Inteligencia artificial, algoritmos y derecho. Una introducción

PID_00270336

Amparo Alonso Betanzos
Verónica Bolón Canedo

Tiempo mínimo de dedicación recomendado: 4 horas

Universitat Oberta de Catalunya

Amparo Alonso Betanzos

Catedrática de Ciencias de la Computación e Inteligencia Artificial de la Universidade da Coruña, donde coordina el grupo LIDIA (Laboratorio de I+D en inteligencia Artificial) del Centro de Investigación en TIC (CITIC). Es Licenciada en Químicas (1984), y doctora en Físicas (1988) por la Universidad de Santiago de Compostela. Ha sido *Postdoctoral Fellow* en el Medical College de Georgia (1988-90), EE. UU. donde trabajó en temas relacionados con el desarrollo de sistemas expertos para aplicaciones médicas.

Su área de investigación es el desarrollo y la aplicación de técnicas de inteligencia artificial en diversas áreas, así como el aprendizaje computacional y las técnicas de ciencia de datos (*big data*). Ha participado y sido investigadora principal en más de 80 proyectos de investigación competitivos y proyectos de transferencia. Es autora de 95 artículos científicos en revistas científicas, 170 artículos en congresos, la mayoría internacionales, así de como 25 libros y capítulos de libros. Recibió en 1998 el Premio *L'Oréal-UNESCO a Women in Science* en España, en 2004 el Premio Galicia TIC a la Innovación, y en 2019 el Premio Galicia TIC a la trayectoria Profesional. Actualmente, y desde 2012, es presidenta de la Asociación Española para la Inteligencia Artificial.

Verónica Bolón Canedo

Ingeniera en Informática (2009) y doctora en Informática (2014) por la Universidade da Coruña (UDC). Después de una estancia posdoctoral en la Universidad de Mánchester (Reino Unido) en 2015, actualmente es profesora ayudante doctora en el Departamento de Ciencias de la Computación y Tecnologías de la Información de la UDC, donde está integrada en el grupo de investigación LIDIA (Laboratorio de I+D en Inteligencia Artificial) del CITIC. Ha impartido docencia en la Universidad de Mánchester y en la Universidad de A Coruña, en asignaturas en su mayoría relacionadas con la inteligencia artificial y el aprendizaje máquina.

Es autora de dos libros, varios capítulos de libro, y más de 60 artículos en congresos internacionales y revistas. Ha coorganizado varias sesiones especiales en congresos internacionales en temas relacionados con *big data*. Su tesis doctoral ha recibido el premio extraordinario de doctorado de la UDC, así como el premio a la mejor propuesta predoctoral (2011), mejor tesis española en IA (2014) y premio Frances Allen a la mejor tesis en IA defendida por una mujer (2015), estos tres últimos premios otorgados por la Asociación Española para la Inteligencia Artificial.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por la profesora: Mònica Vilasau Solana (2020)

Primera edición: febrero 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción

Actualmente nos encontramos inmersos en un cambio social y tecnológico sin precedentes. La digitalización es un proceso intenso y progresivo, que genera grandes cantidades de datos de prácticamente cualquier actividad en la que podamos pensar. La conectividad se ha convertido en una necesidad, que genera retos de adaptación social, pero también inmensas oportunidades de mercado en cualquier campo. Estos datos pueden ser analizados además de forma rápida y económicamente viable debido al abaratamiento de la computación en nube, y a los avances tanto en hardware como en software, para obtener conocimiento útil a partir de ellos. Como consecuencia, la demanda de sistemas inteligentes por parte de las empresas aumenta cada año.

La expansión de estos sistemas inteligentes en diversos dominios está trayendo consigo grandes cambios y no solo de tipo económico, sino también grandes retos de adaptación social que en consecuencia acarrean la necesidad de reglamentación y legislación. Por una parte, habrá cambios en la forma en la que se producen comunicaciones e interacciones en las compañías; por ejemplo, se prevé que un porcentaje importante (sobre el 20 %) de los contenidos publicados por las mismas (documentación legal, comunicados, informes...) serán elaborados por sistemas inteligentes. Se está produciendo un cambio importante en los canales de servicio al cliente, que se gestionarán en un porcentaje muy amplio usando sistemas inteligentes (alrededor del 85 %), la mayoría de ellos en forma de *chatbots*, que se ocuparán de forma más directa y personalizada de los gustos y necesidades de los clientes, a los que pueden atender durante los 365 días del año, durante las 24 horas. Los ejecutivos de las empresas usarán software de reconocimiento de voz para que sus asistentes personales inteligentes les ayuden a organizar su trabajo de manera más eficaz y eficiente. Los vehículos autónomos serán una realidad a corto plazo para el transporte de mercancías, pero quizás también lo sean a medio plazo para nuestros desplazamientos. Estas opciones permitirán una mejora sustancial en la forma de gestionar el tráfico en nuestras colapsadas ciudades, además de mejorar nuestra vida personal y laboral, proporcionándonos tiempo libre, que ahora mismo perdemos conduciendo, intentando buscar aparcamiento o sufriendo atascos. Surgen nuevas áreas de aplicación de la inteligencia artificial (IA), como las relacionadas con las empresas de tecnologías financieras (FinTech), o los seguros relacionados con el empleo de estos sistemas en cualquier ámbito, y muy especialmente en temas de educación o salud.

Obviamente, estos cambios generan la necesidad de más reglamentación y legislación. La mayoría de los países han desarrollado planes estratégicos que no solo se enfrentan al reto tecnológico y a la imprescindible recuperación y fomento del talento humano, sino también a temas relacionados con la ética, la reglamentación y los cambios del modelo de empleo. En concreto, la Unión Europea (UE) propone una visión estratégica basada en una IA ética, sostenible, robusta, confiable y de vanguardia *made in Europe*, que se plasma en el documento de Directrices éticas para una IA confiable. Esta IA se apoya en tres componentes que deben satisfacerse a lo largo de todo el ciclo de vida del sistema:

Lectura recomendada

Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial (2019). *Directrices éticas para una IA fiable*. Comisión Europea.

1) El primer componente es que la IA debe ser lícita, y en consecuencia debe cumplir con las normas vinculantes europeas y de ámbito nacional e internacional que ya son aplicables al desarrollo y uso de los sistemas que utilizan técnicas de inteligencia artificial. Entre ellas se encuentran el derecho primario (tratados de la UE y Carta de Derechos Fundamentales), y el derecho secundario (Reglamento General de Protección de Datos, Directiva sobre máquinas, Directiva sobre discriminación, etc.) de la UE, los Convenios del Consejo de Europa, el Tratado de Derechos Humanos de la ONU, así como normas de carácter nacional en cada uno de los Estados miembros, o sectorial, como puede ser el caso del Reglamento sobre productos sanitarios, entre otros reglamentos y leyes.

2) El segundo componente se refiere a que la IA ha de ser ética, de modo que se garantice el respeto de los principios y valores éticos.

3) Finalmente, el tercer componente establece que debe ser robusta, tanto desde el punto de vista técnico como social, puesto que los sistemas de IA, incluso si las intenciones son buenas, pueden provocar daños accidentales.

Cada uno de estos componentes es en sí mismo necesario pero no suficiente para el logro de una IA fiable. Lo ideal es que todos ellos actúen en armonía y de manera simultánea. Pero no todos los posibles aspectos son susceptibles de ser reglamentados, y en la práctica puede ser relativamente común que aparezcan tensiones entre objetivos diferentes (por ejemplo, la transparencia de los algoritmos puede abrir la puerta al mal uso de los mismos, o identificar y corregir sesgos en los datos podría contrastar con la privacidad). Es importante protegerse de estas situaciones, comunicarlas y documentarlas.

Existen también conexiones entre la Reglamentación y la investigación. En mayo de 2018 entró en vigor el nuevo Reglamento Europeo de Protección de Datos (RGPD), que no solo regula y controla el uso de los datos personales de los europeos, sino que también establece su derecho a requerir una explicación cuando se tomen sobre ellos decisiones que han sido realizadas en todo o en parte por algoritmos inteligentes. La reglamentación ha propiciado la investigación para poder dotar de transparencia y explicabilidad a algoritmos que, hasta ahora, proporcionan respuestas y juicios de forma opaca. Un ejemplo representativo son los disruptivos modelos de aprendizaje profundo (*deep learning*), cuyos desarrolladores recibieron la medalla Turing, patrocinada por Google desde 2014, y que es el equivalente al premio Nobel en este campo. Esos modelos de aprendizaje profundos soy hoy en día un componente crítico de la computación, porque han demostrado producir resultados de vanguardia y de gran exactitud en diferentes tareas. Pero esos modelos tan exactos tienen el problema de comportarse como una «caja negra», es decir, son opacos, ocultando su lógica interna al usuario, y poseen además una complejidad importante. Este aspecto establece una tensión entre la exactitud y la interpretabilidad, y supone no solo un tema de índole práctica, sino también un asunto ético. La capacidad de interpretar un modelo es extremadamente importante, especialmente en algunos campos, como por ejemplo la bioinformática, la medicina personalizada o los vehículos autónomos, ya que en general aumenta la confianza de los usuarios, apoya la comprensión del proceso que se está modelando y proporciona información sobre cómo se puede mejorar el modelo. Es decir, es relevante no solo por temas de confianza, sino también porque puede contribuir al descubrimiento científico y al progreso de la investigación en diversos campos. Ahora mismo, es relevante también en el contexto de la RGPD, que entre otros, como mencionamos anteriormente, legisla el derecho a obtener una explicación entendible cuando una persona recibe una decisión de un sistema basado en IA, por ejemplo, respecto a la concesión de un crédito, de asilo político, de un seguro, etc.

Lecturas recomendadas

Parlamento Europeo y Consejo de la Unión Europea (2018). *Reglamento General de Protección de Datos*. Diario Oficial de la Unión Europea.

Y. LeCun; Y. Bengio; G. Hinton (2015). «Deep learning». *Nature* (521, págs. 436-444).

Existe un acuerdo general en que la necesidad de implementar este principio es urgente y que representa en este momento un reto científico muy importante. Si la tecnología no es capaz de proporcionar esta explicación, este derecho se convierte en papel mojado. Otro riesgo importante que aparece también es la posibilidad de que estos modelos tomen decisiones equivocadas inadvertidamente, debido por ejemplo a que los datos que han usado contienen sesgos y prejuicios. Un ejemplo de esto último es el caso del escándalo que ocurrió recientemente con la herramienta de contratación de Amazon, cerrada en 2017 porque discriminaba a las mujeres, confundiendo su escasez en el área TIC (tecnologías de la información y las comunicaciones) con incapacidad en el trabajo. Las tecnologías de explicación que debemos desarrollar son necesarias para que las compañías creen productos más seguros, más confiables y que puedan manejar cualquier responsabilidad legal en la que puedan incurrir. Pensemos, por ejemplo, en su utilidad en el caso del reciente accidente del coche autónomo de Uber, en el que el fallo se debió a un ajuste de baja sensibilidad ante los objetos de la carretera. El objetivo era que el coche pu-

diera reaccionar y tomar acciones ante objetos sólidos de gran tamaño y no ante objetos inocuos, como bolsas por ejemplo. De esta forma el peatón fue detectado, pero el software decidió no parar, con la grave consecuencia de una persona muerta. Otro problema es que los modelos de redes neuronales profundas (DNN-*deep neural networks*) son capaces de lograr exactitudes importantes en el reconocimiento de texto e imágenes, incluso por encima del nivel humano, pero a la vez también son susceptibles de ser engañados por ataques llamados *adversarial* o de adversarios, que introducen variaciones que pasarían inadvertidas a los humanos, pero que son suficientes para engañar a un clasificador. Por estos y otros motivos, la explicabilidad de los algoritmos es hoy un tema importante en investigación. Por el momento, la mayoría de los escasos métodos desarrollados hasta ahora son locales, es decir, nos proporcionan las razones de una decisión específica. Solamente es interpretable una decisión concreta. Otras limitaciones a tener en cuenta son temporales, ya que si es necesario tomar una decisión rápida (por ejemplo, en un entorno en el que un desastre es inminente), es preferible un modelo de explicación simple para que el usuario humano pueda tomar una decisión. Si el tiempo no es una restricción, por ejemplo, en un problema de concesión de un crédito, puede ser viable un modelo explicativo más complejo y exhaustivo. Otros aspectos a considerar en el tema de la explicabilidad tienen que ver con asuntos éticos, como es el caso de la justicia, garantizando la protección de ciertos grupos frente a la discriminación; y la privacidad, logrando que el modelo de explicación no revele información sensible sobre personas. Finalmente, y entre otros aspectos deseables, la escalabilidad y la portabilidad de los modelos, tanto los de aprendizaje como los de explicación, si fuesen diferentes, son del máximo interés. En muchos de los modelos de explicación se vuelve la mirada hacia los modelos simbólicos, más interpretables en el ámbito local o global, como es el caso de los árboles de decisión o las reglas.

