervisado, semi-supervisado y de aprendizaje multitarea. El método fue evaluado con un banco de preguntas de reseñas de clientes. Los autores consideran que el método que proponen hace las siguientes contribuciones a la comunidad de investigación: incorporación de la interpretación en aplicaciones complejas de aprendizaje profundo en la minería de textos. Las métricas usadas para la evaluación fueron precisión, exactitud, exhaustividad y medida-F_1.

Por otro lado, (Jin et al., 2018) proponen un modelo basado en *LSTM* y *LDA* que denominaron *LSTM-Topic* matrix factorization (*LTMF*) que se integran en un marco de factorización matricial para la revisión de sistemas de información. El conjunto de datos utilizado en los experimentos fue de *Amazon* dividido en aplicaciones para *Android*, instrumentos musicales, productos de oficina, video juegos, entre otros. *LTMF* mostró una mejor capacidad para descubrir tópicos que el *LDA* tradicional. Los autores realizaron pruebas con 5 *baselines* diferentes lo que les permitió comprobar que el método propuesto es capaz de obtener mejores resultados en comparación con los demás. La métrica de evaluación utilizada fue el error cuadrático medio (*MSE*).

En la Tabla 1 se presentan las características más representativas de cada trabajo presentado en esta sección. La mayoría utilizan la técnica *LDA*, pero con la diferencia que incorporan alguna arquitectura de aprendizaje profundo obteniendo resultados alentadores al incorporar modelos neuronales, pero aún consideran necesario incorporar alguna técnica adicional que mejore los resultados proporcionados por las métricas utilizadas durante su experimentación. Todos los trabajos usan córpora en idioma inglés.

Tabla 1. Trabajos relacionados con descubrimiento de tópicos y aprendizaje profundo.

Autor	Algoritmos y recursos	Dominio de datos	Métricas de evaluación
(Chai and Li, 2019)	*Variational autoencoding*, *Supervised Latent Dirichlet Allocation (SLDA)*	Reseñas	Precisión, exactitud, exhaustividad, medida-F_1
(Ma et al., 2016)	*Word2vec, LDA*	Tuits de Redes sociales	Precisión, exhaustividad, medida-F_1
(Shafqat et al., 2020)	*LSTM, LDA*	Reseñas	Exactitud
(Srivastava and Sutton, 2017)	*AEVB, LDA, Neural Variational Document Model*	Noticias	Coherencia del tópico
Eslami et al. (2020)	*HAN, Word2Vec, FastText, NMF*	Fármacos	Precisión, exactitud, medida-F_1, *Kappa*
(Jelodar et al., 2020)	*LSTM, LDA, Gibbs sampling*	COVID-19	Coherencia del tópico
(Bougteb et al., 2019)	*CluStream, DenSream, Dstream,* similitud coseno, *k-means++, autoencoder, Word2Vec*	Noticias y finanzas	Precisión, exhaustividad, medida-F_1
(Pandey, 2021)	*BERT, BIOBERT, SCIBERT, LDA, VADER,*	Tuits de Redes sociales	Sin especificar
(Huang et al., 2020)	*Autoencoder, NMF*	Científico	Sin especificar
(Shahbazi and Byun, 2020)	*Autoencoder, decoder,* Factorización de matrices no negativas, *LDA, GPUDMM* y *PTM*	Científico, deportes, finanzas, negocios.	Coherencia del tópico

Los trabajos analizados proponen usar diferentes características para una misma tarea, es decir descubrir tópicos por medio de un modelo tradicional (LDA) o generando uno nuevo como en

(Bougteb et al., 2019) que descubren tópicos sin usar ninguno de los modelos tradicionales, sino aplicando similitud coseno, modelos de incrustación, agrupamiento y un *autoencoder*.

Hasta el momento se han analizado trabajos relacionados con el descubrimiento de tópicos con aprendizaje profundo. Estas aproximaciones aportan buenos resultados en comparación con los reportados en la literatura con modelos tradicionales (como LDA) para el descubrimiento de tópicos o nuevas técnicas, pero sin usar ningún modelo neuronal.

