Dimitrij Reja[1]

A REVIEW OF MACHINE LEARNING ALGORITHMS FOR ARCHIVAL APPLICATIONS

Abstract

***Purpose**: An overview of machine learning algorithms suitable for use in archival science.*

***Methode/approach**: A review of published articles and literature on machine learning algorithms suitable for the field of archival sciences.*

***Results**: In the multitude of algorithms in machine learning, the solution to choosing an algorithm is only an accurate and clear definition of the problem we want to solve. With the help of a precisely defined problem, the choice of the most suitable algorithm is above all easier.*

***Conclusions/findings**: In the process of processing archival material using metadata, the artificial intelligence or tools used in various procedures help us significantly.*

***Key words**: machine learning, algorithm, archival sciences*

UNA RASSEGNA DEGLI ALGORITMI DI MACHINE LEARNING PER LE APPLICAZIONI DI ARCHIVIAZIONE

Abstract

Scopo: una panoramica degli algoritmi di machine learning adatti all'uso nelle scienze archivistiche.

Metodo/approccio: una revisione degli articoli pubblicati e della letteratura sugli algoritmi di machine learning adatti al campo delle scienze archivistiche.Risultati: nel moltitudine di algoritmi nell'apprendimento automatico, la soluzione alla scelta di un algoritmo è solo una definizione accurata e chiara del problema che vogliamo risolvere. Con l'aiuto di un problema definito con precisione, la scelta dell'algoritmo più adatto è soprattutto più facile.

Conclusioni/risultati: nel processo di elaborazione del materiale d'archivio utilizzando i metadati, l'intelligenza artificiale o gli strumenti utilizzati in varie procedure ci aiutano in modo significativo.

Parole chiave: apprendimento automatico, algoritmo, scienze archivistiche

1 Dimitrij Reja, IT specialist and archivist, graduated in 2002 on Faculty od Social Sciences, Ljubljana. Employed as System Administrator at The Office of the State Prosecutor General of the Republic of Slovenia (2006). CIO at Ministry of Infrastructure Republic of Slovenia (2012). IT specialist and archivist in The Inspectorate of Infrastructure Republic of Slovenia. Student of Archival Science Doctoral Program at Alma Mater Europaea – ECM master's program.

PREGLED ALGORITMOV STROJNEGA UČENJA ZA UPORABO V ARHIVISTIKI

Izvleček

***Namen**: Splošni pregled algoritomov strojnega učenja za uporabo v arhivistiki.*

***Metoda/pristop**: Pregled objavljenih člankov in literature o algoritmih strojnega učenja za področje arhivskih znanosti.*

***Rezultati**: Splošni pregled in enostavna razlaga uporabnih algoritmov na področju arhivistike.*

***Sklepi/ugotovitve**: V množici algoritmov pri strojnem učenju je rešitev glede izbire algoritma le točna in jasna opredelitev problema, ki ga želimo rešiti. S pomočjo natančno definiranega problema je izbira najbolj primernega algoritma nad vse lažja.*

***Ključne besede**: strojno učenje, algoritem, arhivske znanosti*

1 UVOD

Digitalna doba[2], v kateri se nahajamo, ima med drugim definirane naslednje ključne dejavnike, IoT[3], BigData[4] in AI[5]. V poplavi informacij, ki nas vsakodnevno bombardirajo preko številnih komunikacijskih kanalov[6], si je potrebno natančno in skrbno razčistiti posamične pojme na področju strojnega učenja. V pričujočem članku želimo na splošno in laično predstaviti različne algoritme strojnega učenja. Članek je namenjen pregledu vseh zanimivih in najbolj primernih za delo v arhivski znanosti. Ob raziskovanju literature na tem področju nesporno trčimo še na mnogo drugih terminov, ki jih je potrebno opredeliti in razložiti.

