Intel·ligència artificial, algorismes i dret: una introducció

PID_00270379

Amparo Alonso Betanzos
Verónica Bolón Canedo

Temps mínim de dedicació recomanat: 4 hores

Universitat Oberta de Catalunya

Amparo Alonso Betanzos

Catedràtica de Ciències de la Computació i Intel·ligència Artificial a la Universitat de la Corunya, on coordina el grup LIDIA (Laboratori d'R+D en Intel·ligència Artificial) del Centre de Recerca en TIC (CITIC). És llicenciada en Químiques (1984) i doctora en Físiques (1988) per la Universitat de Santiago de Compostel·la. Ha estat *Postdoctoral Fellow* al Medical College de Geòrgia (1988-1990), EUA., on va treballar en temes relacionats amb el desenvolupament de sistemes experts per a aplicacions mèdiques.

La seva àrea de recerca és el desenvolupament i l'aplicació de tècniques d'intel·ligència artificial en diverses àrees i també l'aprenentatge computacional i les tècniques de ciència de dades (*big data*). Ha participat com a investigadora principal en més de vuitanta projectes de recerca competitius i projectes de transferència. És autora de noranta-cinc articles científics en revistes científiques, cent setanta articles en congressos, la majoria internacionals, i vint-i-cinc llibres i capítols de llibres. El 1998 va rebre el premi L'Oréal-Unesco for Women in Science a Espanya, el 2004 el premi Galícia TIC a la Innovació, i el 2019 el premi Galícia TIC a la Trajectòria Professional. Actualment, i des de 2012, és presidenta de l'Associació Espanyola per a la Intel·ligència Artificial.

Verónica Bolón Canedo

Enginyera en Informàtica (2009) i doctora en Informàtica (2014) per la Universitat de la Corunya (UDC). Després d'una estada postdoctoral a la Universitat de Manchester (Regne Unit) el 2015, actualment és professora ajudant doctora en el Departament de Ciències de la Computació i Tecnologies de la Informació de la UDC, on està integrada en el grup de recerca LIDIA (Laboratori d'R+D en Intel·ligència Artificial) del CITIC. Ha impartit docència a la Universitat de Manchester i a la Universitat de la Corunya en assignatures relacionades majoritàriament amb la intel·ligència artificial i l'aprenentatge màquina.

És autora de dos llibres, diversos capítols de llibres i més de seixanta articles en congressos internacionals i revistes. Ha coorganitzat diverses sessions especials en congressos internacionals en temes relacionats amb *big data*. La seva tesi doctoral ha rebut el premi extraordinari de doctorat de la UDC, el premi a la millor proposta predoctoral (2011), el premi a la millor tesi espanyola en IA (2014) i el premi Frances Allen a la millor tesi en IA defensada per una dona (2015), aquests tres últims atorgats per l'Associació Espanyola per a la Intel·ligència Artificial.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Mònica Vilasau Solana

Primera edició: febrer 2020

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

Índex

Introducció

Actualment ens trobem immersos en un canvi social i tecnològic sense precedents. La digitalització és un procés intens i progressiu, que genera grans quantitats de dades de pràcticament qualsevol activitat. La connectivitat s'ha convertit en una necessitat, que genera reptes d'adaptació social, però també oportunitats de mercat immenses en qualsevol camp. A més, aquestes dades poden ser analitzades de manera ràpida i viable econòmicament a causa de l'abaratiment de la computació en núvol i als avenços tant en maquinari com en software per a obtenir coneixement útil a partir d'elles. Com a conseqüència d'això, la demanda de sistemes intel·ligents per part de les empreses augmenta cada any.

L'expansió d'aquests sistemes intel·ligents en diversos dominis està comportant no solament grans canvis de tipus econòmic sinó grans reptes d'adaptació social que impliquen en conseqüència la necessitat de reglamentació i legislació. D'una banda, hi haurà canvis en la forma en què es produeixen comunicacions i interaccions en les companyies; per exemple, es preveu que un percentatge important (prop del 20%) dels continguts publicats per aquestes companyies (documentació legal, comunicats, informes...) seran elaborats per sistemes intel·ligents. S'està produint un canvi important en els canals de servei al client, que es gestionaran en un percentatge molt ampli (prop del 85%) usant sistemes intel·ligents, la majoria en forma de *chatbots*, que s'ocuparan de manera més directa i personalitzada dels gustos i necessitats dels clients, als quals poden atendre durant els 365 dies de l'any, les 24 hores. Els executius de les empreses usaran software de reconeixement de veu perquè els seus assistents personals intel·ligents els ajudin a organitzar el seu treball de manera més eficaç i eficient. Els vehicles autònoms per al transport de mercaderies seran una realitat a curt termini, però per als nostres desplaçaments potser també ho seran a mitjà termini. Aquestes opcions permetran una millora substancial en la manera de gestionar el trànsit a les nostres col·lapsades ciutats, a més de millorar la nostra vida personal i laboral proporcionant-nos temps lliure, que ara mateix perdem conduint, intentant buscar aparcament o patint embussos. Sorgeixen noves àrees d'aplicació de la intel·ligència artificial (IA), com les relacionades amb les empreses de tecnologia financera (FinTech) o les assegurances relacionades amb l'ocupació d'aquests sistemes en qualsevol àmbit, molt especialment en temes d'educació o salut.

Òbviament, aquests canvis generen la necessitat de més reglamentació i legislació. La majoria dels països han desenvolupat plans estratègics que s'enfronten no solament al repte tecnològic i a la imprescindible recuperació i foment del talent humà sinó a temes relacionats amb l'ètica, la reglamentació i els canvis del model d'ocupació. En concret, la Unió Europea (UE) proposa una visió estratègica basada en una IA ètica, sostenible, robusta, confiable i d'avantguarda *made in Europe*, que es concreta en el document de «Directrius ètiques per a una IA confiable». Aquesta IA es recolza en tres components que s'han de satisfer al llarg de tot el cicle de vida del sistema:

Lectura recomanada

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

1) L'IA ha de ser lícita, i en conseqüència ha de complir les normes vinculants europees i d'àmbit nacional i internacional que ja són aplicables al desenvolupament i ús dels sistemes que utilitzen tècniques d'IA. Entre aquestes normes hi ha el dret primari (tractats de la UE i Carta de Drets Fonamentals) i el dret secundari (Reglament general de protecció de sades, Directiva sobre màquines, Directiva sobre discriminació, etc.) de la UE, els convenis del Consell d'Europa, el tractat de Drets Humans de l'ONU i normes de caràcter nacional, en cadascun dels estats membres, o sectorial, com el Reglament sobre productes sanitaris, entre altres reglaments i lleis.

2) L'IA ha de ser ètica, de manera que es garanteixi el respecte als principis i valors ètics.

3) L'IA ha de ser robusta tant des del punt de vista tècnic com social, ja que els sistemes d'IA poden provocar danys accidentals fins i tot si les intencions són bones.

Cadascun d'aquests components és necessari per si mateix però no suficient per a assolir una IA fiable. L'ideal és que tots actuïn en harmonia i de manera simultània. Però no tots els possibles aspectes són susceptibles de ser reglamentats, i en la pràctica pot ser relativament comú que apareguin tensions entre objectius diferents (per exemple, la transparència dels algorismes pot obrir la porta a fer-ne un mal ús, i identificar i corregir biaixos en les dades podria contrastar amb la privadesa). És important protegir-se d'aquestes situacions, comunicar-les i documentar-les.

També hi ha connexions entre la reglamentació i la recerca. El maig de 2018 va entrar en vigor el nou Reglament europeu de protecció de dades (RGPD), que no solament regula i controla l'ús de les dades personals dels europeus, sinó que estableix el seu dret a demanar una explicació quan es prenguin decisions respecte els afectats que hagin estat adoptades totalment o en part per algorismes intel·ligents. El reglament ha facilitat la recerca per a poder dotar de transparència i explicabilitat a algorismes que fins ara proporcionen respostes i judicis de manera opaca. Un exemple representatiu són els models d'aprenentatge profund (*deep learning*) disruptius, els desenvolupadors del qual van rebre la medalla Turing, patrocinada per Google des del 2014 i que és l'equivalent al premi Nobel en aquest camp. Avui dia aquests models

Lectures recomanades

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* (núm. 521, pàg. 436-444).

d'aprenentatge profunds són un component crític de la computació perquè han demostrat produir resultats d'avantguarda i de gran exactitud en diferents tasques. Però aquests models tan exactes tenen el problema de comportar-se com una «caixa negra», és a dir, són opacs, de manera que oculten la lògica interna a l'usuari, i tenen a més una complexitat important. Aquest aspecte estableix una tensió entre l'exactitud i la interpretabilitat, i és no solament un tema d'índole pràctica sinó també un assumpte ètic. La capacitat d'interpretar un model és extremadament important, especialment en alguns camps, com per exemple la bioinformàtica, la medicina personalitzada o els vehicles autònoms, ja que en general augmenta la confiança dels usuaris, dona suport a la comprensió del procés que es modela i proporciona informació sobre com es pot millorar el model. És a dir, és rellevant no solament per temes de confiança, sinó perquè pot contribuir al descobriment científic i al progrés de la recerca en diversos camps. Ara mateix és rellevant també en el context de l'RGPD, que, tal com hem esmentat anteriorment, legisla entre d'altres el dret a obtenir una explicació comprensible quan una persona rep una decisió d'un sistema basat en IA, per exemple, respecte a la concessió d'un crèdit, asil polític, una assegurança, etc.

Hi ha un acord general en el fet que la necessitat d'implementar aquest principi és urgent i que en aquest moment és un repte científic molt important. Si la tecnologia no és capaç de proporcionar aquesta explicació, aquest dret es converteix en paper mullat. Un altre risc important que també apareix és la possibilitat que aquests models prenguin decisions equivocades distretament, per exemple, a causa del fet que les dades que han usat contenen biaixos i prejudicis. Un exemple d'això és el cas de l'escàndol que hi va haver amb l'eina de contractació d'Amazon, tancada el 2017 perquè discriminava les dones i confonia la seva poca presència a l'àrea TIC (tecnologies de la informació i les comunicacions) amb incapacitat en el treball. Les tecnologies explicatives que hem de desenvolupar són necessàries perquè les companyies creïn productes més segurs, més confiables i que puguin manejar qualsevol responsabilitat legal en què puguin incórrer. Per exemple, pensem en la seva utilitat en el cas de l'accident del cotxe autònom d'Uber, en el qual l'error va ser degut a un ajustament de baixa sensibilitat davant els objectes de la carretera. L'objectiu era que el cotxe pogués reaccionar i prendre accions davant objectes sòlids grans i no davant objectes innocus, com bosses per exemple. D'aquesta manera, el vianant va ser detectat, però el software va decidir no parar, amb la greu conseqüència d'una persona morta. Un altre problema és que els models de xarxes neuronals profundes (DNN, *deep neural networks*) són capaços d'aconseguir exactituds importants en el reconeixement de text i imatges, fins i tot per sobre del nivell humà, però alhora també són susceptibles de ser enganyats per atacs anomenats *adversarial* o d'adversaris, que introdueixen variacions que passarien inadvertides per als humans, però que són suficients per a enganyar un classificador. Per aquests i altres motius, l'explicabilitat dels algorismes és un tema important en recerca avui dia. De moment, la majoria dels escassos mètodes desenvolupats fins ara són locals, és a dir, ens proporcionen les raons d'una decisió específica. Solament és interpretable una decisió concreta. Altres

limitacions a tenir en compte són temporals, ja que, si és necessari prendre una decisió ràpida (per exemple, en un entorn en què un desastre és imminent), és preferible un model d'explicació simple perquè l'usuari humà pugui prendre una decisió. Si el temps no és una restricció, per exemple, en un problema de concessió d'un crèdit, pot ser viable un model explicatiu més complex i exhaustiu. Altres aspectes a considerar en el tema de l'explicabilitat tenen a veure amb assumptes ètics, com en el cas de la justícia, que garanteix la protecció de certs grups enfront de la discriminació, i la privadesa, que aconsegueix que el model d'explicació no reveli informació sensible sobre persones. Finalment, entre altres aspectes desitjables, l'escalabilitat i la portabilitat dels models, tant els d'aprenentatge com els d'explicació, són del màxim interès si fossin diferents. En molts dels models d'explicació es torna la mirada cap als models simbòlics, més interpretables en els àmbits local o global, com en el cas dels arbres de decisió o les regles.