3 Aprendizaje profundo y su incursión en el descubrimiento de tópicos

El uso de aprendizaje profundo y su incursión en el descubrimiento de tópicos se ha hecho presente dada su precisión en comparación con un clasificador tradicional. Esto se debe que a pesar de que varias de las técnicas tradicionales existentes para el descubrimiento de tópicos obtienen buenos resultados, ya sea evaluándolos con alguna métrica como la coherencia del tópico o por un experto aún se tenía la necesidad de mejorarlos. Por esta razón, autores como (Rekik and Jamoussi, 2016), (Bougteb et al., 2019), (Eslami et al., 2020), entre otros proponen el uso de algún modelo neuronal porque los elegidos por ellos son capaces de retirar "ruido" es decir, en el caso de un autoencoder eliminar elementos dentro del conjunto de datos que no aportaran información importante y modelos como LDA no son capaces de eliminarlo. Como se puede observar la incursión de aprendizaje profundo en el descubrimiento de tópicos ha favorecido en los resultados esperados, es decir, tópicos coherentes según el dominio a manejar y por lo tanto ese modelo para descubrir tópicos será importante en otras tareas del PLN. Sin embargo, aún existen modelos neuronales por destinar a la tarea de descubrimiento de tópicos y se espera que los resultados obtenidos sean mayores en cuanto a coherencia del tópico.

4 Relación aprendizaje profundo y descubrimiento de tópicos

En la literatura las redes neuronales se describen como aquellas que pueden modelar abstracciones de alto nivel y disminuir las dimensiones utilizando múltiples capas de procesamiento basadas en estructuras y combinándolas con transformaciones no lineales. Por lo que se estima que esa capacidad de procesamiento es la que ha generado que exista una relación con la tarea de descubrimiento de tópicos. Por lo que después de analizar los trabajos expuestos con anterioridad se determina que aún es necesario extender los estudios sobre el descubrimiento de tópicos con los modelos neuronales existentes.

Hasta el momento los modelos neuronales usados en los trabajos descritos son HAN, Autoencoder y LSTM generando buenos resultados para los autores. Sin embargo, aún se enfrentan a problemas como en algunos casos la necesidad de determinados datos previamente validados para llevar a cabo un entrenamiento. El uso de técnicas híbridas de aprendizaje profundo difusas orientado

al descubrimiento de tópicos aún no se encuentra reportado en la literatura. Sin embargo, (Eslami et al., 2020) lo propone como trabajo a futuro, pero orientado además al análisis de sentimientos.

5 Conclusiones

En este artículo se revisaron diferentes trabajos relacionados con descubrimiento de tópicos y el aprendizaje profundo. En general, el idioma utilizado en los experimentos es el inglés. Los dominios fueron noticias, finanzas, COVID-19, fármacos, entre otros.
Se observó una variedad de herramientas utilizadas en el desarrollo de los trabajos de cada autor. Todos los autores aplican algún modelo neuronal en sus metodologías con el propósito de primero obtener clases de cada corpus a usar. Cada trabajo reportó que incluir aprendizaje profundo generó mejoras en los resultados obtenidos en comparación con los proporcionados cuando se aplican técnicas tradicionales de descubrimiento de tópicos sin aprendizaje profundo.
Como trabajo a futuro se contempla profundizar en estado del arte sobre descubrimiento de tópicos con modelos de incrustación de palabras como *Word2Vec, Glove,* entre otros.

Agradecimientos

Los autores agradecen al Laboratorio Nacional de Supercómputo del Sureste de México (LNS), perteneciente al padrón de laboratorios nacionales CONACYT, por los recursos computacionales, el apoyo y la asistencia técnica brindados, a través del proyecto No. 202103090C.

Referencias

Vilca, G. C. V. (2017). Generación automática de resúmenes abstractivos mono documento utilizando análisis semántico y del discurso. PhD thesis, Pontificia Universidad Católica del Peru-CENTRUM Catolica (Perú).

Duan, Z., Xu, Y., Chen, B., Wang, C., Zhou, M., et al. (2021). Topic-net: Semantic graph-guided topic discovery. *Advances in Neural Information Processing Systems*, 34.

Alfaro-Flores, R. (2014). Evaluación del efecto en el algoritmo de análisis semántico latente al utilizar colecciones de datos cada vez más grandes para la detección y extracción de sinónimos y su independencia respecto al lenguaje, por medio de su implementación distribuida. Tesis de maestría, Instituto Tecnológico de Costa Rica.

Mohanapriya, M. D. y Beena, R. (2021). Improving topic modelling for prediction of drug indication and side effects. *Annals of the Romanian Society for Cell Biology*, pages 11542–11558.

Pandey, C. (2021). redbert: A topic discovery and deep sentiment classification model on covid-19 online discussions using bert nlp model. *International Journal of Open Source Software and Processes (IJOSSP)*, 12(3):32–47.

Huang, S.-T., Kang, Y., Hung, S.-M., Kuo, B., y Cheng, I.-L. (2020). Topic diffusion discovery based on deep non-negative autoencoder. In 2020 *IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 405–408. IEEE.

Jelodar, H., Wang, Y., Orji, R., y Huang, H. (2020).Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *arXiv preprint ar-Xiv:2004.11695*, pages 1–23.