Definicijo umetne inteligence je prvi opredelil Alan M. Turing leta 1950. S pomočjo Turingovega testa je Alan M. Turing definiral ali je računalnik sposoben razmišljanja kot človeško bitje (Encyclopedia Britannica 2023). Računalnik poseduje umetno inteligenco v primeru, ko lahko posnema človeške odzive v določenih pogojih (St.George & Gillis 2023).

Turingov test je način, s katerim računalnik opravi preizkus umetne inteligence. Test je pozitiven, če človeški spraševalec po nekaj pisno zastavljenih vprašanjih ne zmore z veliko gotovostjo ugotoviti, ali je avtor pisnih odgovorov človek ali računalnik (Russell et al., 2010, 4).

Za dosego pozitivnega odgovora ali vsaj vzpodbuditi dvom spraševalca je zagotovo potrebno veliko dela na različnih področjih umetne inteligence. V prvi vrsti je potrebno omeniti NLP[7] (obdelavo naravnega jezika), ki bi računalniku omogočala uspešno komunikacijo. Najrazličnejše zbrane podatke je potrebno skladiščiti v BigData skladiščih. Avtomatizacija sklepanja in strojno učenje sta naslednji dve logični enoti razvoja. Za dosego umetne inteligence v smislu popolnoma samostojnega robota potrebujemo računalniški vid in robotiko za manevriranje.

Umetna inteligenca v sebi združuje najrazličnejše pojme, ki so postali samostojna področja raziskovanja. V predmetnem primeru se bomo posebej posvetili področju strojnega učenja in avtomatičnega sklepanja.

2 STROJNO UČENJE

Strojno učenje je sestavni del umetne inteligence. Področje strojnega učenja se ukvarja z vprašanjem, kako sestaviti računalniške programe, ki se samodejno izboljšujejo z izkušnjami (Mitchell 1997). Bolj natančna opredelitev pojma strojnega učenja, »Za računalniški program pravimo, da se uči iz izkušnje E glede na določen razred nalog T in merilo uspešnosti P, če se njegova uspešnost pri nalogah v T, merjeno s P, izboljša z izkušnjo E« (Mitchell, 1997).

Strojno učenje je v osnovi razdeljeno na štiri tipe učenja, in sicer;

- nadzorovano učenje,
- nenadzorovano učenje,
- delno nadzorovano učenje in
- učenje s pomočjo vzpodbude.

2 Digitalna doba je čas v katerem se večina opravil lahko opravi s pomočjo računalnika. Posledično je na voljo ogromna količina informacij.
3 The Internet of Things
4 Big data je pojem, ki se nanašajo na podatke, ki so tako veliki, hitri ali zapleteni, da jih je težko ali nemogoče obdelati s tradicionalnimi metodami
5 Artificial Inteligence, umetna inteligenca
6 Komunikacijski kanal je sistem naprav in signalnih povezav, ki omogočajo prenos digitalne informacije od izvora do uporabnika
7 Natural Language Processing

Velikokrat se v literaturi prepletata termina algoritem in model. Za lažje razumevanje je treba navesti razliko med algoritmom in modelom pri strojnem učenju. Algoritem pri strojnem učenju je postopek, ki se izvaja na podatkih, s pomočjo katerih se ustvari model strojnega učenja. Model je skupek vseh podatkovnih struktur, s pomočjo katerih lahko napovemo bodoče rezultate, in so nastali po izvedbi algoritma strojnega učenja (Brownlee, 2020).

2.1 NADZOROVANO UČENJE

Pri nadzorovanem učenju se stroj uči pod nadzorom človeka. Stroju so podani znani vhodni in poznane možnosti izhodnih podatkov. Na podlagi znanih vhodnih in izhodnih podatkov se zgradi model. Stroj se nato o izhodnem podatku odloča na podlagi različnih algoritomov. Vse skupaj poteka pod budnim očesom nadzornika, v našem primeru človeka.

Stroj se uči na podlagi opazovanja nekaj primerov vhodno-izhodnih parov in se nauči funkcijo, ki jo nato preslika iz vhoda v izhod (Russell et al., 2010, 695).