Asimismo, es necesario que vayamos más allá en temas de regulación y de ética; es necesario legislar en áreas sensibles, como la educación o la salud, pero también fomentar un debate social y político de aspectos también controvertidos de la inteligencia artificial, como es el caso de las noticias falsas, las armas autónomas o el uso de sistemas de reconocimiento facial. También será necesario regular temas que afectan a los modelos de transición de nuestro actual modelo productivo, más basado en la mano de obra humana, y que debe cambiar para afrontar los nuevos modelos de trabajo del futuro más basados en la automatización de muchas tareas rutinarias. Algunas cuestiones se han debatido ya en la UE, y algunos de los países miembros incluso han probado algunas medidas, como la renta básica universal. Las conclusiones del experimento que sobre la implantación de la misma se llevó a cabo en Finlandia, el primer país del mundo en probar la medida a principios del año 2017, son que su uso no ofrece mejores perspectivas laborales, pero sí aumenta considerablemente el nivel de bienestar y sensación de justicia social que perciben los ciudadanos. El experimento escogió una muestra relativamente pequeña de población, dos mil personas desempleadas y en la franja de edad de 25-58 años, a las que se les concedían 560 euros mensuales libres de impuestos sin

tener en cuenta si estaban o no buscando trabajo de forma activa. Los resultados son que los beneficiarios de la renta básica no tuvieron diferencias con los demás ciudadanos a la hora de encontrar trabajo en un mercado laboral abierto, pero tuvieron menos síntomas de estrés, menos dificultad para concentrarse y menos problemas de salud, además de una mayor confianza en el futuro y capacidad para influir en los problemas sociales. ¿Cuál es el problema? La Organización para la Cooperación y el Desarrollo (OCDE) concluyó, en su informe llevado a cabo en cuatro hipotéticos escenarios en Reino Unido, Francia, Italia y Finlandia, que es imposible implantar un sistema de estas características sin una reforma sustancial del modelo tributario, pues si se financiase con el presupuesto actual destinado a ayudas sociales su generalización podría incrementar la pobreza desde el 11,4 % actual al 14,1 %, ya que requeriría aumentar los impuestos a los salarios en casi un 30 %.

La entrada en el mundo del derecho de la inteligencia artificial es ya un hecho en la actualidad, y plantea muchos retos interesantes. Vamos a ver a continuación algunos de sus aspectos básicos, para ayudarnos a entender mejor su relevancia.

1. Historia de la inteligencia artificial

La inteligencia artificial (IA) es un área de las ciencias de la computación, y por tanto de la informática, que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico.

Dentro de la disciplina, se encuentran varias subáreas de trabajo, como pueden ser:

- la robótica,
- el aprendizaje automático,
- el procesamiento del lenguaje natural o
- la visión artificial.

La IA es además cada vez más interdisciplinar, ya que afecta a muchos campos de aplicación, y afecta a numerosos aspectos de la sociedad, por lo que cada vez implica e incorpora a profesionales procedentes de otras disciplinas, como la economía, la filosofía, las matemáticas o el derecho, entre otras.

La IA se conoce con este nombre desde el año 1956, en el que un reducido grupo de científicos (procedentes de distintas disciplinas, como el control, la cibernética o la teoría de autómatas) se reunió en un seminario de verano en Dartmouth College (Estados Unidos) para reflexionar sobre la pregunta «¿pueden pensar las máquinas?», que había enunciado el científico británico Alan Turing, considerado padre de la inteligencia artificial, en un artículo científico publicado en la revista *Mind* en 1950. En este artículo, Turing defendía que cabría esperar que las máquinas puedan competir con los seres humanos en campos puramente intelectuales. Muchas de las ideas que planteaba Turing existen actualmente como modelos de aprendizaje automático (supervisado, no supervisado, evolutivo, por refuerzo), o como modelos usados en robótica (la robótica del desarrollo, que adopta ideas del aprendizaje humano para acortar las escalas de tiempo del aprendizaje automático), mientras que otras, como la creatividad computacional o las interfaces humanizadas, son campos en auge que todavía presentan mucho territorio por explorar.

Lectura recomendada

A. M. Turing (1950). «Computing machinery and intelligence». *Mind* (59, págs. 433-460).

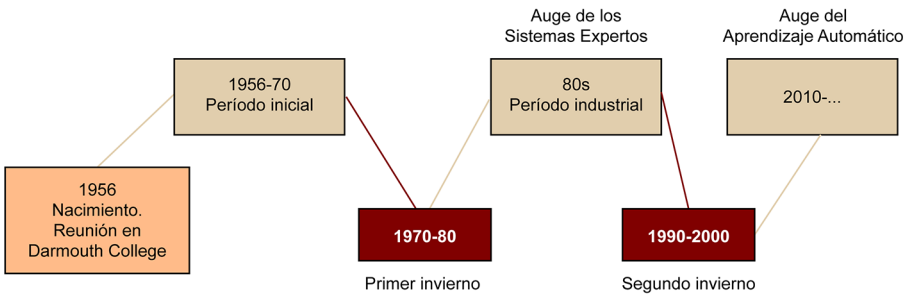
En esa reunión en el Darmouth College, financiada en parte por IBM, los científicos asistentes comienzan ya a plantearse cómo sería posible construir estas máquinas inteligentes y son los cuatro principales organizadores, John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon, quienes escogen el nombre *inteligencia artificial* para denominar este campo nuevo. En esta reunión, la propuesta asumía que cualquier aspecto del aprendizaje o cual-

quier otra característica de la inteligencia podría ser descrito de forma que **una máquina pudiese simularlo**. El objetivo consistía en conseguir avances significativos en áreas como:

- la comprensión y uso del lenguaje natural en las máquinas,
- la construcción por parte de estas de abstracciones y conceptos,
- la resolución de problemas complejos,
- las redes neuronales,
- el aprendizaje o
- la creatividad computacional.

Algunos de los investigadores asistentes fueron McCulloch y Pitts (con su modelo neuronal del cerebro), Von Neumann (proyectos para la construcción del computador ENIAC y del programa EDVAC), Minsky y Edmonds (primer computador neuronal), y Shannon (heurística, programas de ajedrez). Aunque solo hubo diez trabajos en el seminario, los ponentes y sus discípulos marcaron el campo de la IA durante los siguientes treinta años.

Figura 1. Línea de tiempo del desarrollo de la disciplina de inteligencia artificial



Fuente: elaboración propia

Todas las cuestiones que se plantearon en esta reunión se rebelaron de gran calado, y han estado presentes como grandes retos del campo desde su nacimiento. En el seminario, se abordaba un paradigma cognitivo del cerebro humano, que tenía el objetivo de **reproducir las capacidades cerebrales mediante algoritmos**, bajo la premisa de que los estados mentales podrían funcionar de manera análoga a los programas en una computadora. En cuanto a cómo construir estos algoritmos, aparecieron dos líneas diferentes de pensamiento:

- Un grupo, que recibió el nombre de «pulcros», mantenía que la representación simbólica se basaba principalmente en la lógica (es decir, en la sintaxis y el cálculo de predicados).
- El otro grupo, los llamados «desaliñados», pensaba que se basaba principalmente en la semántica.

La primera de las aproximaciones, la «pulcra», optó por soluciones elegantes y cercanas a las matemáticas, mientras que la segunda, los «desaliñados», pensaban que la inteligencia es demasiado complicada, probablemente computacionalmente intratable y, por lo tanto, no se podría resolver con el tipo de

sistema homogéneo que exige requisitos precisos. Y de la mano de cada una de estas dos corrientes ha venido cada una de las dos etapas más brillantes (llamadas primaveras) de la IA (ved línea temporal en la figura 1).

La primera primavera surge de la mano de la aproximación simbólica, relacionada con la rama «desaliñada», con el nacimiento de los llamados «sistemas expertos», que demostraron la factibilidad de usar inteligencia artificial para resolver problemas del mundo real con sistemas como DENDRAL, MYCIN o PROSPECTOR, mientras que la segunda primavera ha llegado con los nuevos enfoques de aprendizaje profundo, el reconocimiento del lenguaje natural y otros enfoques del aprendizaje automático, como el aprendizaje por refuerzo, provenientes de la rama «pulcra». Quizás el futuro nos depare una fusión entre ambas para hacer frente a los nuevos retos que plantea la inteligencia artificial confiable y robusta, con procesos en los que se garantice la transparencia, la reversibilidad y la trazabilidad.

Durante la primera época, tras su nacimiento, la disciplina vive un momento de esplendor, gracias a los avances que se producen en campos como la abstracción de conceptos mediante aprendizaje (desarrollando programas que eran capaces de demostrar algunos teoremas sobre lógica proposicional –como por ejemplo el Logic Theorist–, o sobre geometría), modelos de cómo los humanos resolvemos problemas (como el General Problem Solver: GPS), o métodos de aprendizaje para emplear en áreas de juegos, como el ajedrez o las damas. Pero tras este entusiasmo inicial, comienzan a aparecer, a principios de los años setenta, las primeras críticas a la inteligencia artificial, ya que los algoritmos desarrollados tenían serios problemas de escalabilidad a la hora de poder ser usados en entornos reales complejos, y también contribuye a esta caída en desgracia el relativo fracaso del procesamiento del lenguaje natural, que se asumía como medio de comunicación lógico entre la máquina y el ser humano. La capacidad de los ordenadores del momento estaba muy lejos de permitir la implementación de muchas de las ideas de estos pioneros de la IA.

Ejemplo: la capacidad de los ordenadores

Para hacernos una idea, las aplicaciones prácticas de visión por computadora requieren alrededor de 10.000 a 1.000.000 procesadores MIPS, una cifra mucho mayor que los 80-130 MIPS del supercomputador más rápido del mundo en 1976, Cray-1, un gigante a su vez comparado con un ordenador personal típico de la época, con menos de 1 MIPS.

Aun así, se consiguen también éxitos en esta época, entre los cuales se encuentran los trabajos de Fikes y Nilsson en 1971, que dieron como resultado:

- el sistema de planificación STRIPS (cuyas ideas están vigentes todavía hoy);
- los de Newell y Simon en 1963, que desembocaron en el GPS o solucionador general de problemas, que imitaba protocolos de resolución de problemas de los humanos en la secuencia de objetivos y posibles cursos de acción que generaba;
- el programa de las damas de Samuel que consiguió batir a su creador;
- los trabajos de McCarthy, que definió el lenguaje LISP e inventó el tiempo compartido para conseguir más recursos de cómputo.

Lecturas recomendadas

R. E. Fikes; N. J. Nilsson (1971). «Strips: A new approach to the application of theorem proving to problem solving». *Artificial Intelligence* (núm. 2, págs. 189208).

A. Newell; H. Simon (1963). *GPS, A Program That Simulates Human Thought*. Nueva York: McGraw-Hill.

También se realizaron trabajos importantes que, en dominios limitados, resolvían problemas para los que se requería inteligencia, tales como problemas de álgebra, geometría o cálculo. A pesar de ello, a finales de los años sesenta, la IA se encontró con muchos problemas:

1) Los métodos se habían desarrollado para resolver problemas de carácter muy general, y los programas contenían poco o ningún conocimiento sobre el dominio del problema a modelar. Esto no supone grandes problemas en dominios simples, pero sí en dominios reales en los que los problemas son más variados o de mayor dificultad.

Problemas NP-completos

En un principio, se creyó que la solución sería escalable, pero no resultó así, ya que la mayoría de los problemas interesantes para la IA son NP-completos (de forma sencilla, los problemas que son solubles en tiempo polinomial son tratables y, los que no lo son, son NP-completos, y suelen conocerse como intratables. Este tipo de problemas suelen resolverse mediante aproximaciones).

2) Un campo interesante para la IA en aquel momento era (y continuará siendo) el lenguaje natural, y en especial los traductores automáticos. En este campo, las dificultades aparecían por la falta de un contexto particular y de nociones generales sobre situación y sentido común. Esto provocó la cancelación económica de Estados Unidos de estos proyectos en 1966.

3) En Europa, los científicos más destacados en IA fueron principalmente británicos. En 1971, el Gobierno británico canceló el soporte económico a la IA apoyándose en el *Informe Lighthill*, que concluyó que las aportaciones de esta disciplina no eran relevantes, cortando la financiación en esta área.

La disciplina entra así en su primer invierno, un período de desánimo y de cierta sensación de fracaso, que además trae consigo un fuerte recorte económico. Este primer invierno termina en los años ochenta, en los que surgen con fuerza un tipo de sistemas de IA, llamados **sistemas expertos**, que se adoptan con éxito en muchos entornos reales. Este éxito se basa en un cambio de orientación, se olvida el objetivo de lograr una inteligencia artificial de tipo general, similar a la humana, y se cambia el foco a una IA específica, centrada en campos concretos. La idea era que algunos de los problemas identificados en los métodos que se manejaban por entonces podrían resolverse si se introducía en los sistemas, además del conocimiento general que permite realizar inferencias, conocimiento específico sobre el dominio de aplicación. Esta pos-

tura llevó a los investigadores a afirmar que el conocimiento podía adquirirse de los expertos humanos para ser «transferido» a un computador a través de una representación que este pudiera manipular.

Nacen así los sistemas expertos, que se pueden definir como sistemas cuyo propósito es emular la capacidad de resolución de problemas de un ser humano en un dominio específico, utilizando su mismo conocimiento.

Este primer éxito es mérito de aproximaciones simbólicas, que usan sobre todo modelos basados en reglas de producción que codifican el comportamiento de expertos humanos para resolver problemas concretos. Tras este período fructífero la IA cae en un segundo invierno, debido a la dificultad que planteaba el mantenimiento de estos sistemas expertos especializados y al colapso del mercado hardware especializado (necesario hasta entonces para el desarrollo de los sistemas expertos), debido a la impresionante mejora de prestaciones de los ordenadores de propósito general que conocemos hoy. Tras este invierno, la IA comienza a despertar de nuevo, fragmentándose en campos especializados (sistemas multiagente, robótica, sistemas basados en conocimiento, etc.), que obtienen ciertos éxitos importantes.

Deep Blue

Un ejemplo ilustrativo puede ser que, en el año 1997, Deep Blue se convierte en el primer programa de ajedrez que derrota a un vigente campeón del mundo, Kasparov. Hay que tener en cuenta que esta victoria no se debe tanto a la mejora de la aproximación inteligente como a la mejora del hardware, ya que este nuevo Deep Blue era una versión mejorada de un sistema anterior, que fue derrotado por Kasparov, pero que ahora era capaz de procesar el doble de movimientos por segundo que en la primera versión.