Així mateix, és necessari que anem més enllà en temes de regulació i d'ètica: és necessari legislar en àrees sensibles, com l'educació o la salut, però també fomentar un debat social i polític d'aspectes també controvertits de l'IA, com en el cas de les notícies falses, les armes autònomes o l'ús de sistemes de reconeixement facial. També serà necessari regular temes que afecten els models de transició del nostre model productiu actual, basat més en la mà d'obra humana i que ha de canviar per a afrontar els nous models de treball del futur basats més en l'automatització de moltes tasques rutinàries. Algunes qüestions ja s'han debatut a la UE, i alguns dels països membres fins i tot han provat algunes mesures, com la renda bàsica universal. Les conclusions de l'experiment sobre la implantació d'aquesta renda que es va dur a terme a Finlàndia, el primer país del món a provar la mesura al principi de l'any 2017, són que el seu ús no ofereix millors perspectives laborals, però sí que augmenta considerablement el nivell de benestar i sensació de justícia social que perceben els ciutadans. L'experiment va escollir una mostra relativament petita de població, dues mil persones aturades, i en la franja d'edat de 25-58 anys, a les quals es van concedir 560 euros mensuals lliures d'impostos sense tenir en compte si buscaven treball o no de manera activa. Els resultats són que els beneficiaris de la renda bàsica no van mostrar diferències respecte als altres ciutadans a l'hora de trobar treball en un mercat laboral obert, però van tenir menys símptomes d'estrès, menys dificultat per a concentrar-se i menys problemes de salut, a més d'una major confiança en el futur i una major capacitat per a influir en els problemes socials. Quin és el problema? L'Organització per a la Cooperació i el Desenvolupament (OCDE) va concloure, en el seu informe elaborat en quatre escenaris hipotètics al Regne Unit, França, Itàlia i Finlàndia, que és impossible implantar un sistema d'aquestes característiques sense una reforma substancial del model tributari, ja que la seva generalització podria incrementar la pobresa des de l'11,4% actual fins al 14,1% si es financés amb el pressupost actual destinat a ajudes socials, perquè requeriria augmentar els impostos als salaris en gairebé el 30%.

En l'actualitat l'entrada de l'IA en el món del dret ja és un fet i planteja molts reptes interessants. A continuació en veurem alguns aspectes bàsics per ajudar-nos a entendre millor la seva rellevància.

1. Història de la intel·ligència artificial

L'IA és una àrea de les ciències de la computació, i per tant de la informàtica, que s'ocupa de crear programes informàtics que executen operacions comparables a les que fa la ment humana, com l'aprenentatge o el raonament lògic.

En aquesta disciplina hi ha diverses subàrees de treball:

- la robòtica
- l'aprenentatge automàtic
- el processament del llenguatge natural
- la visió artificial

A més, L'IA afecta molts camps d'aplicació i nombrosos aspectes de la societat i, per tant, és cada vegada més interdisciplinària, ja que implica i incorpora cada vegada més professionals procedents d'altres disciplines, com l'economia, la filosofia, les matemàtiques o el dret, entre d'altres.

L'IA es coneix amb aquest nom des de l'any 1956, quan un reduït grup de científics (procedents de diferents disciplines, com el control, la cibernètica o la teoria d'autòmats) es va reunir en un seminari d'estiu a Dartmouth College (Estats Units) per reflexionar sobre la pregunta «poden pensar les màquines?», que el científic britànic Alan Turing, considerat el pare de l'IA, havia enunciat en un article científic publicat en la revista *Mind* el 1950. En aquest article Turing defensava que molt probablement les màquines podrien competir amb els éssers humans en camps merament intel·lectuals. Moltes de les idees que plantejava Turing existeixen actualment com a models d'aprenentatge automàtic (supervisat, no supervisat, evolutiu, per reforç) o com a models usats en robòtica (la robòtica del desenvolupament, que adopta idees de l'aprenentatge humà per escurçar les escales de temps de l'aprenentatge automàtic), mentre que d'altres, com la creativitat computacional o les interfícies humanitzades, són camps en auge que encara presenten molt territori per explorar.

Lectura recomanada

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

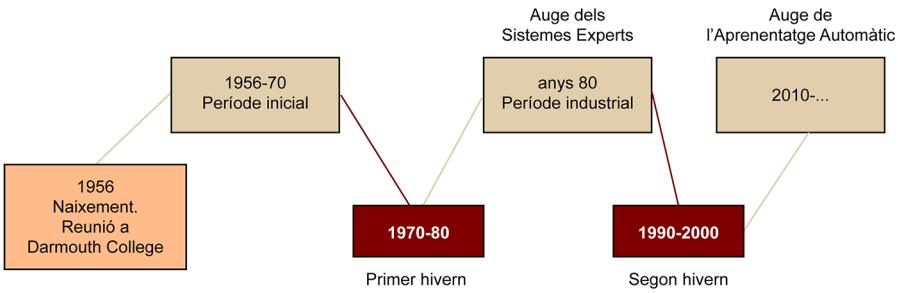
En aquesta reunió al Darmouth College, finançada en part per IBM, els científics assistents ja van començar a plantejar-se com seria possible construir aquestes màquines intel·ligents, i van ser els quatre principals organitzadors, John McCarthy, Marvin Minsky, Nathaniel Rochester i Claude Shannon, que van escollir el nom *intel·ligència artificial* per denominar aquest camp nou. En aquesta reunió la proposta assumia que qualsevol aspecte de l'aprenentatge o

qualsevol altra característica de la intel·ligència podria ser descrit de manera que **una màquina pogués simular-la**. L'objectiu consistia a aconseguir avenços significatius en les àrees següents:

- la comprensió i ús del llenguatge natural en les màquines
- la construcció d'abstraccions i conceptes per part d'aquestes màquines
- la resolució de problemes complexos
- les xarxes neuronals
- l'aprenentatge
- la creativitat computacional

Alguns dels investigadors assistents van ser McCulloch i Pitts (autor del seu model neuronal del cervell), Von Neumann (amb projectes per a construir el computador ENIAC i el programa EDVAC), Minsky i Edmonds (autor del primer computador neuronal) i Shannon (introductor de l'heurística i programes d'escacs). Encara que en el seminari solament va haver-hi deu treballs, els ponents i els seus deixebles van marcar el camp de l'IA durant els trenta anys següents.

Figura 1. Línia de temps del desenvolupament de la disciplina d'IA



Font: elaboració pròpia

Totes les qüestions que es van plantejar en aquesta reunió van demostrar que eren molt importants i han estat presents com a grans reptes del camp des del seu naixement. En el seminari es va abordar un paradigma cognitiu del cervell humà, que tenia l'objectiu de **reproduir les capacitats cerebrals mitjançant algorismes** amb la premissa que els estats mentals podrien funcionar de manera anàloga als programes en una computadora. Quant a la manera de construir aquests algorismes, van aparèixer dues línies de pensament diferents:

- Un grup, que va rebre el nom de «pulcres», mantenia que la representació simbòlica es basava principalment en la lògica (és a dir, en la sintaxi i el càlcul de predicats).
- L'altre grup, anomenats «descurats», pensava que es basava principalment en la semàntica.

La primera de les aproximacions, la «pulcra», va optar per solucions elegants i properes a les matemàtiques, mentre que la segona, dels «descurats», pretenia que la intel·ligència és massa complicada, probablement intractable computacionalment i, per tant, que no es podria resoldre amb el tipus de sistema ho-

mogeni que exigeix requisits precisos. I sota el guiatge de cadascuna d'aquests dos corrents han aparegut les dues etapes més brillants (anomenades primaveres) de l'IA (vegeu-ne la línia temporal en la figura 1).

La primera primavera ha sorgit sota el guiatge de l'aproximació simbòlica, relacionada amb la branca «descurada», amb el naixement dels anomenats «sistemes experts», que van demostrar la factibilitat d'usar IA per a resoldre problemes del món real amb sistemes com DENDRAL, MYCINO o PROSPECTOR, mentre que la segona primavera ha arribat amb els nous enfocaments de l'aprenentatge profund, el reconeixement del llenguatge natural i altres enfocaments de l'aprenentatge automàtic, com l'aprenentatge per reforç, provinents de la branca «pulcra». El futur potser ens oferirà una fusió entre ambdues per fer front als nous reptes que planteja l'IA confiable i robusta, amb processos en què es garanteixi la transparència, la reversibilitat i la traçabilitat.

Durant la primera època, després del seu naixement, la disciplina va viure un moment d'esplendor gràcies als avenços que es produïren en camps com l'abstracció de conceptes mitjançant aprenentatge (amb el desenvolupament de programes que eren capaços de demostrar alguns teoremes sobre lògica proposicional –com per exemple el Logic Theorist– o sobre geometria), models sobre com els humans resolem problemes (com el General Problem Solver: GPS) o mètodes d'aprenentatge per a emprar en àrees de jocs, com els escacs o les dames. Però després d'aquest entusiasme inicial, al principi dels anys setanta, començaren a aparèixer les primeres crítiques a l'IA, ja que els algorismes desenvolupats tenien greus problemes d'escalabilitat a l'hora de poder ser usats en entorns reals complexos, i també contribuí a aquesta caiguda en desgràcia el relatiu fracàs del processament del llenguatge natural, que s'assumia com a mitjà de comunicació lògic entre la màquina i l'ésser humà. La capacitat dels ordinadors del moment era molt lluny de permetre implementar moltes de les idees d'aquests pioners de l'IA.

Exemple de la capacitat dels ordinadors

Per a fer-nos una idea, les aplicacions pràctiques de la visió per computadora requereixen prop de 10.000-1.000.000 de processadors MIPS, una xifra molt més gran que els 80-130 MIPS del supercomputador més ràpid del món el 1976, Cray-1, un gegant comparat amb un ordinador personal típic de l'època, amb menys d'1 MIPS.

Així i tot, també es van aconseguir èxits en aquesta època, entre els quals hi ha els treballs de Fikes i Nilsson el 1971, que van donar com a resultat:

- El sistema de planificació STRIPS (les idees del qual estan vigents encara avui);

- els sistemes de Newell i Simon el 1963, que van desembocar en el GPS o solucionador general de problemes, imitaven els protocols de resolució de problemes dels humans en la seqüència d'objectius i possibles cursos d'acció que generava;

- el programa de les dames de Samuel, que va aconseguir batre el seu creador;

- els treballs de McCarthy, qui va definir el llenguatge LISP i va inventar el temps compartit per aconseguir més recursos de còmput.

Lectures recomanades

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

A. Newell; H. Simon (1963). *GPS, a program that simulates human thought*. Nova York: McGraw-Hill.

També es van fer treballs importants que en dominis limitats resolien problemes per als quals es requeria intel·ligència, com problemes d'àlgebra, geometria o càlcul. Malgrat això, a la fi dels anys seixanta l'IA es va trobar amb molts problemes:

1) Els mètodes s'havien desenvolupat per resoldre problemes de caràcter molt general, i els programes contenien poc o cap coneixement sobre el domini del problema a modelar. Això no comporta grans problemes en dominis simples, però sí en dominis reals, en els quals els problemes són més variats o de major dificultat.

Problemes NP-complets

Al principi es va creure que la solució seria escalable, però no va resultar així, ja que la majoria dels problemes interessants per l'IA són NP-complets. De manera senzilla, els problemes que són solubles en temps polinomial són tractables, i els que no ho són són NP-complets i se solen conèixer com a intractables (aquest tipus de problemes se solen resoldre amb aproximacions).

2) En aquell moment un camp interessant per a l'IA era (i continuaria essent) el llenguatge natural i especialment els traductors automàtics. En aquest camp les dificultats apareixien per la falta d'un context particular i de nocions generals sobre situació i sentit comú. Això va provocar la cancel·lació econòmica d'aquests projectes per part dels Estats Units el 1966.

3) A Europa els científics més destacats en IA van ser principalment britànics. El 1971 el Govern britànic va cancel·lar el suport econòmic a l'IA basant-se en l'*Informe Lighthill*, que va concloure que les aportacions d'aquesta disciplina no eren rellevants i, per tant, va tallar el finançament en aquesta àrea.

La disciplina entra així en el seu primer hivern, un període de descoratjament i de certa sensació de fracàs, que a més comporta una forta retallada econòmica. Aquest primer hivern acaba als anys vuitanta, en els quals sorgeixen amb força un tipus de sistemes d'IA anomenats **sistemes experts**, que s'adopten amb èxit en molts entorns reals. Aquest èxit es basa en un canvi d'orientació: s'oblida l'objectiu d'aconseguir una IA de tipus general, similar a la humana, i redirigeix el focus a una IA específica, centrada en camps concrets. La idea era que alguns dels problemes identificats en els mètodes que es manejaven en aquells dies es podrien resoldre si s'introduïa en els sistemes, a més del coneixement general que permet fer inferències, coneixement específic sobre el domini d'aplicació.

Aquesta postura va portar els investigadors a afirmar que el coneixement es podia adquirir dels experts humans per a ser «transferit» a un computador mitjançant una representació que aquest pogués manipular.

Neixen així els sistemes experts, que es poden definir com a sistemes amb el propòsit d'emular la capacitat de resolució de problemes d'un ésser humà en un domini específic utilitzant-ne el mateix coneixement.

Aquest primer èxit és mèrit d'aproximacions simbòliques, que usen sobretot models basats en regles de producció que codifiquen el comportament d'experts humans per resoldre problemes concrets. Després d'aquest període fructífer l'IA cau en un segon hivern a causa de la dificultat que planteja mantenir aquests sistemes experts especialitzats i al col·lapse del mercat de maquinari especialitzat (necessari fins llavors per a desenvolupar els sistemes experts) amb motiu de la impressionant millora de prestacions dels ordinadors de propòsit general que coneixem avui. Després d'aquest hivern l'IA comença a despertar de nou i es fragmenta en camps especialitzats (sistemes multiagent, robòtica, sistemes basats en coneixement, etc.), que obtenen certs èxits importants.