2.2 NENADZOROVANO UČENJE

Pri nenadzorovanem učenju je razlika od nadzorovanega učenja v odsotnosti nadzornika in uporabi drugačnih algoritmov kot pri nadzorovanem učenju.

Pri nenadzorovanem učenju se stroj nauči vzorcev na vhodu, čeprav povratne informacije niso eksplicitno na voljo (Russell et al., 2010, 694).

2.3 DELNO NADZOROVANO UČENJE

Delno nadzorovano učenje je nekje med nadzorovanim in nenadzorovanim učenjem. Pravzaprav večina delno nadzorovanih učnih strategij temelji na podaljšanju bodisi nenadzorovanega bodisi nadzorovanega učenja. Poleg tega vključuje še nasprotno učno strategijo in s tem dosega boljše rezultate (Zhu & Goldberg, 2009, 9).

2.4 VZPODBUJEVALNO UČENJE

Vzpodbujevalno učenje deluje na principu nagrade in kazni. Vzpodbujevalno učenje je učenje kaj narediti, kako povezati situacijo z akcijo na takšen način, da maksimiziramo število nagrad (Sutton & Barto, 2018, 1). Stroju ni zaupano, katero akcijo naj opravi, ampak jo mora odkriti samostojno s pomočjo nagrajevanja. Akcija, ki dobi največ nagrad je pravilna odločitev.

3 ALGORITMI PRI RAZLIČNIH OBLIKAH STROJNEGA UČENJA

Znotraj vsakega tipa strojnega učenja so razvrščeni različni algoritmi. Algoritmi kot sestavni del strojnega učenja so končno zaporedje ukazov, s katerimi, če jim sledimo v določenem vrstnem redu, opravimo nalogo (Univerza v Ljubljani, b. d.). Na področju nadzorovanega strojnega učenja se algoritmi najpogosteje delijo glede na uporabo tehnike regresije ali klasifikacije.

LINEARNA REGRESIJA

Linearna regresija napoveduje številsko vrednost na podlagi predhodno opazovanih podatkov (Maulud & Abdulazeez, 2020). Linearna regresija se deli na tri različne podvrste linearne regresije, in sicer na enostavna, večkratno in polinomsko regresijo. Linearna regresija se uporablja za prikazovanje ekonomske rasti, rasti cene produktov, prodajo nepremičnin, itd.

KLASIFIKACIJA

Klasifikacija je kategorizacija v najpreprostejšem primeru na dve vrednosti, pozitivno in negativno. Vsekakor imamo lahko tudi več alternativnih ravni (Dietterich, 1997). Tipični primer uporabe odločitve je npr. Ali elektronsko sporočilo sodi med nezaželena sporočila? Z najrazličnejšimi filtri, pregledovalniki vsebine določamo vrednost elektronskemu sporočilu, ali sodi med nezaželeno pošto ali ne?

ODLOČITVENO DREVO

Algoritem odločitvenega drevesa je danes eden najbolj priljubljenih algoritmov pri strojnem učenju. Algoritem nadzorovanega učenja se najpogosteje uporablja za razvrščanje problemov. Dobro deluje pri razvrščanju tako kategoričnih kot zvezno odvisnih spremenljivk (linearnih). Algoritem razdeli podatke na dva ali več homogenih nizov na podlagi najpomembnejših atributov/neodvisnih spremenljivk (Tavasoli, 2016).

PODPORNI VEKTORSKI ALGORITEM

SVM (Support Vector Machine) je algoritem, ki analizira podatke glede na njihove značilnosti. SVM algoritem je binarna tehnika razvrstitve in bo vsak nov element razvrstil v enega od dveh razredov. Dva razvrščena razreda vhodnih podatkov, ki ju SVM algoritem nariše na dvodimenzionalno ravnino (hiperravnino) sta med seboj ločena z robom. Določitev mej (robov) med dvema kategorijama na hiperravnini podatkov mora biti maksimizirana do te mere, da sta ravnini med seboj jasno ločeni. Vsaka hiperravnina mora biti jasno ločena z navidezno premico. Sredinska premica med obema mejnima premicama je naša meja med dvema razredoma.