No es hasta alrededor del año 2010 cuando la IA salta a la industria como una tecnología imprescindible, debido a la conjugación de varios factores que veremos a continuación:

1) En primer lugar, **disponemos de enormes cantidades de datos**, debido al progresivo e intenso proceso de digitalización que está teniendo lugar. Prácticamente cualquier experiencia humana está digitalizada: los viajes, la música, los servicios de salud, etc. Además, disponemos también de una creciente cantidad de sensores, cada vez más exactos, que registran datos procedentes de prácticamente cualquier proceso, y que por lo tanto nos permiten saber cómo se comporta y evoluciona nuestro entorno.

2) Además, el cambio social y tecnológico por el que estamos atravesando no tiene precedentes. Todos nosotros interactuamos con nuestros varios dispositivos móviles en un **mundo cada vez más interconectado**, y en el que el disponer de conectividad a cualquier hora y en cualquier lugar es imprescindible. Esta conectividad genera tanto una oportunidad de mercado para las empresas como un reto de adaptación social.

3) En tercer lugar, además ahora disponemos de la **capacidad de cómputo** necesaria para que los programas que diseña la IA puedan alcanzar soluciones de gran exactitud en un tiempo adecuado. Diversos elementos permiten el procesado rápido y económicamente rentable de enormes volúmenes de datos heterogéneos:

- el avance de las tecnologías de computación de altas prestaciones,
- el abaratamiento de la computación en nube, y
- la disponibilidad de uso de nuevas plataformas de computación paralela y distribuida.

4) También es decisivo el hecho de que disponemos de **avances muy relevantes en software**. Hemos conseguido desarrollar nuevos tipos de bases de datos que nos permiten almacenar y tratar datos estructurados y no estructurados, más allá del clásico dato científico de las bases de datos clásicas. Al mismo tiempo ha sido disruptiva la aparición de nuevos desarrollos teóricos, fundamentalmente matemáticos, como los que se han obtenido en el campo del aprendizaje profundo (*deep learning*), el aprendizaje por refuerzo, o el reconocimiento del lenguaje natural, que han generado resultados de alto nivel de precisión, situando la IA como una tecnología madura de éxito y de gran impacto.

5) Como consecuencia de todo lo anterior, **aumenta la demanda real de las empresas**, que multiplican anualmente sus inversiones económicas en el campo. Esta inversión ayuda también a que la tecnología de IA se consolide, ya que podemos constatar que aquellas empresas que son usuarias o desarrolladoras de la tecnología son las más exitosas en sus respectivos campos. En el año 2019, siete de las diez empresas más importantes del mundo en capital bursátil son usuarias o desarrolladoras de esta tecnología. Esta situación era muy diferente tan solo diez años atrás, cuando nada más que tres de las diez primeras empresas del mundo pertenecían al sector.

Inversión en inteligencia artificial

La inversión global en nuevas empresas de inteligencia artificial se ha multiplicado por nueve entre los años 2011 y 2015, según el Foro Económico Mundial, y ha seguido creciendo aún más desde entonces.

Pero además, a través de su historia, la IA se ha convertido en una disciplina transversal, que afecta a muchos campos de aplicación, como la salud y medicina, la educación, el medio ambiente, la industria, el turismo, etc. Como consecuencia, la inteligencia artificial es un campo cada vez más interdisciplinar, incorporando profesionales procedentes de otras disciplinas como el derecho, la psicología, la sociología, la economía, etc.

En la reunión de Darmouth College, se estableció también que el objetivo último de la inteligencia artificial era conseguir reproducir inteligencia en una máquina, de manera que esta, adecuadamente programada, pudiera replicar la inteligencia humana y exhibir un comportamiento inteligente de tipo general. Este es todavía un objetivo muy ambicioso, que parece que estamos aún lejos de conseguir. Pero sí que hemos podido desarrollar una IA específica, en la que algoritmos y máquinas son capaces de realizar tareas asociadas a la inteligencia

humana como aprender, entender o razonar, y que requieren inteligencia en un ámbito concreto y especializado, pero que no exhiben un comportamiento inteligente de tipo general.

Ejemplo: un campo de juegos

Hemos conseguido desarrollar programas que juegan al ajedrez o al Go, y que son capaces de batir a los campeones del mundo y a los grandes maestros. Pero no tienen una inteligencia general; si cambiamos las reglas del juego, o pretendemos jugar a un juego similar, necesitaríamos en principio desarrollar un algoritmo distinto (aunque se está trabajando intensamente en el tema de la transferibilidad del aprendizaje). Sin embargo, cualquier jugador humano de ajedrez es capaz de aprovechar sus conocimientos para aprender a jugar de forma rápida a las damas, por ejemplo.

A pesar de esta limitación, existen muchos dominios en los que la IA supera a la inteligencia humana, como es el caso de áreas específicas de la medicina, los sistemas de recomendación, los robots, los vehículos autónomos, los asistentes personales o los traductores automáticos. Todos estos sistemas están cambiando nuestra forma de interaccionar con el entorno, y traen consigo importantes cambios económicos y sociales, que necesitan de regulación y reglamentación.

2. Ramas de la inteligencia artificial

Existen varias clasificaciones posibles que podemos usar para hablar de los diferentes campos de la inteligencia artificial. Vamos a ver algunas de estas divisiones, y así también vamos definiendo algunos conceptos básicos. Una forma de clasificación clásica es distinguir entre la inteligencia artificial simbólica y la inteligencia artificial subsimbólica.

Lectura recomendada

S. Russel; P. Norvig (2018). *Artificial Intelligence: A Modern Approach*. Pearson.

2.1. Inteligencia artificial simbólica

También llamada simbólico-deductiva o convencional.

Esta rama surge desde los inicios de la disciplina, y se basa en la idea que se define en la hipótesis de sistemas de símbolos físicos, de Newell y Simon, que básicamente defiende que la mayoría de los aspectos de la inteligencia se pueden modelar usando representaciones de alto nivel simbólicas de los problemas a modelar, usando como herramientas la lógica matemática y la búsqueda.

El paradigma simbólico fue el dominante desde los años cincuenta hasta los ochenta, y produjo el exitoso paradigma de los sistemas expertos, que como hemos visto marca la llegada de la primera de las primaveras que ha vivido la inteligencia artificial. Por sus características, los sistemas expertos podían aplicarse a muchas áreas de la actividad humana, y de ahí su gran éxito, siendo uno de los principales campos de experimentación y aplicación el de la medicina, pero después se han extendido a prácticamente cualquier área de la actividad humana en la que se necesite conocimiento experto, y entre ellas estaría, cómo no, el de las varias ramas del derecho. Los sistemas expertos se definen como sistemas inteligentes que contienen conocimiento explícito de alto nivel de un campo de aplicación complejo, aunque restringido. La forma en la que este conocimiento explícito se representa está basado en la lógica, y el paradigma de representación más utilizado son las **reglas de producción**.

Lectura recomendada

A. J. González; D. D. Dankel (1993). *The Engineering of Knowledge-based systems* . Prentice-Hall.

Las reglas de producción son elementos que tienen una estructura con un antecedente (parte condición de la regla, o parte *if*) y un consecuente (parte *then* o parte conclusión de la regla), y dependiendo del lenguaje de programación usado pueden tener una parte de conclusión alternativa (parte *else*), que se deduciría si fallase la parte antecedente, es decir, si no se cumplen las cláusulas. Cualquiera de las partes puede anidar varias cláusulas en la forma:

REGLA X

IF condición1 AND condición 2 AND...condición n
THEN conclusión 1 AND conclusión 2 AND... conclusión m
ELSE conclusión 1' AND ...conclusión k'

donde las cláusulas pueden anidarse mediante operadores:

- AND (todas deben cumplirse),
- OR (es suficiente con que se cumpla una de ellas), y
- NOT (es la negación de la condición la que debe ser cierta para que se cumpla la cláusula).

Además de conclusiones sobre posibles hipótesis, las partes consecuente también pueden contener acciones, que se llevarían a cabo si la parte antecedente se cumple.

Ejemplo de regla

Veamos un ejemplo de una regla en un sistema experto relacionado con el diagnóstico clínico del estado fetal cuando una paciente embarazada se somete a un test antenatal. La idea es que el experto clínico analiza varios parámetros y, en función de su valor, decide un estado de normalidad/anormalidad fetal, que iría acompañado de las correspondientes actuaciones posteriores:

REGLA 12

IF Linea-de-base-cardiaca fetal=normal
AND Variabilidad-frecuencia-cardiaca-fetal= normal
AND aceleraciones-frecuencia-cardiaca-fetal>=3
AND deceleraciones-frecuencia-cardiaca-fetal=0
THEN estado-fetal= normal

REGLA 7

IF desviación estándar-frecuencia-cardiaca-fetal < 2
THEN Variabilidad- frecuencia-cardiaca-fetal=ausente

REGLA 8

IF desviación estándar-frecuencia-cardiaca-fetal > 2
AND desviación estándar-frecuencia-cardiaca-fetal < 5
THEN Variabilidad-frecuenciacardiaca-fetal= decrecida

REGLA 9

IF desviación estándar-frecuencia-cardiaca-fetal >= 5
AND desviación estándar-frecuencia-cardiaca-fetal < 10
THEN Variabilidadfrecuencia-cardiaca-fetal= normal

REGLA 3

IF latidos-minuto-fetal >= 120
AND latidos-minuto-fetal <=160
THEN Línea-de-base-cardiaca fetal=normal

De esta forma se pueden codificar áreas de conocimiento experto de alto nivel, cuyas reglas pueden establecer cadenas de razonamiento o inferencias, en dos modos diferentes:

- modo de encadenamiento progresivo (desde los datos de un caso concreto hacia las conclusiones), o
- modo de encadenamiento regresivo (estableciendo una hipótesis inicial, y comprobando si los datos del problema nos permiten establecer esa hipótesis como correcta).

Ejemplo: encadenamiento progresivo y encadenamiento regresivo

Relacionado con las reglas anteriores, en el caso de un encadenamiento progresivo, trabajaríamos encadenando las reglas desde los valores de los parámetros hasta el diagnóstico. Así, por ejemplo, si tenemos como valores disponibles del caso que los latidos por minuto fetales son 128, la desviación estándar de la frecuencia cardíaca es 7, no hay deceleraciones y hay 4 aceleraciones, las reglas que se activarían serían la 3, la 9 y la 12, concluyendo con esa cadena de razonamiento que el estado fetal es normal.

Pero también podemos trabajar en modo regresivo, comenzando con una hipótesis de partida. Por ejemplo, podemos suponer que el estado fetal es normal (lo establecemos como hipótesis de partida), y luego vamos analizando qué parámetros serían necesarios para establecer esa hipótesis como cierta. Cuando esos parámetros se averigüen, el proceso regresivo pasa a un proceso similar al anterior, desde los parámetros a las conclusiones, para establecer la veracidad de la hipótesis de partida. En este caso, la hipótesis que planteamos está en la parte de conclusión de la regla 12, y para poder establecer esa conclusión como cierta, necesitaríamos conocer los valores de los parámetros de su parte IF o condición, que en este ejemplo son línea-de-base-cardiaca-fetal, variabilidad-frecuencia-cardiacafetal, aceleraciones-frecuencia-cardiaca-fetal y deceleraciones-frecuenciacardiaca-fetal. Como no sabemos los valores de esos parámetros, debemos averiguarlos para poder establecer la hipótesis como cierta.

Comencemos con el primer parámetro; tendríamos que comprobar que «línea-de-base-cardiaca-fetal=normal», este parámetro desconocido por ahora está formando parte de la conclusión de la regla 3. Establecemos esta situación como nueva subhipótesis y, para comprobarla, debemos saber de nuevo si se cumple la parte IF de la misma regla. En esta regla 3, habría que conocer el valor de latidos-minuto-fetal. Supongamos que podemos conocerlo, y que es 130. Entonces, el proceso evocativo se convierte en progresivo, y nos permite establecer la conclusión de la regla 3. Pero aún nos quedan los demás parámetros de la regla 12 para comprobar, con los que iríamos realizando un proceso análogo, configurando un proceso evocativo (regresivo), hasta que una de las reglas nos dé un valor para el parámetro hipotetizado, y nos permita cambiar el sentido del encadenamiento hacia delante, puesto que todos los valores de los parámetros involucrados son conocidos, y en el caso de que se ajusten a los valores esperados, podríamos establecer la hipótesis inicial como cierta (conclusión del razonamiento). Si alguno de los parámetros no tiene el valor especificado en la regla, entonces la hipótesis no puede establecerse como conclusión. Por ejemplo, si seguimos con el caso anterior, y comprobamos la veracidad del segundo parámetro, «Variabilidad-frecuencia-cardiaca-fetal= normal», veríamos que es la conclusión de la regla 9, y para poder establecer esa subhipótesis como verdadera, necesitaríamos conocer el valor de los parámetros de la parte IF de esa regla, en este caso «desviación estándar-frecuencia-cardiaca-fetal». Si ese valor es por ejemplo 4, la regla 9 no se cumple, sino que se cumpliría la 8, y en este ejemplo por lo tanto lo que ocurre es que la regla 2 no se cumple, y la hipótesis de partida no puede establecerse como cierta.