Deep Blue

Com a exemple il·lustratiu, l'any 1997 Deep Blue es convertí en el primer programa d'escacs que derrotà un vigent campió del món, Kasparov. Cal tenir en compte que aquesta victòria no va ser deguda tant a la millora de l'aproximació intel·ligent com a la millora del maquinari, ja que aquest nou Deep Blue era una versió millorada d'un sistema anterior, que va ser derrotat per Kasparov però que ara era capaç de processar el doble de moviments per segon que en la primera versió.

No és fins al voltant de l'any 2010 que l'IA salta a la indústria com una tecnologia imprescindible per la conjugació de diversos factors:

1) Disposem d'enormes quantitats de dades a causa del procés de digitalització progressiu i intens que té lloc. Pràcticament qualsevol experiència humana està digitalitzada: els viatges, la música, els serveis de salut, etc. A més, disposem també d'una creixent quantitat de sensors, cada vegada més exactes, que registren dades procedents de pràcticament qualsevol procés, i que ens permeten saber, per tant, com es comporta i evoluciona el nostre entorn.

2) El canvi social i tecnològic que travessem no té precedents. Tots nosaltres interactuem amb els nostres dispositius mòbils en un **món cada vegada més interconnectat** i en el qual disposar de connectivitat a qualsevol hora i en qualsevol lloc és imprescindible. Aquesta connectivitat genera tant una oportunitat de mercat per a les empreses com un repte d'adaptació social.

3) Ara disposem de la **capacitat de còmput** necessària perquè els programes que dissenya l'IA puguin trobar solucions de gran exactitud en un temps adequat. Diversos elements permeten el processament ràpid i rendible econòmicament d'enormes volums de dades heterogènies:

- l'avenç de les tecnologies de computació d'altes prestacions
- l'abaratiment de la computació en núvol
- la disponibilitat d'ús de noves plataformes de computació paral·lela i distribuïda

4) També és decisiu el fet que disposem d'**avenços molt rellevants en software**. Hem aconseguit desenvolupar nous tipus de bases de dades que ens permeten emmagatzemar i tractar dades estructurades i no estructurades més enllà de la clàssica dada científica de les bases de dades clàssiques. Al mateix temps ha estat disruptiva l'aparició de nous desenvolupaments teòrics, fonamentalment matemàtics, com els que s'han obtingut en el camp de l'aprenentatge profund (*deep learning*), l'aprenentatge per reforç o el reconeixement del llenguatge natural, que han generat resultats d'alt nivell de precisió, de manera que han situat l'IA com una tecnologia madura exitosa i de gran impacte.

5) Com a conseqüència de tot això, **augmenta la demanda real de les empreses**, que multipliquen anualment les seves inversions econòmiques en aquest camp. Aquesta inversió ajuda també a que la tecnologia d'IA es consolidi, ja que podem constatar que les empreses que són usuàries o desenvolupadores d'aquesta tecnologia són les que triomfen més en els camps respectius. L'any 2019 set de les deu empreses més importants del món en capital borsari són usuàries o desenvolupadores d'aquesta tecnologia. Aquesta situació era molt diferent tan sols deu anys enrere, quan solament tres de les deu primeres empreses del món pertanyien a aquest sector.

Inversió en intel·ligència artificial

La inversió global en noves empreses d'IA s'ha multiplicat per nou entre els anys 2011 i 2015 segons el Fòrum Econòmic Mundial i ha continuat creixent encara més des de llavors.

Però, a més, al llarg de la seva història l'IA s'ha convertit en una disciplina transversal, que afecta molts camps d'aplicació, com la salut i medicina, l'educació, el medi ambient, la indústria, el turisme, etc. Com a conseqüència d'això, l'IA és un camp cada vegada més interdisciplinari, que incorpora professionals procedents d'altres disciplines com el dret, la psicologia, la sociologia, l'economia, etc.

En la reunió de Darmouth College es va establir també que l'objectiu últim de l'IA era aconseguir reproduir intel·ligència en una màquina de manera que pogués, programada adequadament, replicar la intel·ligència humana i exhibir un comportament intel·ligent de tipus general. Aquest és encara un objectiu molt ambiciós, que sembla que encara estem lluny d'aconseguir. Però sí que hem pogut desenvolupar una IA específica, en la qual algorismes i màquines són capaços de fer tasques associades a la intel·ligència humana com aprendre, entendre o raonar i requereixen intel·ligència en un àmbit concret i especialitzat però no exhibeixen un comportament intel·ligent de tipus general.

Exemple: un camp de jocs

Hem aconseguit desenvolupar programes que juguen a escacs o a go i que són capaços de batre els campions del món i els grans mestres. Però no tenen una intel·ligència general: si canviem les regles del joc o pretenem jugar a un joc similar, en principi necessitarem desenvolupar un algorisme diferent (encara que es treballa intensament en la transferibilitat de l'aprenentatge). No obstant això, qualsevol jugador d'escacs humà és capaç d'aprofitar els seus coneixements per a aprendre a jugar de manera ràpida a les dames, per exemple.

Malgrat aquesta limitació, hi ha molts dominis en què l'IA supera la intel·ligència humana, com en el cas d'àrees específiques de la medicina, els sistemes de recomanació, els robots, els vehicles autònoms, els assistents personals o els traductors automàtics. Tots aquests sistemes canvien la nostra manera d'interaccionar amb l'entorn i comporten importants canvis econòmics i socials, que necessiten regulació i reglamentació.

2. Branques de la intel·ligència artificial

Hi ha diverses classificacions possibles que podem usar per a parlar dels diferents camps de la intel·ligència artificial. Ara veurem algunes d'aquestes divisions i així anirem definint alguns conceptes bàsics. Una forma de classificació clàssica és distingir entre la intel·ligència artificial simbòlica i la intel·ligència artificial subsimbòlica.

Lectura recomanada

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

2.1. Intel·ligència artificial simbòlica

També s'anomena simbòlic-deductiva o convencional.

Aquesta branca sorgeix a l'inici de la disciplina i es basa en la idea definida en la hipòtesi de sistemes de símbols físics, de Newell i Simon, que defensa bàsicament que la majoria dels aspectes de la intel·ligència es poden modelar usant representacions simbòliques d'alt nivell dels problemes a modelar, amb les eines de la lògica matemàtica i la cerca.

El paradigma simbòlic va ser el dominant des dels anys cinquanta fins als vuitanta i va produir l'afortunat paradigma dels sistemes experts, que, tal com hem vist, marca l'arribada de la primera de les primaveres que ha viscut l'IA. Per les seves característiques, els sistemes experts es podien aplicar a moltes àrees de l'activitat humana, i d'aquí el seu gran èxit, i un dels principals camps d'experimentació i aplicació ha estat la medicina, però després s'han estès a pràcticament qualsevol àrea de l'activitat humana en què es necessita coneixement expert, i entre aquestes hi ha naturalment diverses branques del dret. Els sistemes experts es defineixen com a sistemes intel·ligents que contenen coneixement explícit d'alt nivell d'un camp d'aplicació complex, encara que restringit. La manera en què aquest coneixement explícit es representa està basat en la lògica, i el paradigma de representació més utilitzat són les **regles de producció**.

Lectura recomanada

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

Les regles de producció són elements que tenen una estructura amb un antecedent (part condició de la regla o part *if*) i un conseqüent (part *then* o part conclusió de la regla), i, depenent del llenguatge de programació usat, poden tenir una part de conclusió alternativa (part *else*), que es deduiria si fallés la part antecedent, és a dir, si no es complissin les clàusules. Qualsevol de les parts pot acollir diverses clàusules amb la forma següent:

REGLA X

IF condició 1 AND condició 2 AND...condició n
THEN conclusió 1 AND conclusió 2 AND... conclusió m
ELSE conclusió 1' AND ...conclusió k'

on les clàusules es poden acollir amb operadors:

- AND (s'han de complir totes),
- OR (n'hi a prou que es compleixi una), i
- NOT (és la negació de la condició la que ha de ser certa perquè es compleixi la clàusula).

A més de conclusions sobre possibles hipòtesis, les parts conseqüents també poden contenir accions, que es durien a terme si la part antecedent es complís.

Exemple de regla

Vegem un exemple d'una regla en un sistema expert relacionat amb el diagnòstic clínic de l'estat fetal quan una pacient embarassada se sotmet a un test prenatal. La idea és que l'expert clínic analitza diversos paràmetres i, en funció del seu valor, decideix un estat de normalitat o anormalitat fetal, que aniria acompanyat de les actuacions posteriors corresponents:

REGLA 12

IF Línia-de-base-cardíaca fetal=normal
AND Variabilitat-freqüència-cardíaca-fetal=normal
AND acceleracions-freqüència-cardíaca-fetal>=3
AND desceleracions-freqüència-cardíaca-fetal=0
THEN estat-fetal=normal

REGLA 7

IF desviació estàndard-freqüència-cardíaca-fetal < 2
THEN Variabilitat- freqüència-cardíaca-fetal=absent

REGLA 8

IF desviació estàndard-freqüència-cardíaca-fetal > 2
AND desviació estàndard-freqüència-cardíaca-fetal < 5
THEN Variabilitat-freqüènciacardíaca-fetal=decrescuda

REGLA 9

IF desviació estàndard-freqüència-cardíaca-fetal >= 5
AND desviació estàndard-freqüència-cardíaca-fetal < 10
THEN Variabilitatfreqüència-cardíaca-fetal=normal

REGLA 3

IF batecs-minut-fetal >= 120
AND batecs-minut-fetal <=160
THEN Línia-de-base-cardíaca fetal=normal

D'aquesta manera es poden codificar àrees de coneixement expert d'alt nivell, les regles de les quals poden establir cadenes de raonament o inferències de dos modes diferents:

- mode d'encadenament progressiu (des de les dades d'un cas concret cap a les conclusions)
- mode d'encadenament regressiu (establint una hipòtesi inicial i comprovant si les dades del problema ens permeten establir aquesta hipòtesi com a correcta).

Exemple d'encadenament progressiu i encadenament regressiu

Relacionat amb les regles anteriors, en el cas d'un encadenament progressiu treballaríem encadenant les regles des dels valors dels paràmetres fins al diagnòstic. Així, per exemple, si tenim com a valors disponibles del cas que els batecs fetals per minut són 128, la desviació estàndard de la freqüència cardíaca és 7, no hi ha desceleracions i hi ha 4 acceleracions, les regles que s'activarien serien la 3, la 9 i la 12, i es conclouria amb aquesta cadena de raonament que l'estat fetal és normal.

Però també podem treballar en mode regressiu començant amb una hipòtesi de partida. Per exemple, podem suposar que l'estat fetal és normal (ho establim com a hipòtesi de partida) i després anem analitzant quins paràmetres serien necessaris per a establir aquesta hipòtesi com a certa. Quan aquests paràmetres s'esbrinen, el procés regressiu passa a un procés similar a l'anterior, dels paràmetres a les conclusions, per establir la veracitat de la hipòtesi de partida. En aquest cas la hipòtesi que plantegem està en la part de conclusió de la regla 12, i, per a poder establir aquesta conclusió com a certa, necessitaríem conèixer els valors dels paràmetres de la seva part IF o condició, que en aquest exemple són línia-de-base-cardíaca-fetal, variabilitat-freqüència-cardíacafetal, acceleracions-freqüència-cardíaca-fetal i desceleracions-freqüènciacardíaca-fetal. Com que no sabem els valors d'aquests paràmetres, hem d'esbrinar-los per a poder establir la hipòtesi com a certa.

Comencem amb el primer paràmetre. Hauríem de comprovar que «línia-de-base-cardíaca-fetal=normal», paràmetre desconegut ara com ara, forma part de la conclusió de la regla 3. Establim aquesta situació com a nova subhipòtesi i, per a comprovar-la, hem de saber de nou si es compleix la part IF de la mateixa regla. En aquesta regla 3 caldria conèixer el valor de batecs-minut-fetal. Suposem que podem conèixer-lo i que és 130. Llavors el procés evocatiu es converteix en progressiu i ens permet establir la conclusió de la regla 3. Però encara ens queden els altres paràmetres de la regla 12 per comprovar, amb els quals aniríem fent un procés anàleg configurant un procés evocatiu (regressiu) fins que una de les regles ens donés un valor per al paràmetre hipotetitzat i ens permetés canviar el sentit de l'encadenament cap endavant, ja que tots els valors dels paràmetres involucrats són coneguts, i en cas que s'ajustessin als valors esperats podríem establir la hipòtesi inicial com a certa (conclusió del raonament). Si algun dels paràmetres no té el valor especificat en la regla, la hipòtesi no es pot establir com a conclusió. Per exemple, si seguim amb el cas anterior i comprovem la veracitat del segon paràmetre, «variabilitat-freqüència-cardíaca-fetal=normal», veurem que és la conclusió de la regla 9, i, per a poder establir aquesta subhipòtesi com a veritable, necessitaríem conèixer el valor dels paràmetres de la part IF d'aquesta regla, en aquest cas «desviació estàndard-freqüència-cardíaca-fetal». Si aquest valor és 4, per exemple, la regla 9 no es compleix, sinó que es compleix la 8, i en aquest exemple el que passa, per tant, és que la regla 2 no es compleix, i la hipòtesi de partida no es pot establir com a certa.