NAIVNI BAYESOV ALGORITEM

Naivni Bayesov algoritem uporablja Bayesov izrek[8] za klasificiranje podatkov. Klasifikatorji na osnovi Baysovega izreka so linearni in so znani po tem, da so zelo učinkoviti (Raschka, 2017). Najpogosteje se ga uporablja pri razvrščanju dokumentov, diagnosticiranju bolnikov, vremenskih napovedih in prepoznavanju obrazov. Algoritem na osnovi Bayesovega izreka podaja pogojno verjetnost dogodka A glede na dogodek B.

K-NAJBLIŽJIH SOSEDOV ALGORITEM

KNN[9] je ne parametrični klasifikacijski algoritem. Znan je po svoji enostavnosti in učinkovitosti. Algoritem shrani vse razpoložljive primere in nato vse nove primere klasificira na podlagi večine glasov k sosedov (Taunk et al., 2019). Sodi v področje nadzorovanega strojnega učenja. Najlažje razumemo njegovo delovanje na primeru iz resničnega življenja. V želji pridobivanja informacij o želeni osebi je najlažje pridobiti informacije o osebi s pomočjo njej najbližjim osebam.

3.1 NAJPOGOSTEJŠI ALGORITMI PRI NENADZOROVANEM UČENJU

Algoritmi pri nenadzorovanem strojnem učenju se ločijo na uporabo gručenja (ang. *clustering*) in asociacije. Algoritem na principu gručenja objektov deluje tako, da objekt z največ podobnostmi ostanejo skupaj in imajo manj ali skoraj nič podobnosti z drugimi predmeti druge skupine. Algoritem asociacije išče povezave med spremenljivkami v velikih podatkih. Poišče nabor elementov, ki se skupaj pojavljajo v naboru podatkov (Javatpoint 2021).

8 Bayesov izrek je ki se uporablja za izračun verjetnosti dogodka, ki temelji na njegovi povezavi z drugim dogodkom.

9 KNN - K-Nearest Neighbor

K-VODITELJEV ALGORITEM

K-voditeljev algoritem je tipičen algoritem združevanja v skupine, ki se največkrat uporablja pri rudarjenju podatkov. Poleg tega se zelo pogosto uporablja pri združevanju v skupine pri zelo velikih nizih podatkov (Na idr., 2010). Uporablja matematične mere za združevanje niza podatkov. Manjša kot je razdalja, bolj so si podobne posamične podatkovne točke. Združevanje v skupine je urejeno tako, da so deležniki posamične skupine homogeni do točke K in nehomogeni do ostalih deležnikov v skupinah (Bonthu, 2021). Z drugimi besedami, K-voditelj poišče opazovanja, ki imajo skupne pomembne značilnosti in jih razvrsti v gruče (Jeffares, 2019).

K-NAČIN ALGORITEM

K-način algoritem se uporablja za združevanje kategoriziranih spremenljivk. Uporablja razlike popolnega neujemanja med podatkovnimi točkami. Manjše kot so razlike, bolj so si podobne podatkovne točke (Sharma & Gaud 2015). Razlika med K-voditeljev in K-načini je v uporabi pomenov in načinov pri izračunavanju.

HIERARHIČNO ZDRUŽEVANJE

Hierarhično združevanje v gruče je nenadzorovan algoritem strojnega učenja, ki uporablja neoznačen niz podatkov. Algoritem generira hierarhijo gruč v obliki drevesne strukture, ki jo imenujemo dendogram[10]. S pomočjo dendograma se objekti najlažje dodelijo posamični gruči. Poznamo dva pristopa združevanja v gruče, in sicer aglomerativni in razdelilni. Aglomerativni pristop začne od spodaj navzgor, razdelilni pristop deluje obratno, torej od zgoraj navzdol. Aglomerativni pristop prične združevati vse podatkovne točke v eno gručo, dokler ne ostane samo ena gruča, in sicer najvišja (Murtagh & Contreras, 2012). Algoritem hierarhičnega združevanja je med drugim zelo uporaben na področju obdelave slik in tekstovne klasifikacije (Steinbach et al., 2000).