Durante los años ochenta, los sistemas expertos supusieron un enorme éxito de la inteligencia artificial, que por primera vez podía trabajar en campos reales de aplicación. Estos sistemas se aplicaron en multitud de dominios, y debido a su importancia histórica, cabe citar:

1) DENDRAL (Buchanan, Feigenbaum, Lederberg, Stanford University, 1969) fue el primer sistema experto con éxito. La meta del proyecto era desarrollar un programa capaz de obtener el mismo nivel de rendimiento de un químico experto en la determinación de las estructuras moleculares basándose en datos procedentes de un espectrógrafo de masas. El proyecto estaba financiado por la NASA para usarlo en una nave espacial enviada a Marte y determinar la estructura molecular del suelo. La primera versión del programa generaba todas las posibles estructuras que se podrían corresponder con la molécula, predecía para cada una de ellas las observaciones del espectro de masas y las comparaba con el espectro real. El problema fundamental era que se podrían generar millones de estructuras posibles. Observaron, entonces, que los químicos muy experimentados podían reducir el número de candidatos posibles a una cantidad tratable mediante el uso de heurísticas adquiridas con la experiencia; por ello, se decidió consultar con químicos analíticos y tratar de simular el comportamiento de estos expertos en el sistema. De esta forma, el número de estructuras candidatas se reducía considerablemente, y DENDRAL se convirtió en el primer sistema intensivo en conocimiento que conseguía funcionar en un entorno real complejo. Este éxito marcó un cambio drástico en la investigación en IA, que pasó de centrarse en los métodos de propósito general, con escaso conocimiento del campo de aplicación y con métodos débiles de búsqueda (es decir, métodos que no usan ningún conocimiento específico del problema concreto para encontrar soluciones), a hacerlo en técnicas específicas del dominio e intensivas en conocimiento. Además, después de esta experiencia se inició el Proyecto de Programación Heurística en la Universidad de Stanford, encaminado a analizar cómo esta nueva metodología podía aplicarse en otras áreas. Además de los primeros sistemas expertos, surge así también la disciplina conocida como ingeniería del conocimiento (IC), que engloba la captura, el análisis y la implementación del conocimiento experto en un programa de computador.

2) El siguiente proyecto importante, desde el punto de vista histórico, y que supuso un éxito en este caso en el campo del diagnóstico médico, fue **MYCIN** (Feigenbaum, Shortliffe, Stanford University, 1972). MYCIN era un sistema basado en reglas de producción para el diagnóstico y la terapia de enfermedades sanguíneas infecciosas. Consiguió niveles de comportamiento tan buenos como los de los médicos expertos. La característica especial que diferenciaba a MYCIN de DENDRAL era que su conocimiento en forma de reglas estaba claramente separado de su mecanismo de razonamiento, por lo cual el sistema era fácilmente extensible y actualizable mediante la inserción o el borrado de reglas. Las reglas de MYCIN se extraían de la experiencia médica, mientras que las reglas de DENDRAL se derivaban de modelos teóricos. Además, las reglas de MYCIN incorporaban incertidumbre mediante un esquema novedoso llamado factores de certeza. El razonamiento bajo incertidumbre, una forma natural de razonar en cualquier experto humano, era una de las partes más importantes del sistema.

Lecturas recomendadas

Staff of the Heuristic Programming Project (1980). «The Stanford heuristic programming project: Goals and activities». *Artificial Intelligence Magazine* (núm. 1, págs. 25-30).

G. Schreiber; H. Akkermans; A. Anjewierden; R. de Hoog; N. Shadbolt; W. van de Velde; B. Wielinga (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press.

A. Alonso Betanzos; B. Guijarro Berdiñas; A. Lozano Tello; J. T. Palma Méndez; M. J. Taboada Iglesias (2004). *Ingeniería del Conocimiento: Aspectos metodológicos* . Pearson.

3) Otro sistema que generó mucha publicidad fue **PROSPECTOR** (Duda, Stanford University, 1979), un sistema inteligente probabilístico para la explotación mineral. Para la representación del conocimiento, el sistema usaba un esquema híbrido que incorporaba reglas y otra forma de representación del conocimiento llamada redes semánticas. PROSPECTOR incorporaba el tratamiento de la incertidumbre usando un esquema de tipo probabilístico, el esquema bayesiano. El sistema era capaz de trabajar al nivel de un geólogo experto, y cuando encontró un depósito de molibdeno (Mb) en el estado de Washington de un valor de más de 100 millones de dólares, no se pudo encontrar mejor justificación para su utilización.

4) XCON. Desarrollado conjuntamente por Digital Equipment Corporation y la Universidad de Carnegie-Mellon, ayudaba en la configuración de nuevos pedidos de computadores VAX. Fue utilizado internamente por la compañía Digital.

Lectura recomendada

V. Barker; D. O'Conner (1989). «Expert systems for configuration at digital: XCON and beyond». *Communications of the ACM* (vol. 32, núm. 3, págs. 298-318).

5) LES. Fue desarrollado por MITRE Corporation y NASA-KSC (NASA Kennedy Space Centre). Su misión era la de monitorizar y diagnosticar los procesos de carga de oxígeno líquido en el tanque principal del transbordador espacial.

Lectura recomendada

E. A. Scarl; J. R. Jamieson; C. I. Delaune (1987). «Diagnosis and sensor validation through knowledge of structure and function». *IEEE Transactions on Systems, Man, and Cybernetics* (vol. 17, núm. 3, págs. 360-368).

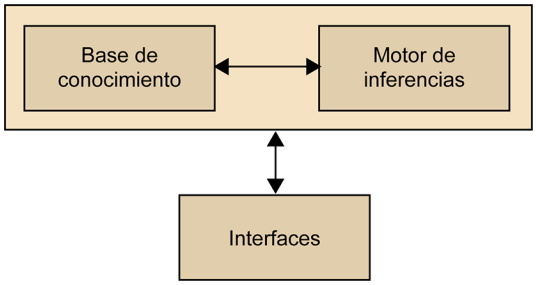
6) INTERNIST. Fue desarrollado en la Universidad de Pittsburgh. En su fecha, fue el sistema médico con mayor número de reglas y su misión era la de asistir al médico en la elaboración de diagnósticos múltiples y complejos relacionados con la medicina interna.

Lectura recomendada

R. A Miller: H. E. Pople Jr.; J. D. Myers (1982). «Internist-1, an experimental computer-based diagnostic consultant for general internal medicine». *New England Journal of Medicine* (307, págs. 468-476).

Estos sistemas proponían la organización de una aplicación en dos componentes: un componente declarativo, denominado base de conocimiento, que representa lo que conoce un agente especializado en un determinado tipo de problemas, y un procedimiento diseñado para simular un proceso de razonamiento (ved figura 2).

Figura 2. Arquitectura general de un sistema experto.



Fuente: elaboración propia

La base de conocimiento contiene los hechos y las reglas del dominio que se están modelando. El motor de inferencias contiene los modelos de razonamiento utilizados, entre ellos los mencionados mecanismos de encadenamiento progresivo y regresivo. Las interfaces son necesarias para interaccionar con los usuarios, otros equipos hardware o programas software, etc.).

Lectura recomendada

D. H. Berman; C. D. Hafner (1989). «The potential of artificial intelligence to help solve the crisis in our legal system». *Communications of the ACM* (núm. 32, págs. 928-938).

También existen algunos sistemas expertos «clásicos» en el mundo del derecho, algunos de los cuales serían:

1) Taxman II, que contiene conocimiento jurídico del ámbito del derecho fiscal en Estados Unidos, en concreto, la tributación de reorganizaciones corporativas. El sistema actúa en un papel asesor, de forma que acepta la descripción de un caso que introduce un usuario en forma de hechos declarados siguiendo un procedimiento estándar, y produce un análisis del mismo a un nivel de abstracción adecuado a la conclusión legal deseada.

Lectura recomendada

L. T. McCarty (1977). «Reflections on Taxman: An experiment in artificial intelligence and legal reasoning». *Harvard Law Review* (núm. 90, págs. 837-893).

2) Sistema LDS de Rank Corporation. En este caso, los desarrolladores describían el razonamiento involucrado en la liquidación de reclamaciones de responsabilidad civil en Estados Unidos. Para poder construir el sistema, los autores estudiaron cómo los abogados y ajustadores evaluaban las demandas civiles en el área de responsabilidad del producto, y elaboraron un esquema que organiza los hechos y detalles de un caso. Tanto el esquema como un amplio conjunto de reglas necesarias para el razonamiento de los casos forman la base de un sistema experto que modela la toma de decisiones legales. Este sistema puede ayudar a los investigadores y a los litigantes a comprender mejor cómo se lleva a cabo la evaluación de las reclamaciones, ya que proporciona una base para generar y organizar hipótesis sobre los métodos de los litigantes para establecer la liquidación. Los resultados obtenidos fueron muy positivos, demostrando que los sistemas expertos basados en reglas pueden captar gran parte de la riqueza y flexibilidad del razonamiento legal en este campo.

Lectura recomendada

D. A. Waterman; L. A. Peterson (1984). «Evaluating civil claims: an expert systems approach». *Expert Systems* (núm. 1, págs. 65-76).

Ejemplo de reglas en este sistema

En un lenguaje seminatural, para su mejor comprensión.

REGLA 4

IF Demandante recibe herida en un ojo
AND ojos afectados= 1
AND tratamiento del ojo requirió cirugía
AND recuperación de la herida casi completa
AND agudeza visual se redujo ligeramente
AND la condición de la herida es estable
THEN aumentar factor de trauma de herida en 10.000 $

REGLA 6

IF Demandante tiene probabilidad de contraer enfermedad seria
AND valor probabilidad >= 5 %
AND valor probabilidad <= 15 %
THEN aumentar factor de trauma futuro del valor del parámetro contraer-enfermedad en 30 %

REGLA 8

IF Demandante no llevaba gafas antes de la herida
AND la herida recibida requiere llevar gafas
AND edad del demandante en el momento de la herida > 25
AND la apariencia física es importante para su trabajo
THEN aumentar factor de desfiguración en 5.000 $

3) Sistema legal Research System, que tiene la función de ayudar a los operadores jurídicos a recuperar informaciones relativas a las decisiones judiciales y a la legislación en el campo del derecho de los títulos de comercio.

Lectura recomendada

C. D. Hafner (1987). «Conceptual organization of case law knowledge bases». En: *Proceedings of the 1st international conference on Artificial intelligence and Law- ICAIL* (págs. 35-42). ACM.

Existen algunos otros sistemas expertos conocidos en el campo del derecho, y varios artículos contienen reflexiones acerca de su uso, sus limitaciones y su futuro. El crecimiento de las aplicaciones para problemas del mundo real y la proliferación de los sistemas expertos provocaron el aumento de la demanda de nuevos esquemas de representación de conocimiento y razonamiento, que intentaban mejorar el rendimiento tanto de los sistemas como de su caro y complejo proceso de construcción. Aun así, en la actualidad son los sistemas basados en aprendizaje automático, es decir, los que aprenden directamente de los datos, los que están triunfando en la mayoría de las aplicaciones reales, estando los sistemas simbólicos restringidos a partes pequeñas de un dominio determinado.

Lecturas recomendadas

Ashley, K. D. (2000). «Designing electronic casebooks that talk back: The cato pro- gram». *Jurimetrics Journal* (núm. 40, págs. 275-319).

Barrio Andrés, M. (2018). *Robótica, inteligencia artificial y derecho.*

Cáceres, E. (2008). «EXPERTIUS: A mexican judicial decision-support system in the field of family law». *Proceedings of the 2008 conference on Legal Knowledge and Information Systems: JURIX 2008* (págs. 78-87).

Casanovas, P. (2010). «Inteligencia artificial y derecho: a vuelapluma». *Revista de Pensamiento Jurídico* (núm. 7, págs. 203-221).

Hernández Jiménez, M. (2019). «Inteligencia artificial y derecho penal». *Actualidad Jurídica Iberoamericana* (núm. 10bis, págs. 792-843).

2.2. IA subsimbólica

También llamada inteligencia artificial computacional o inductiva, engloba los métodos de aprendizaje automático y la computación evolutiva. La idea que subyace a la aproximación es la de que una de las características más distintivas de la inteligencia humana es la capacidad de aprendizaje.

Los algoritmos utilizan para aprender procesos de inducción que, a partir de ejemplos particulares, son capaces de generalizar comportamientos o reconocer patrones, aunque también existen métodos de aprendizaje automático basados en refuerzo (inspirados en la psicología conductista, y que se basan en determinar qué acciones debe escoger un agente software en un entorno dado con el fin de maximizar alguna noción de «recompensa» o premio acumulado), o en la imitación de procesos evolutivos biológicos, usando algoritmos evolutivos, que son optimizadores matemáticos que funcionan generando muchas variaciones singulares de un individuo, de forma que se obtiene una población que sufre cruces, mutaciones, etc. y que, finalmente, evoluciona manteniendo a los mejores individuos de la población, que selecciona utilizando una función de ajuste o *fitness*.

En general, el **área del aprendizaje automático** trata de dar una respuesta a la necesidad de construcción de sistemas computerizados adaptables al entorno desde una perspectiva diferente a la que vimos en el caso de los sistemas expertos, ya que en este caso, o bien la experiencia humana de partida necesaria no existe o no es fácil de extraer. Por ese motivo, no es posible el desarrollo de sistemas basados en reglas que recojan el conocimiento de los expertos humanos, como los que vimos en el apartado anterior. En estos casos, las técnicas subsimbólicas nos permiten programar sistemas inteligentes utilizando datos del proceso que está transcurriendo, o bien datos procedentes de experiencias pasadas. Así, se podrán identificar ciertos patrones o regularidades en los datos que se utilizarán para construir buenas aproximaciones al problema.

Ejemplo: una fotografía

Consideremos la cuestión de reconocer a alguien a través de una fotografía. Esta es una tarea que los seres humanos realizamos sin ninguna dificultad, aunque la foto esté oscura, o la pose dificulte el reconocimiento, entre otros factores. No obstante, nos resulta tremendamente difícil de explicar cómo lo hacemos, por lo que afrontar la tarea desde una aproximación simbólica sería poco menos que imposible. Sin embargo, también sabemos que un rostro no es algo aleatorio, sino que tiene una estructura con ciertas características que serán comunes para una persona en las diferentes imágenes que tengamos

de ella. En esta aproximación subsimbólica, la idea consistiría en recoger un número importante de muestras del caso concreto; en este ejemplo, la muestra sería de fotografías de distintas personas, y aprender las similitudes o patrones específicos de cada una de ellas, para más tarde comparar las nuevas fotografías que el sistema inteligente aún no ha visto con los patrones descubiertos en las anteriores, para poder realizar la identificación de las diferentes personas en las nuevas fotos.