Durant els anys vuitanta els sistemes experts van significar un èxit enorme de l'IA, que per primera vegada podia treballar en camps d'aplicació reals. Aquests sistemes es van aplicar en molts dominis, i cal esmentar els exemples següents a causa de la seva importància històrica:

1) DENDRAL (Buchanan, Feigenbaum, Lederberg, Stanford University, 1969) va ser el primer sistema expert que va tenir èxit. La meta del projecte era desenvolupar un programa capaç d'obtenir el mateix nivell de rendiment d'un químic expert en la determinació de les estructures moleculars basant-se en dades procedents d'un espectrògraf de masses. El projecte estava finançat per la NASA per usar-lo en una nau espacial enviada a Mart i determinar l'estructura molecular del sòl. La primera versió del programa generava totes les estructures possibles que es podrien correspondre amb la molècula, predeia per a cadascuna les observacions de l'espectre de masses i les comparava amb l'espectre real. El problema fonamental era que es podrien generar milions d'estructures possibles. Llavors van observar que els químics molt experimentats podien reduir el nombre de candidats possibles a una quantitat tractable mitjançant l'ús d'heurístiques adquirides amb l'experiència; per això es va decidir consultar químics analítics i tractar de simular el comportament d'aquests experts en el sistema. D'aquesta manera, el nombre d'estructures candidates es reduïa considerablement, i DENDRAL es va convertir en el primer sistema intensiu en coneixement que aconseguia funcionar en un entorn real complex. Aquest èxit va marcar un canvi dràstic en la recerca en IA, que va passar de centrar-se en els mètodes de propòsit general, amb escàs coneixement del camp d'aplicació i amb mètodes de cerca febles (és a dir, mètodes que no usen cap coneixement específic del problema concret per a trobar solucions), a fer-ho en tècniques específiques del domini i intensives en coneixement. A més, després d'aquesta experiència es va iniciar el Projecte de Programació Heurística a la Universitat de Stanford, encaminat a analitzar com aquesta nova metodologia es podia aplicar en altres àrees. A més dels primers sistemes experts, sorgí també la disciplina coneguda com a enginyeria del coneixement (IC), que engloba la captura, anàlisi i implementació del coneixement expert en un programa de computador.

2) El següent projecte important des del punt de vista històric, que va significar un èxit en el camp del diagnòstic mèdic, va ser **MYCIN** (Feigenbaum, Shortliffe, Stanford University, 1972). MYCIN era un sistema basat en regles de producció per al diagnòstic i la teràpia de malalties sanguínies infeccioses. Va aconseguir nivells de comportament tan bons com els dels metges experts. La característica especial que diferenciava MYCIN de DENDRAL era que el seu coneixement en forma de regles estava separat clarament del seu mecanisme de raonament, per la qual cosa el sistema era extensible i actualitzable fàcilment mitjançant la inserció o l'esborrament de regles. Les regles de MYCIN s'extreien de l'experiència mèdica, mentre que les regles de DENDRAL es derivaven de models teòrics. A més, les regles de MYCIN incorporaven incertesa mitjançant un esquema nou anomenat factors de certesa. El raonament amb incertesa, una manera natural de raonar de qualsevol expert humà, era una de les parts més importants del sistema.

3) Un altre sistema que va generar molta publicitat va ser **PROSPECTOR** (Duda, Stanford University, 1979), un sistema intel·ligent probabilístic per a l'explotació mineral. Per a representar el coneixement, el sistema usava un es-

Lectures recomanades

Staff of the heuristic programming project (1980). «The Stanford heuristic programming project: goals and activities». *Artificial Intelligence Magazine* (núm. 1, pàg. 25-30).

G. Schreiber; H. Akkermans; A. Anjewierden *et al.* (2000). *Knowledge Engineering and Management: the CommonKADS Methodology*. MIT Press.

A. Alonso Betanzos; B. Guijarro Berdiñas; A. Lozano Tello *et al.* (2004). *Ingeniería del conocimiento: aspectos metodológicos*. Pearson.

quema híbrid que incorporava regles i una altra forma de representació del coneixement anomenada xarxes semàntiques. PROSPECTOR incorporava el tractament de la incertesa usant un esquema de tipus probabilístic, l'esquema bayesià. El sistema era capaç de treballar al nivell d'un geòleg expert, i el fet que descobrís un dipòsit de molibdè (Mb) per un valor de més de 100 milions de dòlars a l'estat de Washington va ser la millor justificació per a utilitzar-lo.

4) XCON. Desenvolupat conjuntament per Digital Equipment Corporation i la Universitat de Carnegie-Mellon, ajudava a configurar noves comandes de computadors VAX. Va ser utilitzat internament per la companyia Digital.

Lectura recomanada

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

5) LES. Va ser desenvolupat per MITRE Corporation i NASA-KSC (NASA Kennedy Space Centre). La seva missió era monitorar i diagnosticar els processos de càrrega d'oxigen líquid al tanc principal del transbordador espacial.

Lectura recomanada

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, pàg. 360-368).

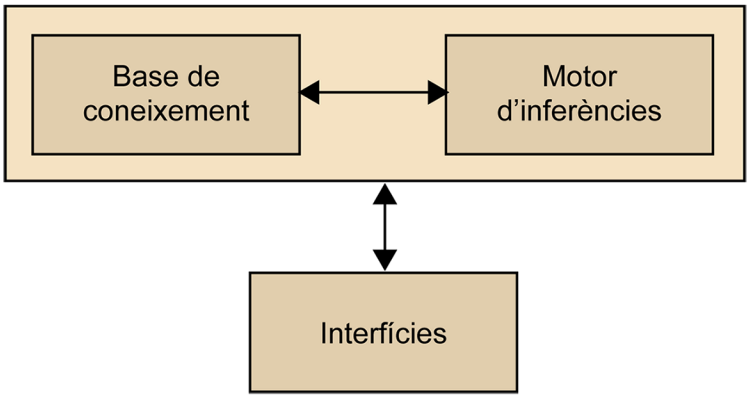
6) INTERNIST. Va ser desenvolupat a la Universitat de Pittsburgh. En el seu temps va ser el sistema mèdic amb un major nombre de regles i tenia la missió d'assistir el metge en l'elaboració de diagnòstics múltiples i complexos relacionats amb la medicina interna.

Lectura recomanada

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* (núm. 307, pàg. 468-476).

Aquests sistemes proposaven organitzar una aplicació en dos components: un component declaratiu, denominat base de coneixement, que representa el que coneix un agent especialitzat en un determinat tipus de problemes, i un procediment dissenyat per a simular un procés de raonament (vegeu la figura 2).

Figura 2. Arquitectura general d'un sistema expert.



Font: elaboració pròpia

La base de coneixement conté els fets i les regles del domini que es modelen. El motor d'inferències conté els models de raonament utilitzats, entre els quals hi ha els mecanismes d'encadenament progressiu i regressiu esmentats. Les interfícies són necessàries per a interaccionar amb els usuaris, altres equips maquinari o programes software, etc.).

Lectura recomanada

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, pàg. 928-938).

També hi ha alguns sistemes experts «clàssics» en el món del dret:

1) Taxman II, que conté coneixement jurídic de l'àmbit del dret fiscal als Estats Units, en concret la tributació de reorganitzacions corporatives. El sistema té un paper d'assessor, de manera que accepta descriure un cas que introdueix un usuari en forma de fets declarats seguint un procediment estàndard, i produeix una anàlisi d'aquest cas a un nivell d'abstracció adequat a la conclusió legal desitjada.

Lectura recomanada

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

2) Sistema LDS de Rank Corporation. En aquest cas els desenvolupadors descriuen el raonament involucrat en la liquidació de reclamacions de responsabilitat civil als Estats Units. Per a poder construir el sistema, els autors van estudiar com els advocats i ajustadores avaluaven les demandes civils en l'àrea de responsabilitat del producte i van elaborar un esquema que organitza els fets i detalls d'un cas. Tant l'esquema com un ampli conjunt de regles necessàries per a raonar els casos formen la base d'un sistema expert que modela la presa de decisions legals. Aquest sistema pot ajudar els investigadors i els litigants a comprendre millor com es duu a terme l'avaluació de les reclamacions, ja que proporciona una base per a generar i organitzar hipòtesis sobre els mètodes dels litigants per a establir la liquidació. Els resultats obtinguts van ser molt positius, ja que van demostrar que els sistemes experts basats en regles poden captar una gran part de la riquesa i flexibilitat del raonament legal en aquest camp.

Lectura recomanada

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

Exemple de regles en aquest sistema

En un llenguatge seminatural per a comprendre'l millor.

REGLA 4

IF Demandant és ferit en un ull
AND ulls afectats=1
AND tractament de l'ull va requerir cirurgia
AND recuperació de la ferida gairebé completa
AND agudesa visual es va reduir lleugerament
AND la condició de la ferida és estable
THEN augmentar factor de trauma de ferida en 10.000 $

REGLA 6

IF Demandant té probabilitat de contreure malaltia seriosa
AND valor probabilitat >= 5%
AND valor probabilitat <= 15%
THEN augmentar factor de trauma futur del valor del paràmetre contreure-malaltia en 30%

REGLA 8

IF Demandant no portava ulleres abans de la ferida
AND la ferida rebuda requereix portar ulleres
AND edat del demandant al moment de la ferida > 25
AND l'aparença física és important per al seu treball
THEN augmentar factor de desfiguració en 5.000 $

3) Sistema legal Research System, que té la funció d'ajudar els operadors jurídics a recuperar informacions relatives a les decisions judicials i a la legislació en el camp del dret dels títols de comerç.

Lectura recomanada

C. D. Hafner (1987). «Conceptual organization of case law knowledge bases». *Proceedings of the 1st International Conference on Artificial Intelligence and Law- ICAIL* (pàg. 35-42). ACM.

Hi ha alguns altres sistemes experts coneguts en el camp del dret, i diversos articles contenen reflexions sobre el seu ús, les seves limitacions i el seu futur. El creixement de les aplicacions per a problemes del món real i la proliferació dels sistemes experts van provocar l'augment de la demanda de nous esquemes de representació de coneixement i raonament, que intentaven millorar el rendiment tant dels sistemes com del seu procés de construcció, car i complex. Així i tot, en l'actualitat els sistemes basats en aprenentatge automàtic, és a dir, els que aprenen directament de les dades, triomfen en la majoria de les aplicacions reals, i els sistemes simbòlics són restringits a parts petites d'un domini determinat.

Lectures recomanades

K. D. Ashley (2000). «Designing electronic casebooks that talk back: the cato program». *Jurimetrics Journal* (núm. 40, pàg. 275-319).

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

E. Cáceres (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* (pàg. 78-87).

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

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

2.2. IA subsimbòlica

És anomenada també intel·ligència artificial computacional o inductiva i engloba els mètodes d'aprenentatge automàtic i la computació evolutiva. La idea que hi ha al darrere d'aquesta aproximació és que una de les característiques més distintives de la intel·ligència humana és la capacitat d'aprenentatge.

Els algorismes utilitzen, per a aprendre, processos d'inducció que a partir d'exemples particulars són capaços de generalitzar comportaments o reconèixer patrons, encara que també hi ha mètodes d'aprenentatge automàtic basats en reforç (inspirats en la psicologia conductista i que es basen a determinar quines accions ha d'escollir un agent software en un entorn donat amb la finalitat de maximitzar alguna noció de «recompensa» o premi acumulat) o en la imitació de processos evolutius biològics, usant algorismes evolutius, que són optimizadors matemàtics que funcionen generant moltes variacions singulars d'un individu, de manera que s'obté una població que pateix creuaments, mutacions, etc. i que evoluciona finalment mantenint els millors individus de la població, que selecciona utilitzant una funció d'ajustament o *fitness*.

En general, l'**àrea de l'aprenentatge automàtic** tracta de donar una resposta a la necessitat de construir sistemes computeritzats adaptables a l'entorn des d'una perspectiva diferent a la que vam veure en el cas dels sistemes experts, ja que en aquest cas, o bé l'experiència humana de partida necessària no existeix o bé no és fàcil d'extreure. Per aquest motiu, no és possible desenvolupar sistemes basats en regles que recullin el coneixement dels experts humans, com els que s'han vist en l'apartat anterior. En aquests casos les tècniques subsimbòliques ens permeten programar sistemes intel·ligents utilitzant dades del procés que transcorre o bé dades procedents d'experiències passades. Així, es podran identificar certs patrons o regularitats en les dades que s'utilitzaran per a construir bones aproximacions al problema.