4 UPORABNA VREDNOST ALGORITMOV V ARHIVSKI ZNANOSTI

V članku smo predstavili nekaj najpogosteje uporabljenih algoritmov s področja nadzorovanega in nenadzorovanega strojnega učenja. Pri raziskovanju algoritmov so se pojavila vprašanja, ali obstaja le en algoritem, ki je primeren za delo v arhivski znanosti. Odgovor na vprašanje ali lahko enoznačno določimo algoritem za uporabo v arhivskih znanosti, je negativen. Arhivska znanost kot multidisciplinarna veda in kot zelo diverzificirana veda se dotika zelo različnih področij, posledično je razpršena tudi uporaba najrazličnejših algoritmov.

Z veliko gotovostjo lahko trdimo, da se v arhivskih znanostih uporabljajo najrazličnejši algoritmi za izdelavo modelov strojnega učenja. Prvi izmed možnih kriterijev izbire vrste algoritma je količina podatkov. Količina podatkov je prvi kriterij, na podlagi katerega bomo izbirali. Količina podatkov in tip algoritma nam narekujeta razpoložljiva sredstva, ki jih potrebujemo za obdelavo. Ob tem ne smemo pozabiti tudi na kakovost vhodnih podatkov. Glede na kakovost podatkov smo vezani na učinkovitost delovanja algoritma in modela. Pri obdelavi podatkov za veliko nacionalno debato v Franciji, ki je potekala v letu 2019, je bila izpostavljena problematika zajema podatkov. Izkazalo se je, da bi bilo prispevke državljanov bolj učinkovito prebrati in jih nato obdelati preko programske opreme za prepoznavanje glasu kot obdelati s OCR tehnologijo (Chabin, 2020, 243).

10 Dendogram je diagram v obliki drevesa

Za lažjo predstavitev se osredotočimo na besedilne dokumente. Danes je svetovni splet glavni vir besedilnih dokumentov in približno 80 % teh podatkov je v nestrukturirani obliki. Posledično je generiranje strukturiranih podatkov, ki predstavljajo dokumente, metapodatke z veliko natančnostjo in zanesljivostjo je bistvenega pomena (Khan et al., 2010).

Na omenjenem področju največ obetajo metode na podlagi algoritmov, kot so Bayesov algoritem, odločitveno drevo, K-voditelji in podporni vektor. Ob tem je treba omeniti, da se priznani raziskovalci na tem področju izogibajo BOW[11] pristopom. Ta pristop predstavlja dokument kot neurejeno vrečko besed, s čimer se izgubi vsaka slovnična relacija med besedami. Besede se nato ponderirajo glede na pogostost pojavljanja. Na podlagi uteži je nato ustvarjen vektor za vsak dokument. Novo nastali vektorji se nato primerjajo za iskanje podobnih dokumentov ali odlomkov(Yeh et al., 2003).

Uporaba različnih algoritmov, kot na primer hibridni algoritmi, je morda ena izmed boljših možnosti na področju avtomatičnega klasificiranja dokumentov. Na tem mestu je treba izpostaviti Podporni vektor, Navini Baysov in K-voditelji algoritme. Naivni Baysov algoritem je zelo uporaben pri delu z elektronsko pošto in določitvi vsiljenih sporočil.

5 LITERATURA

Bonthu, H. (13. 6. 2021). KModes Clustering Algorithm for Categorical data. *Analytics Vidhya*. Pridobljeno na https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data.

Brownlee, J. (29. 4. 2020). Difference Between Algorithm and Model in Machine Learning.*Machine Learning Mastery.* Pridobljeno na https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/.

Chabin, M. A. (2020). The potential for collaboration between AI and archival science in processing data from the French great national debate. *Records Management Journal*, *30*(2), 241–252. Pridobljeno na https://doi.org/10.1108/RMJ-08-2019-0042.