Otra situación en la que esta aproximación resulta interesante es aquella en la que el problema a resolver varía en el tiempo (por ejemplo, la detección de intrusos en redes de ordenadores), o depende del entorno particular (contexto específico) en el que se trabaje. Para ser considerado inteligente, un sistema tendrá que tener la **habilidad de aprender**. Así, podremos tener sistemas de propósito más general capaces de adaptarse a estas circunstancias, en lugar de tener que escribir cada vez programas explícitos para cada situación.

Ejemplo: tráfico de paquetes

Un ejemplo de este tipo podría ser un sistema que redirija el tráfico de paquetes en una red, de manera que se maximice la rapidez del servicio. El camino que maximiza la calidad del servicio entre un origen y un destino varía continuamente ya que depende del tráfico de la red. Un programa de aprendizaje podría monitorizar y adaptarse al entorno cambiante del tráfico, y así suministrar el mejor camino en cualesquiera circunstancias.

Otros ejemplos podrían ser las interfaces inteligentes que pueden adaptarse al perfil del usuario en función de ciertos comportamientos, como sus hábitos de trabajo; o el contenido de páginas web, que podrían adaptarse en función de los perfiles de intereses del usuario. En la actualidad, esta es la rama más exitosa de la inteligencia artificial, debido a la disponibilidad de datos en prácticamente cualquier campo, debido al proceso de digitalización que se está llevando a cabo, que hace posible representar digitalmente la música, la cultura, los viajes o el cuerpo humano, entre otras cosas. Este proceso de convertir casi todo en datos es posible gracias a los **sensores** que registran los eventos y actividades que llevan el mundo físico al digital como, por ejemplo, un secuenciador de genoma. Hace diez años costaba alrededor de 200 millones de euros secuenciar el genoma de un individuo. A día de hoy, cuesta menos de 500 euros y este precio de digitalizar el genoma tenderá a disminuir aún más. Como las entidades digitales pueden ser fácilmente replicadas, almacenadas, transmitidas, modificadas o vendidas, este traspaso hace que sectores enteros –en este caso, la salud– se transformen en servicios de información y conocimiento. Lo mismo sucede en el mundo financiero. Desde una transacción hasta la mayoría de los elementos que conforman la relación con el cliente (o del cliente con el mundo) son información que vive en el mundo digital.

La era del *big data*

Esta proliferación de los diferentes sensores, tanto como la posibilidad de almacenar datos no estructurados (como imágenes escaneadas, documentos, fotos, etc.), origina que en los últimos años hayamos entrado claramente en lo que se conoce como la era de *big data*, en la que el volumen de datos ha crecido de forma exponencial, a una velocidad muy alta, que hace que la cantidad de datos disponible se duplique cada año, y en una variedad importante de tipos (datos estructurados, semiestructurados y no estructurados) que han empequeñecido los sistemas de almacenamiento y procesado tradicionales, transformando las organizaciones y demandando formas de procesado de información innovadoras y eficientes en coste para obtener información y conocimiento nuevos que permitan añadir valor de negocio a los datos disponibles.

En este contexto, los métodos automáticos de análisis de datos se hacen imprescindibles. Por otra parte, también estamos experimentando muchos avances en software; como ya hemos mencionado han aparecido aproximaciones disruptivas, como es el caso de los algoritmos de aprendizaje profundo, que han posibilitado resultados muy precisos.

Lectura recomendada

I. Goodfellow; Y. Bengio; A. Courville (2017). *Deep Learning* . MIT Press.

2.3. Otras clasificaciones

Existen otras divisiones que podemos hacer de la inteligencia artificial, por ejemplo, en función del tipo de problema que queremos resolver. En este caso, estaríamos hablando de problemas relacionados con:

1) Percepción, es decir, la capacidad de comprender información no estructurada en forma de imágenes o vídeo (área de visión por computador), voz (reconocimiento del habla o generación artificial de voz) o textos (procesamiento del lenguaje natural).

Ejemplos de sistemas de este tipo

Los sistemas de reconocimiento facial, los etiquetadores automáticos de fotografías o vídeos (reconociendo tipos de plantas, personas concretas, etc.), los *chatbots* (programas inteligentes con los que es posible mantener una conversación, y que se pueden utilizar en muchos entornos, por ejemplo, para comunicar accidentes a nuestro seguro 24 horas al día, 7 días a la semana), o los asistentes personales (como Siri, Cortana o Alexa).

2) Planificación y búsqueda, que se ocuparían de encontrar la mejor solución entre un gran conjunto de alternativas posibles, en entornos que pueden ser o no totalmente observables, deterministas, finitos, estáticos o dinámicos, discretos o continuos, etc.

Ejemplo de sistemas de este tipo

Los planificadores de rutas de vehículos, que deben trabajar en condiciones reales y en tiempo real, o los planificadores de horarios y turnos en una empresa.

3) Representación del conocimiento y razonamiento automático, que tratan sobre la capacidad de almacenar, expresar y manipular el conocimiento adquirido sobre un dominio, pudiendo hacer uso del conocimiento existente para extraer conclusiones.

Ejemplo de sistemas de este tipo

Los reconocedores de actividades, que se utilizan sobre todo en el campo de la domótica, el ahorro energético o la atención remota a personas dependientes.

4) Aprendizaje, que enfoca la capacidad de generar nuevo conocimiento a partir de nuevas observaciones, etc.

Ejemplo de sistemas de este tipo

Los sistemas que aprenden a reconocer el correo no deseado, los algoritmos de marketing personalizado, la medicina de precisión o la educación individualizada utilizan aproximaciones basadas en este tipo de capacidad.

Finalmente, es importante resaltar el hecho de que debido a su complejidad, en la mayoría de las situaciones reales los sistemas inteligentes necesitan combinar varias de las capacidades que hemos visto anteriormente. Ejemplos típicos son, entre otros, los sistemas multiagente o la robótica:

1) Sistema multiagente: consta de varios agentes inteligentes que interactuán entre ellos, y la idea es que se utilizan en aquellos problemas que son difíciles o imposibles de resolver para un agente inteligente único. Así, se diseñan varios agentes que se ocupan de una parte del problema; estos agentes necesitan establecer mecanismos de cooperación, negociación y comunicación entre ellos para poder organizar la resolución del problema global. Los sistemas multiagente también pueden verse como una forma de distribuir inteligencia artificial para lograr sistemas más escalables.

En otros problemas, también se pueden utilizar agentes inteligentes para modelar situaciones complejas en las que lo que se pretende obtener es un comportamiento emergente que no necesariamente estaba planeado desde un principio o definido dentro de los agentes mismos, y que surge por la interacción entre ellos. Este proceso de emergencia surge a macroescala debido a los comportamientos de los agentes individuales a microescala. Constituyen modelos computacionales que se usan de forma extensiva en biología (por ejemplo, para analizar la difusión de epidemias, o la evolución de poblaciones), en análisis de redes sociales, en otras áreas de las ciencias sociales y la economía, etc.

2) Otra área interesante de la inteligencia artificial que resulta de interés mencionar es la **robótica**, que trata de los agentes físicos que realizan tareas mediante la manipulación física de su entorno. Para poder realizar esto, los robots tienen sensores de diversos tipos que les permiten percibir el entorno (como cámaras, acelerómetros, aparatos de ultrasonidos, etc.), y también efectores, que les permiten actuar sobre dicho entorno (como articulaciones, ruedas, etc.). Los tipos de robots más conocidos son:

- los manipuladores o brazos robóticos (por ejemplo, los de la estación espacial internacional, o los de las líneas de ensamblaje),

- los robots móviles (vehículos terrestres, aéreos o submarinos no tripulados, o robots que mueven contenedores en puertos como el de Hamburgo, por ejemplo) y

- los robots zoomórficos, cuyos sistemas de locomoción imitan a los diversos seres vivos, y que incluyen los robots humanoides, que suelen además diseñarse para tener aptitudes sociales, y que por ejemplo se pueden utilizar en hospitales para acompañar pacientes.

3. Conceptos básicos sobre algoritmos

En este apartado, explicaremos conceptos básicos para entender cómo son los datos que habitualmente usan los algoritmos de inteligencia artificial, centrándonos en el subcampo del aprendizaje automático (o *machine learning*, en inglés). La idea principal es la identificación de patrones o tendencias que existen en los datos de forma automática. Para esta tarea, hay dos grandes familias de métodos: aprendizaje supervisado y aprendizaje no supervisado.

1) Aprendizaje supervisado: en este caso, los algoritmos trabajan con datos que están etiquetados, lo cual significa que habrá una función que, dados los datos de entrada, encuentre su salida deseada.

Ejemplo de aprendizaje supervisado

Un detector de *spam* trabaja buscando una función que relacione los datos que tenemos sobre un correo electrónico (remitente, tipo de destinatario, asunto, etc.) y les asigne una etiqueta que será «*spam*» o «no *spam*».

El aprendizaje supervisado se suele usar en problemas de clasificación.

2) Aprendizaje no supervisado: en este caso, no se dispone de datos etiquetados para el entrenamiento del algoritmo, por lo que los métodos no supervisados se basan en describir la estructura de los datos para intentar encontrar algún tipo de organización que simplifique el análisis.

Ejemplo de aprendizaje no supervisado

Un caso representativo de aprendizaje no supervisado es el agrupamiento (o *clustering*, en inglés), que busca agrupamientos basados en similitudes.

A continuación, nos centraremos en las tareas de clasificación, ya que son probablemente las más usadas en aprendizaje automático.

3.1. Los conjuntos de datos

Un factor clave para el análisis de los datos son, como puede resultar obvio, los propios datos. En los últimos años, vivimos en una sociedad en la que almacenamos y recolectamos cantidades ingentes de datos sobre casi cualquier materia que podamos imaginar, lo cual ha propiciado la aparición del llamado *big data*. Con este océano de datos en el que nos encontramos inmersos, ha aparecido un perfil profesional que está altamente demandado, el **científico de datos**, cuya misión es extraer información útil a partir de enormes cantidades de datos «en crudo». Pero empecemos por el principio... ¿Cómo podemos definir lo que son los datos?

Los datos se recolectan habitualmente por investigadores en forma de conjunto de datos (o *dataset* en inglés). Un conjunto de datos se puede definir como una colección de datos individuales, a menudo llamados *muestras*, *instancias* o *patrones*. Una muestra se puede ver como la información sobre un ejemplo particular, por ejemplo sobre un paciente en el dominio médico. La información sobre este ejemplo particular se obtiene de las llamadas *características* o *atributos*. Una característica puede ser el sexo del paciente, su presión sanguínea o el color de sus ojos.

Una de las tareas más importantes en el campo del análisis de datos es la clasificación que, como se ha mencionado antes, consiste en asignar a cada muestra una *clase* o categoría específica. Típicamente, las muestras que pertenecen a una misma clase tienen características similares. En la tabla 1 podemos ver, como ejemplo, el conjunto de datos de «jugar al tenis»[1]. En este pequeño ejemplo tenemos quince muestras, y cada muestra tiene cuatro características diferentes que nos dan información que puede ser útil para determinar si es posible jugar al tenis o no (teniendo en cuenta que el tenis es un deporte que se juega en exterior). La última columna representa la variable predictiva o clase, que es la salida deseada de este conjunto de datos en una tarea de clasificación.

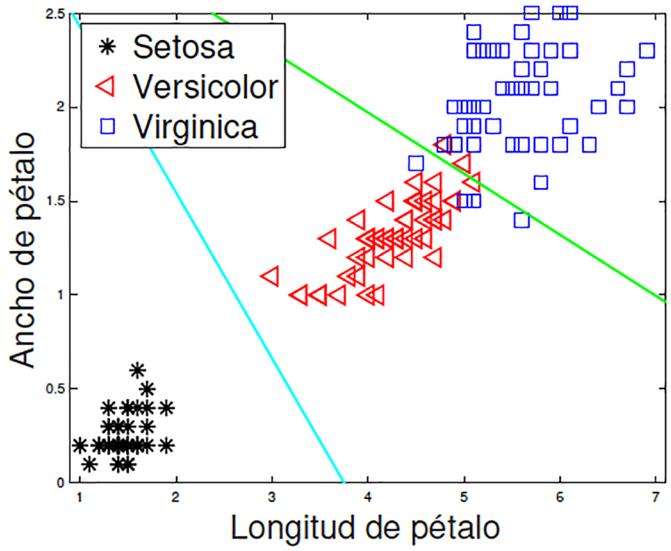
[(1)]Sobre el repositorio de conjunto de datos: **D. Dua; Graff, C.** (2017). *UCI machine learning repository*.

Tabla 1. Conjunto de datos de jugar al tenis

Pronóstico	Temperatura	Humedad	Ventoso	¿Jugar?
soleado	caluroso	alta	falso	no
soleado	caluroso	alta	verdadero	no
soleado	caluroso	alta	verdadero	no
nublado	caluroso	alta	falso	sí
lluvioso	templado	alta	falso	sí
lluvioso	fresco	normal	falso	sí
lluvioso	fresco	normal	verdadero	no
nublado	fresco	normal	verdadero	sí
soleado	templado	alta	falso	no
soleado	fresco	normal	falso	sí
lluvioso	templado	normal	falso	sí
soleado	templado	normal	verdadero	sí
nublado	templado	alta	verdadero	sí
nublado	caluroso	normal	falso	sí
lluvioso	templado	alta	verdadero	no

Uno de los conjuntos de datos más populares en la literatura sobre análisis de datos es el conjunto Iris. Este conjunto de datos se ha usado en miles de trabajos de investigación a lo largo de los años, y consiste en distinguir entre tres clases de planta iris (setosa, virgínica y versicolor). El conjunto tiene cuatro características, que son el ancho y longitud de pétalo y sépalo, y cincuenta muestras de cada una de las tres clases o categorías. Como se puede ver en la figura 3, una de las clases (setosa) es claramente separable con una línea recta de las otras dos, mientras que en las clases virgínica y versicolor esto no sería posible, ya que los ejemplos se mezclan en la frontera de ambas clases.