Exemple: una fotografia

Considerem la qüestió de reconèixer algú amb una fotografia. Aquesta és una tasca que els éssers humans fem sense cap dificultat, encara que la foto estigui fosca o el posat, entre altres factors, dificulti el reconeixement. No obstant això, ens resulta tremendament difícil d'explicar com ho fem, per la qual cosa afrontar la tasca des d'una aproximació simbòlica seria poc menys que impossible. No obstant això, també sabem que un rostre no és quelcom aleatori, sinó que té una estructura amb certes característiques que seran comunes per a una persona en les diferents imatges que en tinguem. En aquesta aproximació subsimbòlica la idea consistiria a recollir un nombre important de mostres del cas concret (en aquest exemple la mostra seria de fotografies de diferents persones) i aprendre les similituds o patrons específics de cadascuna per comparar més tard les noves

fotografies que el sistema intel·ligent encara no ha vist amb els patrons descoberts en les anteriors i poder fer la identificació de les diferents persones en les noves fotos.

Una altra situació en què aquesta aproximació és interessant és aquella en què el problema a resoldre varia en el temps (per exemple, la detecció d'intrusos en xarxes d'ordinadors) o depèn de l'entorn particular (context específic) en què es treballi. Per a ser considerat intel·ligent, un sistema haurà de tenir l'**habilitat d'aprendre**. Així, podrem tenir sistemes amb un propòsit més general capaços d'adaptar-se a aquestes circumstàncies, en lloc d'haver d'escriure cada vegada programes explícits per a cada situació.

Exemple: trànsit de paquets

Un exemple d'aquest tipus podria ser un sistema que redirigeixi el trànsit de paquets en una xarxa de manera que es maximitzi la rapidesa del servei. El camí que maximitza la qualitat del servei entre un origen i una destinació varia contínuament, ja que depèn del trànsit de la xarxa. Un programa d'aprenentatge podria monitorar i adaptar-se a l'entorn canviant del trànsit, i subministrar així el millor camí en qualsevol circumstància.

Altres exemples podrien ser les interfícies intel·ligents que poden adaptar-se al perfil de l'usuari en funció de certs comportaments, com els seus hàbits de treball, o el contingut de pàgines web, que podrien adaptar-se en funció dels perfils d'interessos de l'usuari. En l'actualitat aquesta és la branca amb més èxit de l'IA perquè les dades són disponibles en pràcticament qualsevol camp a causa del procés de digitalització que es duu a terme, que fa possible representar digitalment la música, la cultura, els viatges o el cos humà, entre altres coses. Aquest procés de convertir gairebé tot en dades és possible gràcies als **sensors** que registren els esdeveniments i activitats que porten el món físic al digital, com per exemple un seqüenciador de genoma. Fa deu anys seqüenciar el genoma d'un individu costava prop de 200 milions d'euros. A dia d'avui costa menys de 500 euros, i aquest preu de digitalitzar el genoma tendirà a disminuir encara més. Com que les entitats digitals poden ser fàcilment replicades, emmagatzemades, transmeses, modificades o venudes, aquest traspàs fa que sectors sencers –en aquest cas la salut– es transformin en serveis d'informació i coneixement. El mateix passa en el món financer. Des d'una transacció fins a la majoria dels elements que conformen la relació amb el client (o del client amb el món), són informació que viu en el món digital.

L'era del *big data*

Aquesta proliferació dels diferents sensors, i també la possibilitat d'emmagatzemar dades no estructurades (com a imatges escanejades, documents, fotos, etc.), fan que en els últims anys hàgim entrat clarament en el que es coneix com a era del *big data*, en la qual el volum de dades ha crescut de manera exponencial a una velocitat molt alta, que fa que la quantitat de dades disponible es dupliqui cada any, i amb una varietat important de tipus (dades estructurades, semiestructurades i no estructurades) que han empetitit els sistemes d'emmagatzematge i processament tradicionals transformant les organitzacions i demandant maneres de processar informació innovadores i eficients en cost per obtenir informació i coneixement nous que permetin afegir valor de negoci a les dades disponibles.

En aquest context, els mètodes automàtics d'anàlisis de dades són imprescindibles. D'altra banda, també experimentem molts avenços en software; tal com hem dit, han aparegut aproximacions disruptives, com és el cas dels algorismes d'aprenentatge profund, que han fet possible resultats molt precisos.

Lectura recomanada

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

2.3. Altres classificacions

Podem fer altres divisions de l'IA, per exemple, en funció del tipus de problema que volem resoldre. En aquest cas estaríem parlant de problemes relacionats amb els sistemes següents:

1) Percepció, que tenen la capacitat de comprendre informació no estructurada en forma d'imatges o vídeo (àrea de visió per computador), veu (reconeixement del parla o generació artificial de veu) o textos (processament del llenguatge natural).

Exemples de sistemes d'aquest tipus

Són els de reconeixement facial, els etiquetadors automàtics de fotografies o vídeos (amb reconeixement de tipus de plantes, persones concretes, etc.), els bots (*chatbots*, programes intel·ligents amb els quals és possible mantenir una conversa i que es poden utilitzar en molts entorns, per exemple, per a comunicar accidents a la nostra asseguradora 24 hores al dia, 7 dies a la setmana) o els assistents personals (com Siri, Cortana o Alexa).

2) Planificació i cerca, que s'ocupen de trobar la millor solució entre un gran conjunt d'alternatives possibles, en entorns que poden ser o no ser totalment observables, deterministes, finits, estàtics o dinàmics, discrets o continus, etc.

Exemple de sistemes d'aquest tipus

Són els planificadors de rutes de vehicles, que han de treballar en condicions reals i en temps real, o els planificadors d'horaris i torns en una empresa.

3) Representació del coneixement i raonament automàtic, que tracten sobre la capacitat d'emmagatzemar, expressar i manipular el coneixement adquirit sobre un domini i poden fer ús del coneixement existent per a extreure conclusions.

Exemple de sistemes d'aquest tipus

Són els reconeixedors d'activitats, que s'utilitzen sobretot en el camp de la domòtica, l'estalvi energètic o l'atenció remota a persones dependents.

4) Aprenentatge, que enfoquen la capacitat de generar nou coneixement a partir de noves observacions, etc.

Exemple de sistemes d'aquest tipus

Són els sistemes que aprenen a reconèixer el correu no desitjat, els algorismes de màrqueting personalitzat, la medicina de precisió o l'educació individualitzada i utilitzen aproximacions basades en aquest tipus de capacitat.

Finalment, és important remarcar el fet que en la majoria de les situacions reals els sistemes intel·ligents necessiten combinar, a causa de la seva complexitat, diverses de les capacitats que hem vist anteriorment. Exemples típics són, entre d'altres, els sistemes multiagent o la robòtica:

1) Sistema multiagent: consta de diversos agents intel·ligents que interactuen entre ells, i la idea és que s'utilitza en els problemes que són difícils o impossibles de resoldre per a un agent intel·ligent únic. Així, es dissenyen diversos agents que s'ocupen d'una part del problema; aquests agents necessiten establir mecanismes de cooperació, negociació i comunicació entre ells per a poder organitzar la resolució del problema global. Els sistemes multiagent també es poden veure com una manera de distribuir IA per aconseguir sistemes més escalables.

En altres problemes també es poden utilitzar agents intel·ligents per a modelar situacions complexes en què el que es pretén obtenir és un comportament emergent que no estava planejat necessàriament des d'un principi o definit en els agents mateixos i que sorgeix per la interacció entre ells. Aquest procés d'emergència sorgeix a macroescala a causa dels comportaments dels agents individuals a microescala. Constitueixen models computacionals que s'usen de manera extensiva en biologia (per exemple, per a analitzar la difusió d'epidèmies o l'evolució de poblacions), en anàlisi de xarxes socials, en altres àrees de les ciències socials i l'economia, etc.

2) Robòtica, que tracta dels agents físics que fan tasques mitjançant la manipulació física del seu entorn. Per a poder fer això, els robots tenen sensors de diversos tipus, que els permeten percebre l'entorn (com càmeres, acceleròmetres, aparells d'ultrasons, etc.), i també efectors, que els permeten actuar sobre aquest entorn (com articulacions, rodes, etc.). Els tipus de robots més coneguts són els següents:

- els manipuladors o braços robòtics (per exemple, els de l'estació espacial internacional o els de les línies d'assemblatge),

- els robots mòbils (vehicles terrestres, aeris o submarins no tripulats, o robots que mouen contenidors en ports com el d'Hamburg, per exemple) i

- els robots zoomòrfics, els sistemes de locomoció dels quals imiten els diversos éssers vius i que inclouen els robots humanoides, que, a més, se solen dissenyar per tenir aptituds socials i que es poden utilitzar, per exemple, en hospitals per a acompanyar pacients.

3. Conceptes bàsics sobre algorismes

En aquest apartat, explicarem conceptes bàsics per a entendre com són les dades que usen habitualment els algorismes de la intel·ligència artificial, centrant-nos en el subcamp de l'aprenentatge automàtic (o *machine learning*, en anglès). La idea principal és la identificació de patrons o tendències que hi ha en les dades de manera automàtica. Per a aquesta tasca, hi ha dues grans famílies de mètodes: aprenentatge supervisat i aprenentatge no supervisat.

1) Aprenentatge supervisat: els algorismes treballen amb dades que estan etiquetades, la qual cosa significa que hi haurà una funció que trobi la sortida desitjada donades les dades d'entrada.

Exemple d'aprenentatge supervisat

Un detector de correu brossa (*spam*) treballa buscant una funció que relacioni les dades que tenim sobre un correu electrònic (remitent, tipus de destinatari, assumpte, etc.) i els assigni una etiqueta que serà «*spam*» o «no *spam*».

L'aprenentatge supervisat se sol usar en problemes de classificació.

2) Aprenentatge no supervisat: no es disposa de dades etiquetades per a entrenar l'algorisme, per la qual cosa els mètodes no supervisats es basen a descriure l'estructura de les dades per intentar trobar algun tipus d'organització que simplifiqui l'anàlisi.

Exemple d'aprenentatge no supervisat

L'agrupament (o *clustering*, en anglès) busca agrupaments basats en similituds.

A continuació ens centrarem en les tasques de classificació, ja que són probablement les més usades en aprenentatge automàtic.

3.1. Els conjunts de dades

Un factor clau per a analitzar les dades són òbviament les mateixes dades. En els últims anys vivim en una societat en què emmagatzemem i recol·lectem quantitats ingents de dades gairebé sobre qualsevol matèria que puguem imaginar, la qual cosa ha propiciat l'aparició de l'anomenat *big data*. Amb aquest oceà de dades en què estem immersos, ha aparegut un perfil professional molt demandat, el **científic de dades**, la missió del qual és extreure informació útil a partir d'enormes quantitats de dades «en cru». Però comencem pel principi. Com podem definir el que són les dades?

Habitualment, investigadors recol·lecten les dades en forma de conjunt de dades (o *dataset*, en anglès). Un conjunt de dades es pot definir com una col·lecció de dades individuals, sovint anomenades *mostres*, *instàncies* o *patrons*. Una mostra es pot veure com la informació sobre un exemple particular, per exemple, sobre un pacient en el domini mèdic. La informació sobre aquest exemple particular s'obté de les *característiques* o *atributs*. Una característica pot ser el sexe del pacient, la pressió sanguínia o el color dels ulls.

Una de les tasques més importants en el camp de l'anàlisi de dades és la classificació, que, com s'ha esmentat abans, consisteix a assignar a cada mostra una *classe* o categoria específica. Típicament, les mostres que pertanyen a una mateixa classe tenen característiques similars. En la taula 1 podem veure com a exemple el conjunt de dades de «jugar a tennis»[1]. En aquest petit exemple tenim quinze mostres, i cada mostra té quatre característiques diferents que ens donen informació que pot ser útil per a determinar si és possible jugar a tennis o no (tenint en compte que el tennis és un esport que es juga al l'exterior). L'última columna representa la variable predictiva o classe, que és la sortida desitjada d'aquest conjunt de dades en una tasca de classificació.

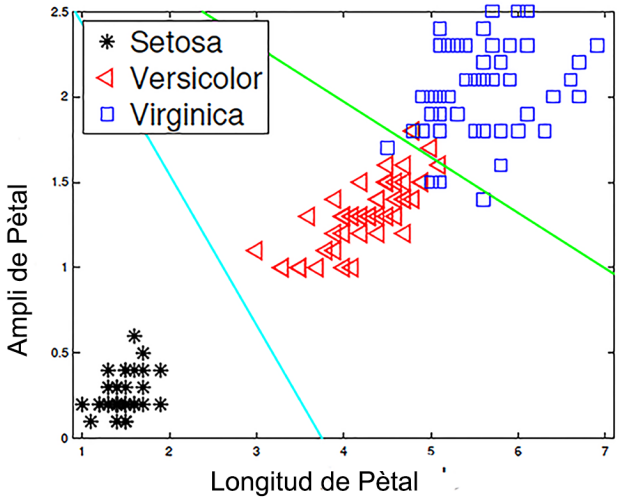
[(1)]Sobre el repositori de conjunt de dades: **D. Dua; Graff, C**. (2017). *UCI machine learning repository*.