Ma, B., Yuan, H., Wan, Y., Qian, Y., Zhang, N., and Ye, Q. (2016). Public opinion analysis based on probabilistic topic modeling and deep learning. PACIS 2016 proceedings, 171:1–11.

Eslami, B., Rezaei, Z., Habibzadeh, M., Fouladian, M., y Ebrahimpour-Komleh, H. (2020). Using deep learning methods for discovering associations between drugs and side effects based on topic modeling in social network. *Social Network Analysis and Mining*, 10:1–17.

Shahbazi, Z. y Byun, Y.-C. (2020). Topic modeling in short-text using non-negative matrix factorization based on deep reinforcement learning. *Journal of Intelligent & Fuzzy Systems*, 39(1):753–770.

Shafqat, W. et al. (2020). A Hybrid Approach for Topic Discovery and Recommendations based on Topic Modeling and Deep Learning. PhD thesis, Jeju National University.

Srivastava, A., y Sutton, C. (2017). "Autoencoding variational inference for topic models". *ICLR* 2017.

Bougteb, Y., Ouhbi, B., Frikh, B., et al. (2019). Deep learning based topics detection. In 2019 Third International Conference on Intelligent *Computing in Data Sciences (ICDS),* pages 1–7. IEEE.

Chai, Y. y Li, W. (2019). Towards deep learning interpretability: A topic modeling approach. *ICIS,* 26:1–10.

Jin, M., Luo, X., Zhu, H., y Zhuo, H. H. (2018). Combining Deep learning and topic modeling for review understanding in context-aware recommendation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1605–1614.

Londoño, O. (2006). Cómo escribir artículos científicos. Bogotá: Universidad Cooperativa de Colombia. Bogota: EDUCC

Índice de autores

Nombre del Autor	Nacionalidad	
Adelina Escobar Acevedo	Mexicana	
Amparo Palomino Merin	Mexicana	
Ana Laura Lezama Sánchez	Mexicana	
Arturo Jasso Garduño	Mexicana	
Daniel Marcelo González Arriaga	Mexicana	
David Pinto	Mexicana	
Edmundo Cortes Vazquez	Mexicana	
Ignacio Muñoz Máximo	Mexicana	
José Alejandro Reyes Ortiz	Mexicana	
Josefina Guerrero García	Mexicana	Editora
María Aurora Diozcora Vargas Treviño	Mexicana	
María Josefa Somodevilla García	Mexicana	
Mireya Tovar Vidal	Mexicana	Editora
Pierre Antoine Delice	Haitiano	
Sergio Vergara Limon	Mexicana	

Compiladores

Mireya Tovar Vidal
Claudia Zepeda Cortés
Darnes Vilariño Ayala
Juan Manuel González Calleros
Josefina Guerrero García

Revisores

Abraham Sánchez López
Amparo Dora Palomino Merino
Beatriz Beltrán Martínez
Carmen Cerón Garnica
Claudia Zepeda Cortés
Darnes Vilariño Ayala
Georgina Flores Becerra
Guillermo De Ita Luna
Helena Gómez Adorno
Hilda Castillo Zacatelco
Iván Olmos Pineda
Ivo H. Pineda Torres
José Andrés Vázquez Flores
José Arturo Olvera López
José de Jesús Lavalle Martínez
José Luis Carballido Carranza
José Raymundo Marcial Romero
Josefa Somodevilla García
Josefina Guerrero García
Juan Manuel González Calleros
Manuel Isidro Martín Ortíz
Maria Aurora Diozcora Vargas Treviño
Maria Auxilio Medina Nieto
Maria de la Concepción Pérez de Celis
Meliza Contreras González
Mireya Tovar Vidal
Omar Flores Sánchez
Rafael de la Rosa Flores
Reyna Carolina Medina Ramírez
Sergio Vergara Limón

Editores

Mireya Tovar Vidal
Claudia Zepeda Cortés
Darnes Vilariño Ayala
Juan Manuel González Calleros
Josefina Guerrero García

Lenguaje, conocimiento y tecnología educativa: nuevos enfoques de aplicación
Coordinado por
Mireya Tovar Vidal
Claudia Zepeda Cortés
Darnes Vilariño Ayala
Juan Manuel González Calleros
Josefina Guerrero García
está disposición en PDF en la página
de la Facultad de Ciencias de la Computación
de la Benemérita Universidad Autónoma de Puebla (BUAP)
https://www.cs.buap.mx/ mtovar/doc/Libros/LibroCDLKE22.pdf
a partir de diciembre de 2022
Peso del archivo: 4.0 MB