Figura 3. Representación de las tres clases del conjunto de datos Iris, con la longitud de pétalo en el eje horizontal y el ancho de pétalo en el eje vertical



Fuente: elaboración propia

El hecho de contar con características que sean linealmente separables nos hace conseguir una precisión de clasificación perfecta, mientras que cuando las clases no son separables es posible que algunos clasificadores cometan algunos errores. Comentaremos en más detalle este asunto en el siguiente apartado.

3.2. Error/precisión de clasificación

Como hemos mencionado anteriormente, la tarea de un clasificador es asignar a qué clase pertenece una muestra determinada. Por tanto, necesitaremos medidas para evaluar cómo de buena ha sido la clasificación.

Una métrica de evaluación ampliamente usada es el **error de clasificación**, que es el porcentaje de muestras incorrectamente clasificadas dividido por el número total de muestras. Análogamente, la **precisión de clasificación** es el porcentaje de muestras correctamente clasificadas dividido por el número total de muestras.

Sin embargo, fijarse solo en el error o precisión de clasificación no es una buena idea. Supongamos que tenemos un conjunto de datos formado por cien imágenes, de las cuales noventa y cinco de ellas son gatos y solo cinco son perros. Usaremos en primer lugar un clasificador, al que llamaremos C_1, que decide que todas las imágenes son gatos, por lo que tendrá una precisión de clasificación del 95 %, lo cual suena muy bien. A continuación usaremos otro clasificador, al que llamaremos C_2, que falla al clasificar cuatro imágenes de gatos y dos de perros, obteniendo finalmente una precisión del 94 %. ¿Qué clasificador es mejor? La respuesta a esta pregunta depende del tipo de datos y del objetivo del aprendizaje, pero, en general, es mejor conseguir una solución de compromiso entre la capacidad de clasificación de las dos clases, por lo que es necesario comprobar también los porcentajes de acierto en cada una de las clases individualmente.

Además, existen otras medidas para evaluar la bondad de los clasificadores que es muy importante tener en cuenta:

- **Verdaderos positivos** (VP): porcentaje de muestras positivas clasificadas como positivas.
- **Falsos positivos** (FP): porcentaje de muestras negativas clasificadas incorrectamente como positivas. También conocido como error de tipo I.
- **Verdaderos negativos** (VN): porcentaje de muestras negativas clasificadas como negativas.
- **Falsos negativos** (FN): porcentaje de muestras positivas incorrectamente clasificadas como negativas. También conocido como error de tipo II.

Por supuesto, el objetivo de cualquier sistema clasificador es mantener muy altas las tasas de VP y VN. Sin embargo, hay que tratar con mucho cuidado los FP y FN, priorizando uno u otro dependiendo del escenario o naturaleza del problema. Veamos dos ejemplos.

Ejemplo: detección de enfermedades

Supongamos que tenemos un sistema de inteligencia artificial para detectar si un paciente tiene una determinada enfermedad, y para ello usaremos un algoritmo de clasificación. En este contexto, un FP significará que le diremos al paciente que tiene la enfermedad cuando en realidad está sano, y un FN significará que le diremos al paciente que está sano cuando realmente tiene la enfermedad. Aunque es deseable que ninguna de estas situaciones ocurran, en este contexto es mejor cometer un FP (ya que lo más probable es que en posteriores pruebas se descubra que el paciente realmente no tiene la enfermedad) a cometer un FN, y que tengamos un paciente enfermo que no esté siendo tratado.

Ejemplo: amenazas en la red

Supongamos ahora que tenemos un sistema de inteligencia artificial para detectar amenazas en una red de computadoras. En este contexto, un FP significará que se ha enviado una alarma de que hay un ataque cuando era una conexión normal, y un FN significará que había una amenaza pero no se ha detectado. Por supuesto, es crucial no dejar pasar ataques inadvertidos, pero una tasa de FP alta puede tener efectos desastrosos, ya que si recibimos muchas alarmas falsas, dejaremos de hacerles caso, y cuando se produzca un ataque real lo ignoraremos y el sistema no servirá para nada.

3.3. Entrenamiento y prueba

A lo largo de este apartado hemos estado hablando de la tarea de clasificación y cómo un algoritmo debe aprender a diferenciar entre las clases existentes. Pero ¿qué pasa cuando llega una nueva muestra para ser clasificada? Por ejemplo, ¿qué pasa cuando llegan los datos de un nuevo paciente y necesitamos saber si está enfermo? Esta es la esencia de los algoritmos de clasificación: ser capaces de **clasificar muestras nuevas** para las cuales no sabemos *a priori* su clase o categoría.

En una situación ideal, se usarían todas las muestras disponibles en nuestro conjunto de datos para las cuales sabemos cuál es su clase (por ejemplo, porque las hemos sacado de un histórico de datos). De esta forma, nuestro algoritmo de clasificación será capaz de aprender las particularidades de los datos y la relación entre los valores de las características y su correspondiente clase. Más tarde, cuando llegue una nueva muestra para la cual no sabemos su clase, nuestro algoritmo de clasificación ya entrenado hará una predicción sobre la clase a la que debe pertenecer esta nueva muestra. Pero, de esta forma, ¿cómo sabremos si nuestro algoritmo de clasificación ha aprendido correctamente a partir de datos que representen correctamente el problema? Por ejemplo, y volviendo a la analogía con el problema médico, ¿qué pasaría si hemos estado entrenando a nuestro algoritmo solo con pacientes mujeres pero en el mundo real los hombres desarrollan otro tipo de síntomas para la enfermedad?

Para asegurarnos un proceso de aprendizaje correcto, existen dos prácticas habituales:

- En primer lugar, asegurarnos de que el conjunto de datos que usaremos para aprender representa correctamente a la población global.
- En segundo lugar, guardar parte de los datos de los que sabemos la clase para usar en lo que se llama comúnmente el conjunto de prueba o test. Este segundo paso es muy importante a la hora de desarrollar una correcta metodología de aprendizaje, ya que es necesario guardar una parte de los datos que el algoritmo de aprendizaje nunca haya visto mientras se ha entrenado para aprender.

> Todos los parámetros que estén involucrados en el proceso de aprendizaje se deben calibrar sobre el conjunto de entrenamiento, y nunca sobre el conjunto de test, así como los procesos de preprocesado de los datos, ya que los datos que se han guardado para la fase de test no deben utilizarse en ninguna de las otras fases de aprendizaje.

Esto incluye la selección de características, que consiste en seleccionar las características relevantes para un problema dado y descartar las irrelevantes o redundantes. Es una práctica relativamente habitual (aunque incorrecta) rea-

lizar la selección de características sobre todos los datos disponibles y, una vez eliminadas las características que no son necesarias, realizar la división de los datos en entrenamiento y test. Esta práctica afectará negativamente al proceso de aprendizaje, ya que todo tipo de preprocesado de los datos tiene que realizarse únicamente sobre el conjunto de entrenamiento, dejando el conjunto de test para evaluar el rendimiento del modelo aprendido.

Algunos de los conjuntos de datos comúnmente usados en la literatura científica ya vienen originalmente divididos en conjunto de entrenamiento y prueba.

Ejemplo: conjunto de datos divididos originalmente en entrenamiento y prueba

El conjunto de datos KDD (Knowledge Discovery and Data Mining Tools Conference) Cup 99, un conjunto de referencia en el campo de los sistemas de detección de intrusos.

Existen conjuntos de datos separados para entrenamiento y prueba, con la particularidad de que el porcentaje de las distintas clases (conexión normal y distintos tipos de ataque) varía significativamente del conjunto de entrenamiento al de prueba, incluso dándose el caso de que en el conjunto de prueba aparecen nuevos ataques que no estaban presentes en el conjunto de entrenamiento.

Cuando los datos no vienen divididos de origen, son los investigadores los que deben decidir cómo dividir los datos. Existen varios protocolos habituales para esta división, y a continuación se describirán los más usados:

- **Validación cruzada con *k* paquetes**. Este es uno de los esquemas de validación más populares. El conjunto se datos se divide en *k* paquetes de igual tamaño. El algoritmo de aprendizaje se entrena con *k* 1 paquetes, y posteriormente se usa el paquete restante como conjunto de test o prueba y se estima el error. Este proceso se repite tantas veces como paquetes *k* tengamos, y el error de validación cruzada se calcula a partir de la media de los *k* errores obtenidos en cada iteración.

- **Validación cruzada dejando uno**. Esta es una variante de la validación cruzada con *k* paquetes, de forma que *k* es el número de muestras del paquete. En cada iteración del proceso, se deja para prueba una única muestra.

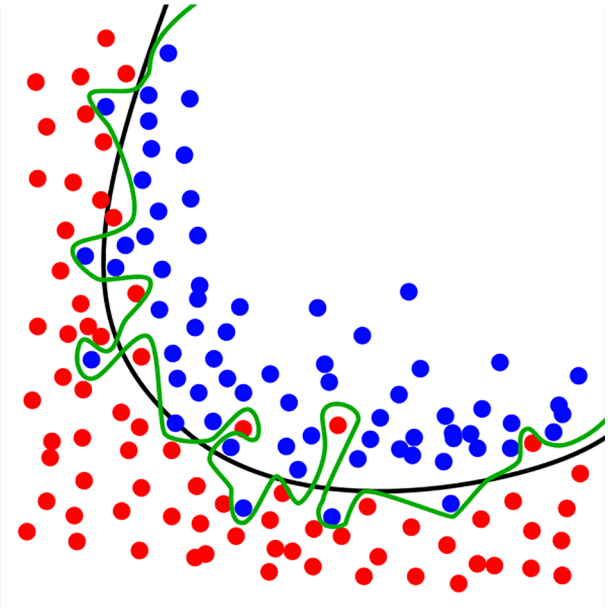
- ***Bootstrap***. Esta es una técnica general de remuestreo. Una muestra *bootstrap* está compuesta de *n* muestras con igual probabilidad de ser escogidas, con reemplazo, del conjunto de datos original. De esta forma, es posible que algunas de las muestras aparezcan repetidas varias veces, mientras que otras muestras no lleguen a aparecer nunca. El algoritmo de aprendizaje se entrena con la muestra de datos *bootstrap* y, posteriormente, se prueba con las muestras que se han quedado fuera. El error se aproxima calculando la

media de las muestras basado en réplicas independientes (normalmente entre veinticinco y cien).

- **Validación simple**. Esta técnica consiste en dividir de forma aleatoria los datos disponibles en dos particiones: entrenamiento y test. Normalmente se divide dejando 2/3 de los datos para entrenamiento y 1/3 para test. El algoritmo de aprendizaje se entrena usando los datos de la partición de entrenamiento y el error se estima calculando la proporción de errores en los datos de prueba. Este tipo de validación se usa cuando, en un estudio, hay conjuntos de datos que ya vienen divididos originariamente en entrenamiento y test pero otros no.

La elección de la técnica de validación apropiada no es en absoluto trivial, y normalmente depende del tamaño de datos que tengamos que tratar. Por ejemplo, si solo contamos con cien muestras (algo habitual con algunos tipos de datos genéticos, por ejemplo), escoger validación simple dejando 2/3 para entrenamiento y 1/3 para test no sería una buena idea, ya que el tamaño del conjunto de entrenamiento puede ser demasiado pequeño y sufrir el efecto de sobreajuste de los datos (es decir, el algoritmo de aprendizaje no podrá generalizar; ved figura 4). Por el contrario, si tenemos un conjunto de datos realmente grande (millones de muestras o características), usar validación cruzada en cualquiera de sus variantes provocará un tiempo de computación excesivamente elevado, por lo que se está volviendo al método clásico de la validación simple.

Figura 4. Ejemplo de sobreajuste de los datos



Fuente: *Diagram showing overfit- ting of a classifier* con licencia Creative Commons vía Wikimedia Commons. Disponible en: https://commons.wikimedia.org/wiki/File:Overfitting.svg

La línea verde representa un clasificador que se adapta perfectamente a los datos (separando las dos clases –azul y roja– correctamente), pero está demasiado adaptado a ellos, y ante la aparición de nuevos datos probablemente obtendrá más errores que el clasificador que está representado por la línea negra.

3.4. Comparación de modelos: test estadísticos

Cuando se presenta un nuevo algoritmo de aprendizaje, es necesario comparar sus resultados con los que obtienen los métodos que conforman el estado del arte, para demostrar que nuestra propuesta es rompedora y competitiva. Para que esta comparativa sea justa, la práctica habitual es usar test estadísticos. En su libro, L. Kuncheva nos da una serie de recomendaciones para comparar varios modelos:

Lectura recomendada

L. I. Kuncheva (2014). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

- Escoger cuidadosamente la división entre entrenamiento y prueba (ver apartado anterior) antes de empezar los experimentos. Si el experimento se publica, es necesario asegurarse de que se dan los detalles necesarios para que pueda ser reproducible.

- Asegurarse de que todos los modelos usan toda la información posible, y por supuesto de que usen los mismos datos para entrenar y posteriormente en la etapa de prueba. Por ejemplo, no es una comparativa justa ejecutar validaciones cruzadas distintas para varios modelos, porque la división aleatoria de los datos puede favorecer a alguno de los modelos. La forma correcta de realizar los experimentos es dividir los datos en los paquetes necesarios y en cada iteración entrenar los distintos modelos con los datos correspondientes de entrenamiento.

- Asegurarse de que los datos reservados para la etapa de prueba no se usen anteriormente en la etapa de entrenamiento.