Taula 1. Conjunt de dades de jugar a tennis

Pronòstic	Temperatura	Humitat	Ventós	Jugar?
assolellat	calorós	alta	fals	no
assolellat	calorós	alta	veritable	no
assolellat	calorós	alta	veritable	no
ennuvolat	calorós	alta	fals	sí
plujós	temperat	alta	fals	sí
plujós	fresc	normal	fals	sí
plujós	fresc	normal	veritable	no
ennuvolat	fresc	normal	veritable	sí
assolellat	temperat	alta	fals	no
assolellat	fresc	normal	fals	sí
plujós	temperat	normal	fals	sí
assolellat	temperat	normal	veritable	sí
ennuvolat	temperat	alta	veritable	sí
ennuvolat	calorós	normal	fals	sí
plujós	temperat	alta	veritable	no

Un dels conjunts de dades més populars en la literatura sobre anàlisi de dades és el conjunt Iris. Aquest conjunt de dades s'ha usat en milers de treballs de recerca al llarg dels anys i consisteix a distingir tres classes de planta iris (setosa, virgínica i versicolor). El conjunt té quatre característiques, que són l'amplada i longitud del pètal i sèpal, i cinquanta mostres de cadascuna de les tres classes o categories. Com es pot veure en la figura 3, una de les classes (setosa) és clarament separable de les altres dues amb una línia recta, mentre que en les classes virgínica i versicolor això no seria possible, ja que els exemples es barregen en la frontera d'ambdues classes.

Figura 3. Representació de les tres classes del conjunt de dades Iris amb la longitud de pètal en l'eix horitzontal i l'amplada de pètal en l'eix vertical



Font: elaboració pròpia

El fet de disposar de característiques que siguin separables linealment ens permet aconseguir una precisió de classificació perfecta, mentre que quan les classes no són separables és possible que alguns classificadors cometin alguns errors. Comentarem amb més detalls aquest assumpte en l'apartat següent.

3.2. Error/precisió de classificació

Tal com hem esmentat anteriorment, la tasca d'un classificador és assignar a quina classe pertany una mostra determinada. Per tant, necessitarem mesures per a avaluar quant bona ha estat la classificació.

Una mètrica d'avaluació usada àmpliament és l'**error de classificació**, que és el percentatge de mostres classificades incorrectament dividit pel nombre total de mostres. Anàlogament, la **precisió de classificació** és el percentatge de mostres classificades correctament dividit pel nombre total de mostres.

No obstant això, fixar-se solament en l'error o precisió de classificació no és una bona idea. Suposem que tenim un conjunt de dades format per cent imatges, noranta-cinc de les quals són gats i solament cinc són gossos. En primer

lloc usarem un classificador, que anomenarem C_1 i que decideix que totes les imatges són gats, de manera que tindrà una precisió de classificació del 95%, la qual cosa sona molt bé. A continuació usarem un altre classificador, que anomenarem C_2 i que falla a classificar quatre imatges de gats i dos de gossos, de manera que obté finalment una precisió del 94%. Quin classificador és millor? La resposta a aquesta pregunta depèn del tipus de dades i de l'objectiu de l'aprenentatge, però en general és millor aconseguir una solució de compromís entre la capacitat de classificació de les dues classes, per la qual cosa és necessari també comprovar individualment els percentatges d'encert en cadascuna de les classes.

A més, hi ha altres mesures per a avaluar la bondat dels classificadors que és molt important tenir en compte:

- **Veritable positiu** (VP): percentatge de mostres positives classificades com a positives.
- **Fals positiu** (FP): percentatge de mostres negatives classificades incorrectament com a positives. També és conegut com a error de tipus I.
- **Veritable negatiu** (VN): percentatge de mostres negatives classificades com a negatives.
- **Fals negatiu** (FN): percentatge de mostres positives classificades incorrectament com a negatives. També és conegut com a error de tipus II.

Per descomptat, l'objectiu de qualsevol sistema classificador és mantenir molt altes les taxes de VP i VN. No obstant això, cal tractar amb molta cura els FP i FN, prioritzant un o un altre depenent de l'escenari o naturalesa del problema. Vegem-ne dos exemples.

Exemple: detecció de malalties

Suposem que tenim un sistema d'intel·ligència artificial per a detectar un algorisme de classificació si un pacient té una determinada malaltia. En aquest context un FP significarà que direm al pacient que té la malaltia quan en realitat està sa, i un FN significarà que direm al pacient que està sa quan té la malaltia realment. Encara que és desitjable que cap d'aquestes situacions no passin, en aquest context és millor cometre un FP (ja que el més probable és que en proves posteriors es descobreixi que el pacient no té la malaltia realment) a cometre un FN i tenir un pacient malalt que no sigui tractat.

Exemple: amenaces a la xarxa

Ara suposem que tenim un sistema de intel·ligència artificial per a detectar amenaces en una xarxa de computadores. En aquest context un FP significarà que s'ha enviat una alarma que hi ha un atac quan era una connexió normal, i un FN significarà que hi havia una amenaça però no s'ha detectat. Per descomptat, és crucial no deixar passar atacs inadvertits, però una alta taxa d'FP pot tenir efectes desastrosos, ja que si rebem moltes alarmes falses deixarem de fer-hi cas, i quan es produeixi un atac real l'ignorarem i el sistema no servirà per a res.

3.3. Entrenament i prova

Al llarg d'aquest apartat hem parlat de la tasca de classificació i com un algorisme ha d'aprendre a diferenciar les classes existents. Però què passa quan arriba una nova mostra per ser classificada? Per exemple, què passa quan arriben les

dades d'un nou pacient i necessitem saber si està malalt? Aquesta és l'essència dels algorismes de classificació: ser capaços de **classificar mostres noves** per a les quals no sabem *a priori* la classe o categoria.

En una situació ideal s'usarien totes les mostres disponibles en el nostre conjunt de dades per a les quals sabem quina és la classe (per exemple, perquè les hem tret d'un històric de dades). D'aquesta manera, el nostre algorisme de classificació serà capaç d'aprendre les particularitats de les dades i la relació entre els valors de les característiques i la seva classe corresponent. Més tard, quan arribi una nova mostra per a la qual no sabem la classe, el nostre algorisme de classificació ja entrenat farà una predicció sobre la classe a què ha de pertànyer aquesta nova mostra. Però, d'aquesta manera, com sabrem si el nostre algorisme de classificació ha après correctament a partir de dades que representin correctament el problema? Per exemple, tornant a l'analogia amb el problema mèdic, què passaria si hem estat entrenant el nostre algorisme solament amb pacients dones però al món real els homes desenvolupen un altre tipus de símptomes per a la malaltia?

Per a assegurar-nos un procés d'aprenentatge correcte, hi ha dues pràctiques habituals:

- En primer lloc, assegurar-nos que el conjunt de dades que usarem per a aprendre representa correctament la població global.
- En segon lloc, guardar una part de les dades de les quals sabem la classe per usar en el que s'anomena comunament conjunt de prova o test. Aquest segon pas és molt important a l'hora de desenvolupar una metodologia d'aprenentatge correcta, ja que és necessari guardar una part de les dades que l'algorisme d'aprenentatge mai no ha vist mentre s'ha entrenat per aprendre.

Tots els paràmetres que estiguin involucrats en el procés d'aprenentatge s'han de calibrar sobre el conjunt d'entrenament, mai sobre el conjunt de test, i sobre els processos de preprocessament de les dades, ja que les dades que s'han guardat per a la fase de test no s'han d'utilitzar en cap de les altres fases d'aprenentatge.

Això inclou seleccionar les característiques rellevants per a un problema donat i descartar les irrellevants o redundants. És una pràctica relativament habitual (encara que incorrecta), fer la selecció de característiques sobre totes les dades disponibles i, una vegada eliminades les característiques que no són necessàries, fer la divisió de les dades en entrenament i test. Aquesta pràctica afectarà negativament el procés d'aprenentatge, ja que tot tipus de preprocessament de les dades s'ha de fer únicament sobre el conjunt d'entrenament deixant el conjunt de test per a avaluar el rendiment del model après.

Alguns dels conjunts de dades usades comunament en la literatura científica ja venen dividides originàriament en conjunt d'entrenament i prova.

Exemple: conjunt de dades dividides originàriament en entrenament i prova

Conjunt de dades KDD (Knowledge Discovery and Data Mining Tools Conference) Cup 99, conjunt de referència en el camp dels sistemes de detecció d'intrusos.

Hi ha conjunts de dades separades per a entrenament i prova amb la particularitat que el percentatge de les diferents classes (connexió normal i diferents tipus d'atac) varia significativament entre el conjunt d'entrenament i el de prova, i es dona el cas que fins i tot en el conjunt de prova apareixen nous atacs que no estaven presents en el conjunt d'entrenament.

Quan les dades no són dividides d'origen, són els investigadors que han de decidir com dividir-les. Hi ha diversos protocols habituals per a aquesta divisió, els més usats dels qual són els següents:

- **Validació creuada amb *k* paquets**. És un dels esquemes de validació més populars. El conjunt de dades es divideix en *k* paquets de la mateixa grandària. L'algorisme d'aprenentatge s'entrena amb *k* - 1 paquets, i posteriorment s'usa el paquet restant com a conjunt de test o prova i s'estima l'error. Aquest procés es repeteix tantes vegades com paquets *k* tinguem, i l'error de validació creuada es calcula a partir de la mitjana dels *k* errors obtinguts en cada iteració.

- **Validació creuada deixant-ne un**. És una variant de la validació creuada amb *k* paquets, de manera que *k* és el nombre de mostres del paquet. En cada iteració del procés es deixa una única mostra per a prova.

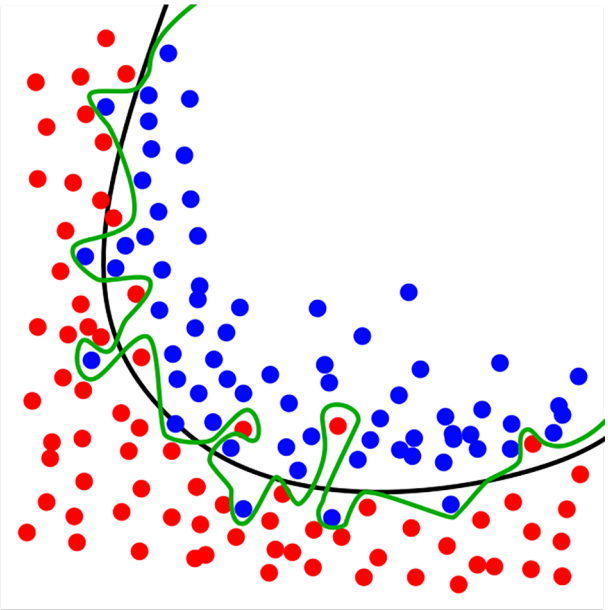
- ***Bootstrap***. És una tècnica general de remostreig. Una mostra *bootstrap* es compon d'*n* mostres amb la mateixa probabilitat de ser escollides, amb reemplaçament, entre el conjunt de dades original. D'aquesta manera, és possible que algunes de les mostres apareguin repetides diverses vegades, mentre que altres mostres no arriben a aparèixer mai. L'algorisme d'aprenentatge s'entrena amb la mostra de dades *bootstrap* i es prova posteriorment amb les mostres que s'han quedat fora. L'error s'aproxima calculant la mitjana de les mostres basat en rèpliques independents (normalment entre vint-i-cinc i cent).

- **Validació simple**. Consisteix a dividir de manera aleatòria les dades disponibles en dues particions: entrenament i test. Normalment es divideix deixant 2/3 de les dades per a entrenament i 1/3 per a test. L'algorisme d'aprenentatge s'entrena usant les dades de la partició d'entrenament i l'error s'estima calculant la proporció d'errors en les dades de prova. Aquest tipus de validació s'usa quan en un estudi hi ha conjunts de dades que ja

són dividides originàriament en entrenament i test però d'altres que no ho són.

L'elecció de la tècnica de validació apropiada no és trivial en absolut i depèn normalment de la grandària de dades que hàgim de tractar. Per exemple, si solament disposem de cent mostres (quelcom habitual amb alguns tipus de dades genètiques, per exemple), escollir validació simple deixant 2/3 per a entrenament i 1/3 per a test no seria una bona idea, ja que la grandària del conjunt d'entrenament pot ser massa petita i patir l'efecte de sobreajustament de les dades (és a dir, l'algorisme d'aprenentatge no podrà generalitzar; vegeu la figura 4). Per contra, si tenim un conjunt de dades realment gran (milions de mostres o característiques), usar validació creuada en qualsevol de les variants provocarà un temps de computació excessivament elevat, per la qual cosa es tornarà al mètode clàssic de la validació simple.

Figura 4. Exemple de sobreajustament de les dades



Font: *Diagram showing overfitting of a classifier*, amb llicència Creative Commons via Wikimedia Commons. Disponible a https://commons.wikimedia.org/wiki/file:Overfitting.svg

La línia verda representa un classificador que s'adapta perfectament a les dades (separant les dues classes, blava i vermella, correctament), però hi està massa adaptat, i davant l'aparició de noves dades obtindrà probablement més errors que el classificador que és representat per la línia negra.