Dietterich, T. G. (1997). Machine-Learning Research. *AI Magazine 18*(4). Pridobljeno na https://doi.org/10.1609/aimag.v18i4.1324.

Encyclopedia Britannica. (2023). *Turing test*. Pridobljeno na https://www.britannica.com/technology/Turing-test.

Javatpoint. (2021). *Unsupervised Machine learning—Javatpoint*. Pridobljeno na https://www.javatpoint.com/unsupervised-machine-learning.

Jeffares, A. (19. 11. 2019). K-means: A Complete Introduction. *Medium.* Pridobljeno na https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c.

Khan, A., Baharudin, B., Lee, L. H. & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology 1*(1), 4–20.

Maulud, D. & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends 1*(4), 140–147.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Murtagh, F. & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery 2*(1), 86–97. Pridobljeno na https://doi.org/10.1002/widm.53.

11 BOW – Bag of Words

Na, S., Xumin, L. & Yong, G. (2010). Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. V *2010 Third International Symposium on Intelligent Information Technology and Security Informatics* (str. 63–67). Pridobljeno na https://doi.org/10.1109/IITSI.2010.74.

Raschka, S. (2017). Naive Bayes and Text Classification I - Introduction and Theory (arXiv:1410.5329). *arXiv*. Pridobljeno na http://arxiv.org/abs/1410.5329.

Russell, S. J., Norvig, P. & Davis, E. (2010). *Artificial intelligence: A modern approach* (3rd ed). Prentice Hall.

Sharma, N. & Gaud, N. (2015). K-modes clustering algorithm for categorical data. *International Journal of Computer Applications 127*(1), 46.

Steinbach, M., Karypis, G. & Kumar, V. (2000). *A Comparison of Document Clustering Techniques* [Report]. Pridobljeno na http://conservancy.umn.edu/handle/11299/215421.

Sutton, R. S. & Barto, A. G. (2018). *Reinforcement Learning, second edition: An Introduction*. MIT Press.

St.George, B. & Gillis, A. S. (2023). Turing Test. *Enterprise AI*. Pridobljeno na https://www.techtarget.com/searchenterpriseai/definition/Turing-test.

Taunk, K., De, S., Verma, S. & Swetapadma, A. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. V *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (str. 1255–1260). Pridobljeno na https://doi.org/10.1109/ICCS45141.2019.9065747.

Tavasoli, S. (9. 11. 2016). Top 10 Machine Learning Algorithms You Need to Know in 2023. *Simplilearn*. Pridobljeno na https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article.

Univerza v Ljubljani, FMF, Oddelek za matematiko in mehaniko. (b. d.). *Definicija algoritma*. http://www.educa.fmf.uni-lj.si/izodel/sola/2000/di/zabukovec/algoritmi/11def.htm.

Yeh, A. S., Hirschman, L., & Morgan, A. A. (2003). Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup. *Bioinformatics*, *19*(Suppl 1), i331–i339. Pridobljeno na https://doi.org/10.1093/bioinformatics/btg1046.

Zhu, X. & Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Springer International Publishing. Pridobljeno na https://doi.org/10.1007/978-3-031-01548-9.

Summary

Artificial intelligence has been influencing our lives for a long time. It has been present in archival science since the second half of the twentieth century, when computers were first used. Originally, they looked for patterns by using tools that could perform a search based on regular expressions. Technology, and consequently also artificial intelligence, has progressed tremendously since then. In the flood of terminology and concepts, experts who are not familiar with terms from the field of artificial intelligence, start doubting. A completely different aspect in archival science is the understanding of new technologies and their operations. It is very important to understand how the tool works, because just like in production, without a proper understanding of how the tools work, we cannot improve or optimize the operations. The present paper is a quick overview of a fragment in the large collection of artificial intelligence. Algorithms, as an integral part of any program, are also an integral part of machine learning here. This paper tries to show the differences between different types of machine learning in a simple way. All this with the aim of expanding the horizon of knowledge in archival science.