- Cuando sea posible, realizar test estadísticos. Es mejor saber si las diferencias entre modelos son significativamente diferentes o no.

4. IA fiable

Como hemos mencionado anteriormente, con la aparición del fenómeno *big data*, hemos sido testigos de un incremento abrumador en el volumen de datos almacenados, al mismo tiempo que estos datos se han dotado de variedad, variabilidad, veracidad (es decir, asegurarnos de que los datos son de calidad y no contienen ruido o anormalidades), etc. En los primeros años, esta situación ha resultado en un empeoramiento en el funcionamiento de algoritmos de IA existentes, en particular, de algoritmos de aprendizaje automático. Como consecuencia de esto, la escalabilidad de algoritmos se convirtió en un requisito fundamental, para poder tener algoritmos eficaces, eficientes y que pudieran ser ejecutados con grandes tamaños de datos.

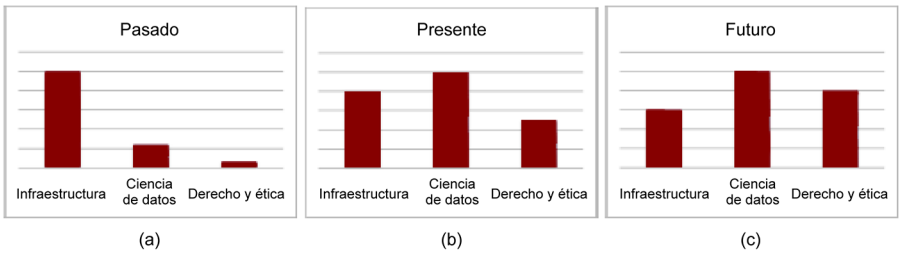
Si miramos este campo de estudio desde otra perspectiva, podemos ver que la evolución de la industria de datos ha sido muy rápida. Durante los primeros años del fenómeno del *big data*, se ha centrado en el desarrollo de las infraestructuras necesarias para procesar estas enormes cantidades de datos. Actualmente, estas plataformas están bien asentadas en el mercado. Sin embargo, hay una nueva preocupación que ha irrumpido con fuerza en el campo de la IA, y es la **necesidad de una regulación ética y legal** que asegure la privacidad de los datos y su uso legítimo, transparencia en las decisiones, etc., que no puede ser pospuesta.

El RGPD

En la Unión Europea, el Reglamento General de Protección de Datos (RGPD) está cambiando el panorama actual, regulando el uso de los datos y el derecho a una explicación en el caso de decisiones automatizadas, con el objetivo de conseguir una IA transparente que evite discriminaciones sobre la base de la raza, el estado de salud, el sexo, etc.

En el futuro, se espera cada vez menos demanda en infraestructura, ya que el mercado está de alguna forma asentado, pero todavía existe una necesidad de soluciones que sean más rápidas y potentes, por lo que aún hay margen para la mejora. El interés en ciencia de datos continuará, ya que estaremos trabajando cada vez en más áreas, usando cada vez más datos que necesitarán preprocesarse antes de trabajar con ellos. Además, también es necesario desarrollar soluciones que permitan explicar y visualizar los resultados obtenidos, así como manejar los datos dentro de los algoritmos respetando su privacidad. Finalmente, el área relacionada con las leyes, reglamentos y ética irá ganando en popularidad, y profesionales de otros campos como el derecho, la economía o la filosofía serán necesarios, configurando de esta forma un área todavía más interdisciplinar. La inteligencia artificial comienza un nuevo camino para ganarse la confianza de la sociedad.

Figura 5. Pasado, presente y futuro en las distintas áreas que cobran protagonismo en el desarrollo de la inteligencia artificial



Fuente: elaboración propia

En abril de 2018, la Comisión Europea publicó un comunicado en el cual se anunciaba una ambiciosa estrategia europea para la inteligencia artificial. Parte de esta estrategia ha sido la creación del **Grupo de Expertos de Alto Nivel en Inteligencia Artificial** (AI HLEG), que incluye representantes del mundo académico, sociedad civil y empresas. Su objetivo es apoyar la implementación de la estrategia europea en IA, incluyendo la elaboración de recomendaciones para futuras políticas de desarrollo relacionadas con IA, así como asuntos éticos, legales y sociales relacionados con la IA. En abril de 2019, el AI HLEG publicó un documento de directrices y ética para una IA fiable. De acuerdo a estas directrices, y como se ha comentado en la introducción, la IA fiable debe ser:

- **Lícita**. Respetando todas las leyes y reglamentos aplicables.
- **Ética**. Asegurando el cumplimiento de los principios y valores éticos.
- **Robusta**. Tanto desde el punto de vista técnico como social, ya que los sistemas de IA, incluso si sus intenciones son buenas, pueden provocar daños por accidente.

Además, las directrices establecen una serie de siete requisitos clave que los sistemas de IA deben cumplir para poder ser considerados fiables.

1) Acción y supervisión humana:

- En primer lugar, es necesario que los sistemas de IA respeten los derechos fundamentales de las personas. Aunque idealmente deben ser concebidos para facilitar la vida de las personas, por ejemplo mejorando la accesibilidad a la educación, es también posible que afecten negativamente a los derechos fundamentales, y este riesgo debe ser evaluado cuidadosamente.

- En segundo lugar, es necesario proporcionar a los usuarios de los sistemas de IA los conocimientos y herramientas necesarios para comprender los sistemas de IA y por qué toman determinadas decisiones. El principio general de autonomía del usuario es fundamental en el diseño del sistema, respetando el derecho a no ser juzgado basándose exclusivamente en la decisión de un proceso automatizado, como así recoge el artículo 22 del RGPD.

- Y, en tercer lugar, es necesario que exista supervisión humana de los sistemas de IA, tanto en la fase de diseño como en la fase de ejecución. De este modo, será posible incluso ignorar una decisión tomada por un sistema si así se estima conveniente.

2) Solidez técnica y seguridad. La solidez técnica está estrechamente vinculada al principio de prevención del daño, y requiere que los sistemas de IA se desarrollen con un enfoque preventivo en relación con los riesgos. Estos riesgos incluyen que el sistema de IA puede ser atacado, como cualquier sistema de software, por lo que es necesario que los sistemas de IA sean seguros y robustos. Por último, para evitar fallos indeseados, es fundamental que los sistemas de IA sean reproducibles. Esto significa que si repetimos varias veces un experimento en las mismas condiciones, el resultado debería ser siempre el mismo, y así es posible describir con exactitud lo que hace un sistema de IA.

3) Gestión de la privacidad y los datos. La privacidad es un derecho fundamental que a menudo es vulnerado por los sistemas de IA. Estos sistemas no solo usan la información inicialmente facilitada por el usuario, sino que también gestionan la información obtenida sobre el usuario en el contexto de su interacción con el sistema. Esto es un tema que debe ser tratado con delicadeza, ya que los sistemas de IA pueden utilizar para su aprendizaje no solo la información original sobre los usuarios, sino también otra información más susceptible de provocar discriminaciones, como su orientación sexual, género, etc. Otro frente abierto muy importante es la calidad de los datos, ya que para que un sistema de IA aprenda correctamente, es necesario que los datos de los que aprenda sean de calidad, evitando sesgos sociales, imprecisiones, o incluso la inclusión de datos malintencionados.

4) Transparencia. Este requisito está muy relacionado con el principio de explicabilidad e incluye todos los elementos susceptibles de ser transparentes para un sistema de IA: los datos, el sistema y los modelos de negocio. Por una parte, un sistema de IA debe ser **trazable**; esto significa que todos los datos y decisiones acerca del sistema deben ser documentados de forma rigurosa, para facilitar la trazabilidad futura y aumentar la transparencia. A su vez, la trazabilidad aumenta la auditabilidad y la explicabilidad. Esta última es una propiedad altamente deseable en un sistema de IA, y debe estar relacionada con la capacidad de explicar tanto los procesos técnicos del sistema como las decisiones humanas asociadas. Es necesario que las decisiones que toma un sistema de IA sean comprensibles para los humanos. En algunos casos, puede ser posible que, en aras de aumentar la explicabilidad del sistema, se sacrifique ligeramente su precisión. Por último, cuando se presenta un sistema de IA a un usuario, hay que dejar claro a este que está interactuando con una máquina y no con otro humano, así como informarle de las capacidades y limitaciones del sistema.

5) Diversidad, no discriminación y equidad. Un sistema de IA fiable tiene que garantizar la inclusión y la diversidad a lo largo de todo su ciclo de vida. Es fundamental que los datos que usan los sistemas de IA no presenten sesgos, ya que, de otra forma, el sistema de IA aprenderá esos sesgos como una característica normal de los datos. Para evitar los sesgos, hay que poner énfasis en identificar información anómala en la fase de recopilación de la información y abogar por un proceso de supervisión que permita analizar de forma transparente el propósito, las restricciones, los requisitos y las decisiones del sistema. Además, también se debe garantizar el acceso a los sistemas de IA por parte de cualquier usuario sin discriminación por su edad, género, capacidades o características.

6) Bienestar social y ambiental. Se debe fomentar la sostenibilidad y la responsabilidad ecológica de los sistemas de IA, así como impulsar la investigación de soluciones de IA para resolver problemas y retos que afectan a toda la población, como los objetivos de desarrollo sostenible. Cuando se desarrolla un sistema de IA, es recomendable que, además de resolver correctamente el problema que se le ha encomendado, también lo haga del modo más respetuoso posible con el medio ambiente. Por ejemplo, hay técnicas de IA que necesitan realizar grandes cómputos y por ello consumen mucha energía, con el consiguiente impacto para el medio ambiente en forma de emisiones de CO_2, por lo que es necesario encontrar alternativas que tengan un menor uso energético. Por último, además de estudiar el impacto que los sistemas de IA ejercen sobre sus usuarios, también es deseable estudiar su impacto sobre las instituciones, la democracia y la sociedad en su conjunto.

7) Rendición de cuentas. Además de los requisitos anteriores, para que un sistema de IA sea fiable, es necesario establecer mecanismos que permitan garantizar la responsabilidad y la rendición de cuentas sobre el sistema y sus resultados, tanto antes de su implantación como después de esta. Los sistemas de IA deben ser auditables, tanto por auditores internos como externos, lo cual contribuye a la fiabilidad de la tecnología. Además, es preciso garantizar tanto la capacidad de informar sobre las acciones o decisiones que contribuyen al resultado del sistema como de responder a las consecuencias de dicho resultado, intentando minimizar sus efectos negativos. Finalmente, si se producen efectos adversos injustos, deben preverse mecanismos accesibles que aseguren una compensación adecuada, para que de este modo los usuarios tengan mayor confianza en los sistemas de IA.

5. Algunas áreas interesantes que utilizan inteligencia artificial en campos del derecho

Automatización de tareas legales

En Lexpo, uno de los mayores eventos europeos sobre innovación legal que se celebra anualmente en Ámsterdam, uno de los conferenciantes predijo ya en 2017 que, en los próximos cinco años, entre el 20 % y el 50 % de las tareas legales rutinarias se verán totalmente reemplazadas por la inteligencia artificial, que será capaz de automatizar muchos de los procesos sin que exista intervención alguna de las firmas de abogados.

Los cambios tecnológicos que están teniendo lugar han ocasionado el nacimiento del nuevo campo del **derecho digital**, ya que es necesario contar con una respuesta jurídica precisa para toda actividad relacionada con los servicios de la sociedad de la información y de la comunicación –que cuenta con su propia regulación, lenguaje y elementos tecnológicos–; y, además debe tenerse en cuenta el efecto transversal que las tecnologías, sobre todo internet, están suponiendo en nuestras vidas y, por ende, en el resto de las «tradicionales» disciplinas del derecho. Pero el otro gran reto es también el de la inteligencia artificial. Como hemos visto ya en la introducción, la inteligencia artificial es una disciplina que va a introducir, ya lo está haciendo, grandes cambios sociales y económicos durante los próximos años. El mundo del derecho, a pesar de su cierto carácter tradicional, no va a ser ajeno a esta revolución.

Veamos algunos campos en plena efervescencia:

1) Asesoramiento legal a emprendedores y *startups* (así como también a *business angels*, y fondos de inversión), que ya ofertan prestigiosos bufetes de abogados en España, asesorando en cuestiones como pueden ser problemas de inscripción o registro de marca, un mal reparto del capital social o la desprotección de las patentes. También asesoran en un aspecto muy importante en las empresas y *startups* digitales: la necesidad de asegurar la gestión de datos. En esta área, la dificultad es que se trabaja con clientes tecnológicos que desarrollan o emplean tecnología punta, y casi siempre el problema radica en que la regulación va por detrás de la tecnología, y es necesario planificar ciertos aspectos, previendo posibles futuros cambios y mejoras tanto como aspectos éticos que puedan confluir en el negocio.

2) Las llamadas iniciativas Legaltech, que intentan dar solución a situaciones relativamente sencillas en las que no se requiere intervención judicial, pero que para resolverlas se requiere la actuación de la justicia, y que van desde contratos laborales a reclamaciones a compañías de servicios (telefónicas, aerolíneas, bancarias, etc.), divorcios, etc. La idea es que la compañía preste un servicio legal a sus usuarios mediante una interfaz utilizable en dispositivos móviles, y que suele adoptar la forma de un *chatbot*. Un *chatbot* es un programa informático en línea que utiliza inteligencia artificial y nos permite mantener una conversación, para solicitar o recibir información o llevar a cabo acciones. Este tipo de empresas están en expansión en España desde hace un par de años, aunque ya tienen bastante más recorrido en países como Estados Unidos o el Reino Unido, y su objetivo es solucionar problemas legales coti-

dianos sin necesidad de acudir a un bufete. La idea central es similar a la de la banca electrónica, que hoy en día es un modo común de interaccionar con nuestro banco en línea.