3.4. Comparació de models: tests estadístics

Quan es presenta un nou algorisme d'aprenentatge és necessari comparar-ne els resultats amb els que obtenen els mètodes que conformen l'estat de l'art, per a demostrar que la proposta és trencadora i competitiva. Perquè aquesta

comparativa sigui justa, la pràctica habitual és usar tests estadístics. En el seu llibre, L. Kuncheva ens dona una sèrie de recomanacions per a comparar diversos models:

- Escollir acuradament la divisió entre entrenament i prova (vegeu l'apartat anterior) abans de començar els experiments. Si l'experiment es publica, és necessari assegurar-se que es donen els detalls necessaris perquè pugui ser reproduïble.

- Assegurar-se que tots els models usen tota la informació possible, i per descomptat que usen les mateixes dades per a entrenar i posteriorment en l'etapa de prova. Per exemple, no és una comparativa justa executar validacions creuades diferents per a diversos models, perquè la divisió aleatòria de les dades pot afavorir algun dels models. La manera correcta de fer els experiments és dividir les dades en els paquets necessaris i entrenar els diferents models amb les dades corresponents d'entrenament en cada iteració.

- Assegurar-se que les dades reservades per a l'etapa de prova no s'usin anteriorment en l'etapa d'entrenament.

- Quan sigui possible, fer tests estadístics. És millor saber si les diferències entre models són significativament diferents o no.

Lectura recomanada

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

4. IA fiable

Tal com hem dit, amb l'aparició del fenomen del *big data*, hem estat testimonis d'un increment aclaparador del volum de dades emmagatzemades, al mateix temps que aquestes dades s'han dotat de varietat, variabilitat, veracitat (és a dir, és segur que són de qualitat i no contenen soroll o anormalitats), etc. Als primers anys aquesta situació ha resultat en un empitjorament del funcionament d'algorismes d'IA existents, en particular d'algorismes d'aprenentatge automàtic. Com a conseqüència d'això, l'escalabilitat d'algorismes s'ha convertit en un requisit fonamental per a poder tenir algorismes eficaços, eficients i que puguin ser executats amb grans volums de dades.

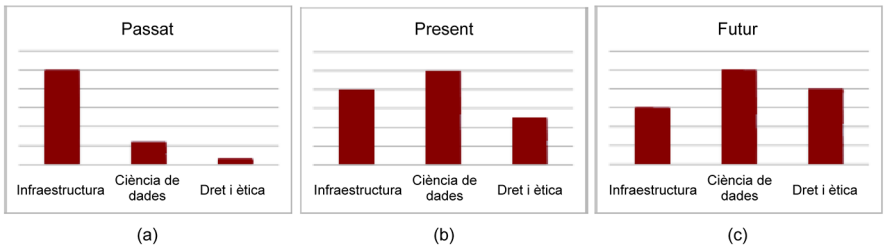
Si mirem aquest camp d'estudi des d'una altra perspectiva, podem veure que l'evolució de la indústria de dades ha estat molt ràpida. Durant els primers anys del fenomen del *big data*, s'ha centrat en el desenvolupament de les infraestructures necessàries per a processar aquestes enormes quantitats de dades. Actualment aquestes plataformes estan ben assentades en el mercat. No obstant això, hi ha una nova preocupació que ha irromput amb força en el camp de l'IA, que és la **necessitat d'una regulació ètica i legal** que asseguri la privadesa de les dades i el seu ús legítim, transparència en les decisions, etc., que no pot ser postposada.

El RGPD

A la Unió Europea el Reglament General de Protecció de Dades (RGPD) està canviant el panorama actual regulant l'ús de les dades i el dret a una explicació en cas de decisions automatitzades, amb l'objectiu d'aconseguir una IA transparent que eviti discriminacions sobre la base de la raça, l'estat de salut, el sexe, etc.

En el futur s'espera cada vegada menys demanda d'infraestructura, ja que el mercat està assentat d'alguna manera, però encara hi ha una necessitat de solucions que siguin més ràpides i potents, per la qual cosa hi ha marge per a millorar. L'interès en ciència de dades continuarà, ja que treballarem cada vegada en més àrees usant cada vegada més dades que necessitaran preprocessar-se abans de ser treballades. A més, també és necessari desenvolupar solucions que permetin explicar i visualitzar els resultats obtinguts, i manejar les dades dins dels algorismes respectant-ne la privadesa. Finalment, l'àrea relacionada amb les lleis, reglaments i ètica anirà guanyant en popularitat, i seran necessaris professionals d'altres camps com el dret, l'economia o la filosofia, que configuraran d'aquesta manera una àrea encara més interdisciplinària. L'IA comença un nou camí per a guanyar-se la confiança de la societat.

Figura 5. Passat, present i futur en les diferents àrees que cobren protagonisme en el desenvolupament de la intel·ligència artificial



Font: elaboració pròpia

L'abril de 2018 la Comissió Europea va publicar un comunicat en què s'anunciava una ambiciosa estratègia europea per a la intel·ligència artificial. Una part d'aquesta estratègia fou crear el **Grup d'Experts d'Alt Nivell en Intel·ligència Artificial** (AI HLEG), que inclou representants del món acadèmic, societat civil i empreses. El seu objectiu és donar suport a la implementació de l'estratègia europea en IA, incloent l'elaboració de recomanacions per a futures polítiques de desenvolupament relacionades amb IA, i assumptes ètics, legals i socials relacionats amb l'IA. L'abril de 2019 l'AI HLEG va publicar un document de directrius i ètica per una IA fiable. D'acord amb aquestes directrius, i com s'ha comentat en la introducció, l'IA fiable ha de ser:

- **Lícita**. Ha de respectar totes les lleis i reglaments aplicables.
- **Ètica**. Ha d'assegurar el compliment dels principis i valors ètics.
- **Robusta**. Tant des del punt de vista tècnic com social, ja que els sistemes d'IA poden provocar danys per accident fins i tot si les seves intencions són bones.

A més, les directrius estableixen set requisits clau que els sistemes d'IA han de complir per a poder ser considerats fiables:

1) Acció i supervisió humana:

- En primer lloc és necessari que els sistemes d'IA respectin els drets fonamentals de les persones. Encara que han de ser concebuts idealment per a facilitar la vida de les persones, per exemple, millorant l'accessibilitat a l'educació, també és possible que afectin negativament els drets fonamentals, i aquest risc ha de ser avaluat acuradament.

- En segon lloc és necessari proporcionar als usuaris dels sistemes d'IA els coneixements i eines necessaris per a comprendre aquests sistemes i la raó per què prenen determinades decisions. El principi general d'autonomia de l'usuari és fonamental en el disseny del sistema, de manera que ha de respectar el dret a no ser jutjat basant-se exclusivament en la decisió d'un procés automatitzat, tal com ho recull l'article 22 de l'RGPD.

- I en tercer lloc cal que hi hagi una supervisió humana dels sistemes d'IA tant en la fase de disseny com en la d'execució. D'aquesta manera, serà

possible fins i tot ignorar una decisió presa per un sistema si s'estima convenient així.

2) Solidesa tècnica i seguretat. La solidesa tècnica està vinculada estretament al principi de prevenció del dany i requereix que els sistemes d'IA es desenvolupin amb un enfocament preventiu en relació amb els riscos. Aquests riscos inclouen que el sistema d'IA pot ser atacat com qualsevol sistema de software, per la qual cosa és necessari que els sistemes d'IA siguin segurs i robustos. Finalment, per a evitar errors indesitjats, és fonamental que els sistemes d'IA siguin reproduïbles. Això significa que si repetim diverses vegades un experiment en les mateixes condicions el resultat hauria de ser sempre el mateix, i així seria possible descriure amb exactitud el que fa un sistema d'IA.

3) Gestió de la privadesa i les dades. La privadesa és un dret fonamental que és vulnerat sovint pels sistemes d'IA. Aquests sistemes no solament usen la informació inicialment facilitada per l'usuari, sinó que gestionen la informació obtinguda sobre l'usuari en el context de la seva interacció amb el sistema. Aquest és un tema que ha de ser tractat amb delicadesa, ja que els sistemes d'IA poden utilitzar no solament la informació original sobre els usuaris per al seu aprenentatge sinó una altra informació més susceptible de provocar discriminacions, com la seva orientació sexual, el gènere, etc. Un altre front obert molt important és la qualitat de les dades, ja que, perquè un sistema d'IA aprengui correctament, és necessari que les dades de les quals aprengui siguin de qualitat, de manera que s'evitin biaixos socials, imprecisions o fins i tot la inclusió de dades malintencionades.

4) Transparència. Aquest requisit està molt relacionat amb el principi d'explicabilitat i inclou tots els elements susceptibles de ser transparents per a un sistema d'IA: les dades, el sistema i els models de negoci. D'una banda, un sistema d'IA ha de ser **traçable**: això significa que totes les dades i decisions sobre el sistema han de ser documentades de manera rigorosa per a facilitar la traçabilitat futura i augmentar la transparència. Al seu torn, la traçabilitat augmenta l'auditabilitat i l'explicabilitat. Aquesta és una propietat altament desitjable en un sistema d'IA i ha d'estar relacionada amb la capacitat d'explicar tant els processos tècnics del sistema com les decisions humanes associades. És necessari que les decisions que pren un sistema d'IA siguin comprensibles per als humans. En alguns casos pot ser possible que, a fi d'augmentar l'explicabilitat del sistema, se'n sacrifiqui lleugerament la precisió. Finalment, quan es presenta un sistema d'IA a un usuari, cal deixar-li clar que interactua amb una màquina i no amb un altre humà, i informar-lo de les capacitats i limitacions del sistema.

5) Diversitat, no discriminació i equitat. Un sistema d'IA fiable ha de garantir la inclusió i la diversitat al llarg de tot el seu cicle de vida. És fonamental que les dades que usen els sistemes d'IA no presentin biaixos, ja que, altrament, el sistema d'IA prendrà aquests biaixos com una característica normal de les dades. Per a evitar els biaixos, cal posar èmfasi a identificar informació anòmala

en la fase de recopilació de la informació i advocar per un procés de supervisió que permeti analitzar de manera transparent el propòsit, les restriccions, els requisits i les decisions del sistema. A més, també s'ha de garantir l'accés als sistemes d'IA per part de qualsevol usuari sense discriminació per edat, gènere, capacitats o característiques.

6) Benestar social i ambiental. S'ha de fomentar la sostenibilitat i la responsabilitat ecològica dels sistemes d'IA, i impulsar la recerca de solucions d'IA per a resoldre problemes i reptes que afecten tota la població, com els objectius de desenvolupament sostenible. Quan es desenvolupa un sistema d'IA és recomanable que, a més de resoldre correctament el problema que se li ha encomanat, també ho faci de la manera més respectuosa possible amb el medi ambient. Per exemple, hi ha tècniques d'IA que necessiten fer grans còmputs i per això consumeixen molta energia, amb el conseqüent impacte per al medi ambient en forma d'emissions de CO_2, per la qual cosa és necessari trobar alternatives que tinguin un menor ús energètic. Finalment, a més d'estudiar l'impacte que els sistemes d'IA exerceixen sobre els seus usuaris, també és desitjable estudiar-ne l'impacte sobre les institucions, la democràcia i la societat en conjunt.

7) Rendició de comptes. A més dels requisits anteriors, perquè un sistema d'IA sigui fiable, cal establir mecanismes que permetin garantir la responsabilitat i la rendició de comptes sobre el sistema i els seus resultats tant abans de la implantació com després. Els sistemes d'IA han de ser auditables per auditors tant interns com externs, la qual cosa contribueix a la fiabilitat de la tecnologia. A més, cal garantir la capacitat d'informar sobre les accions o decisions que contribueixen al resultat del sistema i de respondre a les conseqüències d'aquest resultat, intentant minimitzar els seus efectes negatius. Finalment, si es produeixen efectes adversos injustos s'han de preveure mecanismes accessibles que assegurin una compensació adequada perquè els usuaris tinguin d'aquesta manera una major confiança en els sistemes d'IA.

5. Algunes àrees interessants que utilitzen intel·ligència artificial en camps del dret

Els canvis tecnològics que estan tenint lloc han ocasionat el naixement del nou camp del **dret digital**, ja que és necessari disposar d'una resposta jurídica precisa per a tota activitat relacionada amb els serveis de la societat de la informació i de la comunicació –que té la seva pròpia regulació, llenguatge i elements tecnològics. I, a més, s'ha de tenir en compte l'efecte transversal que les tecnologies, sobretot internet, tenen en les nostres vides i, per tant, en la resta de les disciplines del dret «tradicionals». Però l'altre gran repte és també el de l'IA. Tal com hem vist en la introducció, l'IA és una disciplina que introduirà –ja ho fa– grans canvis socials i econòmics durant els propers anys. El món del dret no serà aliè a aquesta revolució malgrat el seu cert caràcter tradicional.

Automatització de tasques legals

A Lexpo, un dels majors esdeveniments europeus sobre innovació legal que se celebra anualment en Amsterdam, un dels conferenciants va predir, ja el 2017, que als propers cinc anys entre el 20% i el 50% de les tasques legals rutinàries es veuran reemplaçades totalment per l'IA, que serà capaç d'automatitzar molts dels processos sense que hi hagi cap intervenció de les signatures d'advocats.