Ejemplo de Legaltech

La empresa iUrisfy, que es una *app* que asesora y tramita divorcios de mutuo acuerdo, desarrollando toda la negociación entre las dos partes para que puedan comunicarse entre ellos y generar el documento legal del divorcio.

Este tipo de interacción cambia las relaciones entre abogados y clientes, y abarata también los costes; por ejemplo, en el caso de un divorcio, pueden reducirse prácticamente a la mitad. El uso de este tipo de empresas está en auge en nuestro país, y se espera un crecimiento importante en los próximos años.

3) Automatización de tareas rutinarias. Existen numerosos procedimientos en el ámbito del derecho en los que se hace un trabajo prácticamente manual y donde la cualificación del abogado realmente no aporta un valor diferencial. Por lo tanto, y como en otros ámbitos, estas tareas pueden ser realizadas por un sistema automatizado, aunque sea necesaria una supervisión letrada final. Esto se encuadraría dentro de lo que hemos visto en las líneas directrices para una IA fiable de la UE, en las que el humano es siempre el eje central de la toma de decisiones final. Al respecto, las principales firmas de abogados españolas han adquirido ya, o lo están analizando, tecnologías de inteligencia artificial para automatizar parte de sus procesos, y ya existen empresas especializadas en este campo, de las que pueden ser algunos ejemplos RAVN Systems, Luminance o Neota Logic.

Ejemplos: uso de sistemas basados en IA para operaciones comunes

Las operaciones *due diligence*, que consisten en que para realizar el asesoramiento en la posible compra de una compañía, es necesaria la verificación previa de todo tipo de información financiera, contractual y empresarial con el objetivo de fijar el valor real de la misma. Esta tarea puede ser larga, incluso podría llevar meses, pero al tener los clientes un tiempo de decisión limitado, ocasiona que este trabajo manual tenga que realizarse de forma intensiva en los bufetes, con los consiguientes aumentos de coste, por un lado, así como un posible aumento de fallos humanos debidos a sobreexceso de trabajo, que son aún peores en el caso de que no se produzca finalmente la compra de la compañía.

Otro ejemplo podría ser el uso de sistemas inteligentes en la resolución de pleitos como los de las cláusulas suelo, donde ya está reconocido el derecho del cliente, y que en la actualidad están apilados a la espera de ser resueltos, y que podrían resolverse en apenas unas horas.

El software inteligente es mucho más eficaz, pudiendo analizar la documentación en cuestión de horas y con márgenes de error prácticamente nulos, obteniendo costes de servicio más económicos, y lo que no es menos importante, permitiendo que los expertos en derecho usen su tiempo en aquellas tareas en las que realmente su experiencia es útil e imprescindible.

4) Predicción de sentencias usando aprendizaje automático. El sistema judicial español, como el de otros países, necesita ser más eficaz, y una de las posibles soluciones puede encontrarse en el uso de la inteligencia artificial en la predicción de sentencias, una posibilidad que está estudiando actualmen-

te el Consejo General del Poder Judicial. La gran ventaja de introducir esta tecnología se encuentra en la agilización de la velocidad burocrática, ya que permite establecer patrones de eficacia en, aproximadamente, un 85 % de los casos, siempre con supervisión humana experta. Uno de los países que figuran como adoptadores tempranos de la tecnología es Estonia, que ya es líder en la transformación por completo del país a través de la inteligencia artificial. Hoy en día, Estonia tiene una base de datos de 1,3 millones de ciudadanos en las que ya aplica IA y aprendizaje automático. Para Estonia, es importante que tengamos en cuenta la aproximación centrada en las personas de la IA europea (recordemos que en la UE los sistemas de IA deben apoyar a los humanos y respetar los derechos fundamentales, y no disminuir, limitar o desviar la autonomía humana). Esto requerirá mecanismos de supervisión adecuados, para evitar los posibles sesgos algorítmicos y de los datos, e incluyendo aproximaciones de tres tipos:

- capacidad de intervención humana en cada ciclo de decisión del sistema,
- capacidad de intervención humana durante el ciclo de diseño del sistema y la monitorización de las operaciones del sistema, y
- el humano al mando, es decir, capacidad de supervisar la actividad general del sistema de IA, incluidos sus impactos más amplios, y la capacidad de decidir cuándo y cómo usar el sistema en cualquier situación particular.

Contamos también con el Reglamento General de Protección de Datos, que se aplica en territorio de la Unión Europea desde mayo de 2018 y que, entre otros aspectos, regula el derecho a la explicación y la posibilidad de reclamación por parte de las personas. Por lo tanto, en realidad no es el proceso en sí el que cambia de forma sustancial, pero sí lo hace la velocidad de los trámites, permitiendo una aplicación más eficaz y eficiente de la justicia.

Ejemplo: el servicio Jurimetría y otros

Es una herramienta de analítica jurisprudencial que usa inteligencia artificial, que utilizan grandes despachos de abogados en nuestro país, y calcula si un litigio se va a ganar. Para ello, rastrea en unos minutos millones de sentencias (maneja un banco de más de diez millones de sentencias en España), buscando patrones similares al caso que está evaluando, y calcula el resultado más probable de la sentencia, el tiempo que previsiblemente tardará el juez en resolver el caso, e incluso la probabilidad de ganar la apelación en la audiencia, o el porcentaje de éxito del abogado de la parte contraria. Como resultado, ofrece además gráficos que le permiten al letrado evaluar si es mejor estrategia firmar un acuerdo extrajudicial que entrar en sala.

Un grupo de investigadores del University College London, la Universidad de Sheffield y la Universidad de Pennsylvania han desarrollado un algoritmo capaz de analizar los datos de casos de la Corte Europea de Derechos Humanos (CEDH), que ha logrado predecir un 79 % de las resoluciones alcanzadas en 584 asuntos. Procedimientos de este tipo que, por razones puramente formales se extienden durante años, se podrían resolver en pocos meses, agilizando considerablemente el sistema.

Hemos visto solamente algunos de los aspectos en los que la IA puede contribuir a la mejora en la eficacia de los procesos relacionados con el derecho, en ámbitos principalmente relacionados con el derecho de los negocios en los

que existan infracciones económicas o temas de competencia desleal, o en jurisdicciones como la fiscal –casos de contabilidad–, la civil –deudas, aseguradoras–, la administrativa –multas de tráfico– o en temas de marcas y patentes.

Recordemos que la aproximación de la UE es una aproximación centrada en lo humano, y por lo tanto, no acabará ni con la figura del juez ni con la del letrado. Las soluciones finales que recomiende el sistema deberán ser revisadas por el letrado; cualquier fallo generado por esta vía siempre deberá ser verificado y refrendado por un juez y, en caso de desacuerdo, siempre tendrá que existir una posibilidad de recurso.

La idea es disponer de sistemas que sirvan de apoyo a la toma de decisiones, aportando el conjunto de argumentos legales aplicables y proponiendo decisiones que deben ser refrendadas, pero con la ventaja de que eliminan las tareas tediosas y repetitivas y aligerarían los tiempos necesarios para la actuación de la justicia. Para que estas cuestiones planteadas se conviertan en realidad, es necesario también un cambio social; es imprescindible generar un salto cualitativo tanto tecnológico como cultural.

Resumen

Este módulo tiene como objetivo servir como una introducción de la disciplina de la inteligencia artificial para estudiantes de derecho, en el que se ha discutido la relación entre el campo de la inteligencia artificial y el derecho. Aunque *a priori* pueden parecer dos campos que no están muy relacionados, se ha demostrado que sí lo están, y como vimos desde el nacimiento de los sistemas expertos en los primeros años de la disciplina, hasta el momento actual. Pero durante los últimos años, sobre todo desde la entrada en vigor del RGPD, que no solo regula y controla el uso de los datos personales de los ciudadanos europeos, sino que también establece su derecho a requerir explicaciones cuando se han tomado decisiones sobre ellos que involucran a sistemas de inteligencia artificial, la relación se ha estrechado más si cabe, y ha hecho imprescindible el conocimiento de algunas nociones de la disciplina para los estudiantes de Derecho. En primer lugar, se ha descrito brevemente la historia de la inteligencia artificial, desde su nacimiento en la década de 1950 hasta su auge en la actualidad, pasando por distintos periodos en los que ha tenido más o menos popularidad y que se conocen como inviernos y primaveras. A continuación, se han descrito las distintas ramas que componen la disciplina de la inteligencia artificial, que pueden dividirse en dos subcampos: la IA simbólica y la subsimbólica. Ambas disciplinas tienen sus ventajas y sus limitaciones y son ampliamente usadas, aunque hoy en día se está popularizando más el uso de técnicas de aprendizaje automático, que se engloban dentro de la IA subsimbólica. Tomando como base el aprendizaje automático, se han explicado conceptos básicos sobre los algoritmos, recalcando la importancia de contar con unos datos adecuados para el aprendizaje y la conveniencia de usar técnicas y medidas de evaluación correctas. Seguidamente, se ha introducido el concepto de inteligencia artificial fiable, que es aquel sistema de IA en cuyas decisiones podemos confiar, ya que debe cumplir propiedades deseables como ser robusto, explicable, transparente, respetar la privacidad de los datos, ético, auditable, y velar por el bienestar social y ambiental. Para finalizar, se han comentado algunas áreas interesantes que utilizan IA en campos del derecho, como la predicción de sentencias usando técnicas de aprendizaje automático o la automatización de tareas rutinarias. Ante nosotros se abren nuevos retos para el derecho, en muchas de sus ramas y ámbitos clásicos, pero también seguramente en algunas áreas nuevas, ya que los sistemas de inteligencia artificial son la tecnología más transformadora en la actualidad, y la que más influencia tiene en la sociedad 4.0 que se está configurando ante nuestros ojos, a una gran escala y a rápida velocidad.

Bibliografía

Alonso Betanzos, A.; Guijarro Berdiñas, B.; Lozano Tello, A.; Palma Méndez, J. T.; Taboada Iglesias, M. J. (2004). *Ingeniería del Conocimiento: Aspectos metodológicos*. Pearson.

Ashley, K. D. (2000). *Designing electronic casebooks that talk back: The cato program*. Jurimetrics Journal (40, págs. 275-319).

Barker, V.; O'Conner, D. (1989). *Expert systems for configuration at digital: XCON and beyond*. Communications of the ACM (vol. 32, núm. 3, págs. 298-318).

Barrio Andrés, M. (2018). *Robótica, inteligencia artificial y derecho*. <http://www.realinstitutoelcano.org/wps/portal/rielcano_es/contenido?WCM_GLOBAL_CONTEXT=/elcano/elcano_es/zonas_es/ari103-2018-barrioandres-robotica-inteligencia-artificial-derecho>.

Berman, D. H.; Hafner, C. D. (1989). *The potential of artificial intelligence to help solve the crisis in our legal system*. Communications of the ACM (32, págs. 928-938).

Cáceres, E. (2008). *EXPERTIUS: A mexican judicial decision-support system in the field of family law*. En: Proceedings of the 2008 conference on Legal Knowledge and Information Systems: JURIX 2008 (págs. 78-87).

Casanovas, P. (2010). *Inteligencia artificial y derecho: a vuelapluma*. Revista de Pensamiento Jurídico (7, págs. 203-221).

Dua, D.; Graff, C. (2017). *UCI machine learning repository*. <https://archive.ics.uci.edu/ml/index.php>.

Fikes, R. E.; Nilsson, N. J. (1971). *Strips: A new approach to the application of theorem proving to problem solving*. Artificial Intelligence (2, págs. 189-208).

González, A. J.; Dankel, D. D. (1993). *The Engineering of Knowledge-based systems*Prentice-Hall.

Goodfellow, I.; Bengio, Y.; Courville, A. (2017). *Deep Learning*. MIT Press.

Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial (2019). *Directrices éticas para una IA fiable*Comisión Europea.

Hafner, C. D. (1987). *Conceptual organization of case law knowledge bases*. En: Proceedings of the 1st international conference on Artificial intelligence and Law- ICAIL (págs. 35-42). ACM.

Hernández Jiménez, M. (2019). *Inteligencia artificial y derecho penal*. Actualidad Jurídica Iberoamericana (10bis, págs. 792-843).

Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*John Wiley & Sons.

LeCun, Y.; Bengio, Y.; Hinton, G. (2015). *Deep learning*. Nature (521, págs. 436-444).

McCarty, L. T. (1977). *Reflections on taxman: An experiment in artificial intelligence and legal reasoning*. Harvard Law Review (90, págs. 837-893).

Miller, R. A.; Pople Jr., H. E.; Myers, J. D. (1982). *Internist-1, an experimental computer-based diagnostic consultant for general internal medicine*. New England Journal of Medicine (307, págs. 468-476).

Newell, A.; Simon, H. (1963). *GPS, A Program That Simulates Human Thought*Nueva York: McGraw-Hill.

Parlamento Europeo y Consejo de la Unión Europea (2018). *Reglamento General de Protección de Datos*Diario Oficial de la Unión Europea.

Russel, S.; Norvig, P. (2018). *Artificial Intelligence: A Modern Approach*Pearson.

Scarl, E. A.; Jamieson, J. R.; Delaune, C. I. (1987). *Diagnosis and sensor validation through knowledge of structure and function*. IEEE Transactions on Systems, Man, and Cybernetics (vol. 17, núm. 3, págs. 360-368).

Schreiber, G.; Akkermans, H.; Anjewierden, A.; de Hoog, R.; Shadbolt, N.; Van de Velde, W.; Wielinga, B. (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press.

Staff of the Heuristic Programming Project (1980). *The Stanford heuristic programming project: Goals and activities*. Artificial Intelligence Magazine (1, págs. 25-30).

Turing, A. M. (1950). *Computing machinery and intelligence*. Mind (59, págs. 433-460).

Waterman, D. A.; Peterson, L. A. (1984). *Evaluating civil claims: an expert systems approach*. Expert Systems (1, págs. 65-76).