Vegem alguns camps legals en plena efervescència:

1) Assessorament legal a emprenedors i *startups* (i també a *business angels* i fons d'inversió), que ja ofereixen bufets d'advocats prestigiosos a Espanya que assessoren en qüestions com problemes d'inscripció o registre de marca, un mal repartiment del capital social o la desprotecció de les patents. També assessoren en un aspecte molt important a les empreses i *startups* digitals: la necessitat d'assegurar la gestió de dades. En aquesta àrea la dificultat és que es treballa amb clients tecnològics que desenvolupen o utilitzen tecnologia punta, i el problema rau gairebé sempre en el fet que la regulació va per darrere de la tecnologia i és necessari planificar certs aspectes preveient possibles futurs canvis i millores i aspectes ètics que puguin confluir en el negoci.

2) Les anomenades iniciatives Legaltech, que intenten donar solució a situacions relativament senzilles en què no es requereix una intervenció judicial però que per a resoldre-les es requereix l'actuació de la justícia, i que són contractes laborals, reclamacions a companyies de serveis (telefòniques, aerolínies, bancàries, etc.), divorcis, etc. La idea és que la companyia presti un servei legal als usuaris mitjançant una interfície utilitzable en dispositius mòbils, que sol adoptar la forma d'un *chatbot*. Un *chatbot* és un programa informàtic en línia que utilitza IA i ens permet mantenir una conversa per a sol·licitar o rebre informació o dur a terme accions. Aquest tipus d'empreses estan en expansió a Espanya des de fa un parell d'anys, encara que ja tenen bastant més recorregut en països com Estats Units o el Regne Unit, i el seu objectiu és solucionar problemes legals quotidians sense necessitat d'acudir a un bufet. La idea central és similar a la de la banca electrònica, que avui dia és una manera comuna d'interaccionar amb el nostre banc en línia.

Exemple de Legaltech

L'empresa iUrisfy, que és una *app* que assessora i tramita divorcis de mutu acord desenvolupant tota la negociació entre les dues parts perquè puguin comunicar-se entre elles i generar el document legal del divorci.

Aquest tipus d'interacció canvia les relacions entre advocats i clients i també abarateix els costos; per exemple, en el cas d'un divorci, es poden reduir pràcticament a la meitat. L'ús d'aquest tipus d'empreses està en auge al nostre país i se n'espera un creixement important als propers anys.

3) Automatització de tasques rutinàries. Hi ha molts procediments en l'àmbit del dret en els quals es fa un treball pràcticament manual i en els quals la qualificació de l'advocat no aporta realment un valor diferencial. Per tant, com en altres àmbits, aquestes tasques poden ser fetes per un sistema automatitzat encara que sigui necessària una supervisió lletrada final. Això s'enquadraria dins del que hem vist en les línies directrius per una IA fiable de la UE, en les quals l'humà és sempre l'eix central de la presa de decisions final. Pel que fa a aquest tema, les principals firmes d'advocats espanyoles ja han adquirit tecnologies d'IA per a automatitzar una part dels seus processos o ho estan analitzant, i ja hi ha empreses especialitzades en aquest camp, alguns exemples de les quals són RAVN Systems, Luminance i Neota Logic.

Exemples: ús de sistemes basats en IA per a operacions comunes

Les operacions *due diligence*, que consisteixen en el fet que per a fer l'assessorament en la possible compra d'una companyia, és necessari verificar prèviament tot tipus d'informació financera, contractual i empresarial amb l'objectiu de fixar-ne el valor real. Aquesta tasca pot ser llarga, fins i tot pot requerir mesos, però, com que els clients tenen un temps de decisió limitat, ocasiona que aquest treball manual s'hagi de fer de manera intensiva als bufets, amb els consegüents augments de cost, d'una banda, i un possible augment d'errors humans deguts a un sobreexcés de treball, de l'altra, que encara són pitjors en cas que no es produeixi finalment la compra de la companyia.

Un altre exemple pot ser l'ús de sistemes intel·ligents en la resolució de plets com els de les clàusules terra, respecte de les quals el dret del client ja és reconegut, que en l'actualitat estan acumulats a l'espera de ser resolts i que es podrien resoldre solament amb unes hores.

El software intel·ligent és molt més eficaç, ja que pot analitzar la documentació en qüestió d'hores i amb marges d'error pràcticament nuls, de manera que s'obtenen costos de servei més econòmics i, el que no és menys important, es permet que els experts en dret usin el seu temps en les tasques en què la seva experiència és realment útil i imprescindible.

4) Predicció de sentències mitjançant aprenentatge automàtic. El sistema judicial espanyol, com el d'altres països, necessita ser més eficaç, i una de les possibles solucions és l'ús de l'IA en la predicció de sentències, una possibilitat que estudia actualment el Consell General del Poder Judicial. El gran avantatge d'introduir aquesta tecnologia rau en l'agilitació de la velocitat burocràtica, ja que permet establir patrons d'eficàcia en el 85% dels casos aproximadament, sempre amb supervisió humana experta. Un dels països que figuren com a primers que han adoptat aquesta tecnologia és Estònia, que ja és capdavantera en la transformació completa del país mitjançant l'IA. Avui dia Estònia té una

base de dades d'1,3 milions de ciutadans en la qual ja aplica IA i aprenentatge automàtic. Per a Estònia, és important que tinguem en compte l'aproximació centrada en les persones de l'IA europea (recordem que a la UE els sistemes d'IA han de donar suport als humans i respectar els drets fonamentals, i no disminuir, limitar o desviar l'autonomia humana). Això requerirà mecanismes de supervisió adequats per a evitar els possibles biaixos algorítmics i de les dades i inclourà aproximacions de tres tipus:

- Capacitat d'intervenció humana en cada cicle de decisió del sistema.
- Capacitat d'intervenció humana durant el cicle de disseny del sistema i el monitoratge de les operacions del sistema.
- La persona al comandament, és a dir, capacitat de supervisar l'activitat general del sistema d'IA, inclosos els impactes més amplis i la capacitat de decidir quan i com usar el sistema en qualsevol situació particular.

També disposem del Reglament general de protecció de dades, que s'aplica en territori de la Unió Europea des del maig de 2018 i que regula, entre altres aspectes, el dret a l'explicació i la possibilitat de reclamació per part de les persones. Per tant, en realitat no és el procés per si mateix que canvia de manera substancial, però sí que ho fa la velocitat dels tràmits, de manera que permet una aplicació més eficaç i eficient de la justícia.

Exemple: el servei Jurimetría i altres

És una eina d'analítica jurisprudencial utilitzada per grans despatxos d'advocats al nostre país que usa IA i calcula si un litigi es guanyarà. Per a això, rastreja en uns quants minuts milions de sentències (maneja un banc de més de deu milions de sentències a Espanya) buscant patrons similars al cas que avalua, i calcula el resultat més probable de la sentència, el temps que el jutge tardarà previsiblement a resoldre el cas, i fins i tot la probabilitat de guanyar l'apel·lació en l'audiència o el percentatge d'èxit de l'advocat de la part contrària. Com a resultat, ofereix també gràfics que permeten al lletrat avaluar si signar un acord extrajudicial és una estratègia millor que entrar en sala.

Un grup d'investigadors del University College London, la Universitat de Sheffield i la Universitat de Pennsylvania han desenvolupat un algorisme capaç d'analitzar les dades de casos de la Cort Europea de Drets Humans (CEDH) que ha aconseguit predir el 79% de les resolucions de 584 assumptes. Procediments d'aquest tipus, que per raons merament formals s'estenen durant anys, es podrien resoldre en pocs mesos, la qual cosa agilitzaria considerablement el sistema.

Hem vist solament alguns dels aspectes en què l'IA pot contribuir a millorar l'eficàcia dels processos relacionats amb el dret en àmbits relacionats principalment amb el dret dels negocis en què hi ha infraccions econòmiques o competència deslleial, en jurisdiccions com la fiscal –en casos de comptabilitat–, la civil –en deutes, asseguradores– o l'administrativa –en multes de trànsit–, i en temes de marques i patents.

Recordem que l'aproximació de la UE està centrada en la persona i, per tant, no posarà fi a la figura del jutge ni a la del lletrat. Les solucions finals que recomani el sistema hauran de ser revisades pel lletrat; qualsevol error generat per aquesta via haurà de ser verificat i confirmat sempre per un jutge i, en cas de desacord, sempre caldrà que hi hagi una possibilitat de recurs.

La idea és disposar de sistemes que serveixin de suport a la presa de decisions, que aportin el conjunt d'arguments legals aplicables i proposin decisions que han de ser confirmades, però amb l'avantatge d'eliminar les tasques tedioses i repetitives i alleugerir els temps necessaris per a l'actuació de la justícia. Perquè aquestes qüestions plantejades es converteixin en realitat, cal també un canvi social; és imprescindible generar un salt qualitatiu tant tecnològic com cultural.

Resum

En aquest mòdul, que té com a objectiu servir d'introducció a la disciplina de l'IA per a estudiants tant de dret com d'altres que no tenen necessàriament coneixements tecnològics, s'ha discutit la relació entre el camp de l'IA i el del dret. Encara que *a priori* poden semblar dos camps que no estan molt relacionats, s'ha demostrat que sí que ho estan, i ho estan des del naixement dels sistemes experts als primers anys de la disciplina fins al moment actual, tal com hem vist. Però durant els últims anys aquesta relació s'ha fet encara més estreta i ha fet imprescindible conèixer algunes nocions de la disciplina per als estudiants esmentats, sobretot des de l'entrada en vigor de l'RGPD, que no solament regula i controla l'ús de les dades personals dels ciutadans europeus, sinó que estableix el seu dret a requerir explicacions quan s'han pres decisions sobre ells que involucren sistemes d'IA. En primer lloc s'ha descrit breument la història de l'IA des del seu naixement a la dècada de 1950 fins al seu auge en l'actualitat passant per diferents períodes en què ha tingut més o menys popularitat i que es coneixen com a primaveres i hiverns. A continuació s'han descrit les diferents branques que componen la disciplina de l'IA, que es poden dividir en dos subcamps: l'IA simbòlica i la subsimbòlica. Ambdues disciplines tenen els seus avantatges i les seves limitacions i són usades àmpliament, encara que avui dia es popularitza més l'ús de tècniques d'aprenentatge automàtic, que s'engloben dins de l'IA subsimbòlica. Prenent com a base l'aprenentatge automàtic, s'han explicat conceptes bàsics sobre els algorismes recalcant la importància de disposar d'unes dades adequades per a l'aprenentatge, i la conveniència d'usar tècniques i mesures d'avaluació correctes. Seguidament s'ha introduït el concepte d'IA fiable, que és el sistema d'IA en les decisions del qual podem confiar, ja que ha de complir propietats desitjables com ser robust, explicable, transparent, respectar la privadesa de les dades, ètic, auditable, i vetllar pel benestar social i ambiental. Per finalitzar, s'han comentat algunes àrees interessants que utilitzen IA en camps del dret, com la predicció de sentències usant tècniques d'aprenentatge automàtic o l'automatització de tasques rutinàries. Davant nostre apareixen nous reptes per al dret en moltes de les branques i àmbits clàssics, però segurament també en algunes àrees noves, ja que els sistemes d'IA són la tecnologia actual més transformadora i la que té més influència en la societat 4.0 que s'està configurant davant nostre a gran escala i a gran velocitat.

Bibliografia

Alonso Betanzos, A.; Guijarro Berdiñas, B.; Lozano Tello, A.*et al.* (2004). *Ingeniería del conocimiento: aspectos metodológicos*. Pearson.

Ashley, K. D. (2000). «Designing electronic casebooks that talk back: the cato program». *Jurimetrics Journal* (núm. 40, pàg. 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, pàg. 298-318).

Barrio Andrés, M. (2018). *Robótica, inteligencia artificial y derecho*. <http://www.realinstitutoelcano.org/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* (núm. 32, pàg. 928-938).

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* (pàg. 78-87).

Casanovas, P. (2010). «Inteligencia artificial y derecho: a vuelapluma». *Revista de Pensamiento Jurídico* (núm. 7, pàg. 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* (núm. 2, pàg. 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». *Proceedings of the 1st International Conference on Artificial Intelligence and Law - ICAIL* (pàg. 35-42). ACM.

Hernández Jiménez, M. (2019). «Inteligencia artificial y derecho penal». *Actualidad Jurídica Iberoamericana* (10 bis, pàg. 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* (vol. 521, pàg. 436-444).

McCarty, L. T. (1977). «Reflections on Taxman: an experiment in artificial intelligence and legal reasoning». *Harvard Law Review* (núm. 90, pàg. 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* (núm. 307, pàg. 468-476).

Newell, A.; Simon, H. (1963). *GPS, a program that simulates human thought*. Nova 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, pàg. 360-368).

Schreiber, G.; Akkermans, H.; Anjewierden, A. *et al.* (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* (núm. 1, pàg. 25-30).

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

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