OTKRIVANJE ZNANJA I METODE RUDARENJA PODATAKA U PROIZVODNIM SISTEMIMA

Seid Žapčević[1], Peter Butala[2]
[1]Univerzitet u Bihaću, Tehnički fakultet, Dr. I. Ljubijankića bb, BA-77000 Bihać,
zapcevic.seid@gmail.com
[2]Univerza v Ljubljani, Fakulteta za strojništvo, Aškerčeva 6, SI-1000 Ljubljana,
peter.butala@fs.uni-lj.si

Ključne riječi: proizvodni sistem, otkrivanje znanja, rudarenje podataka.

SAŽETAK:

Važan preduvjet za prilagođavanje proizvodnih sistema na visoko konkurentnim globalnim tržištima je njihova sposobnost da uče. Proces učenja u proizvodnim sistemima je zasnovan na otkrivanju novog znanja i njegovom rastu. U radu je predstavljen pregled literature na području otkrivanja znanja i metoda za rudarenje podataka. Predstavljeni su: model procesa otkrivanja znanja, opisne i prediktivne metode za rudarenje podataka.

1. UVOD

Sa razvojem i primjenom kompjuteriziranih informacionih sistema na raspolaganju je sve više podataka, koji se skupljaju za vrijeme različitih procesa i uskladištavaju se u baze podataka. Tradicionalne statističke metode i alati za upravljanje podataka nisu više sposobni za analiziranje tako velikih količina podataka. Nekoliko domena gdje su velike količine podataka uskladištene u centraliziranim ili distribuiranim bazama podataka su [1]: financijska ulaganja, zdravstvena zaštita, proizvodnja, telekomunikacije, World Wide Web. U proizvodnoj industriji prikupljaju se podaci iz skoro svih procesa organizacije takvi kao projektiranje proizvoda i procesa, planiranje i upravljanje materijala, sklapanje, raspoređivanje, održavanje, recikliranje itd. Ove baze podataka nude ogroman potencijal kao izvori novog znanja. Rudarenje podataka je metodološko rješenje za transformiranje podataka u korisno znanje. Dakle, nove metode za rješavanje problema sa velikim količinama podataka u bazama podataka su postale ekstremno važne sa ciljem izvlačenja korisnih informacija i znanja za donošenje odluke. Ove metode uključuju otkrivanje znanja u bazama podataka i rudarenja podataka.

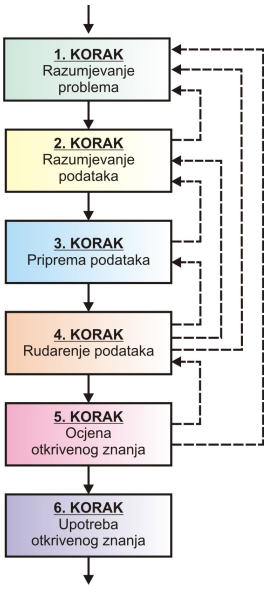
2. OTKRIVANJE ZNANJA I METODE RUDARENJA PODATAKA

Pristup otkrivanja znanja u bazama podataka traži novo znanje u određenom području primjene [2]. Ovaj pristup se definiše kao netrivijalni proces identificiranja validnih, novih, potencijalno korisnih, i na kraju razumljivih uzoraka u podacima. Proces generalizira izvore podataka koji nisu baze podataka, iako on naglašava baze podataka kao primarni izvor podataka. Na najvišem nivou taksonomije rudarenja podataka možemo razlikovati dva osnovna tipa [3 i 4]: (1) rudarenje orjentirano na verifikaciju i (2) rudarenje orjentirano na otkrivanje. Prvi tip je namjenjen potvrđivanju hipoteze, a drugi je namjenjen autonomnom traženju uzoraka i pravila.

U našem istraživanju upotrebljavamo metode, koje su orjentirane na otkrivanje novog znanja iz podataka, koji se prikupljaju u proizvodnji [5].

Proces otkrivanja znanja u podacima, također nazvan kao KDD (engl. knowledge discovery in data) se sastoji od višestrukih koraka, koji se izvršavaju u nizu. Svaki sljedeći korak je pokrenut na uspješnom završetku prethodnog koraka, i zahtjeva rezultat generiran prethodnim korakom kao njegov ulaz. Procedure koje se izvode u procesu otkrivanja znanja u bazama podataka opisane su modelom. Nekoliko različitih modela se predlaže u literaturi [6 – 10]. Svaki od ovih modela ima svoje prednosti i nedostatke koji zavise od područja primjene i pojedinih ciljeva poslovanja.

U sklopu ovog istraživanja prihvaćen je model procesa u šest koraka za otkrivanje znanja u bazama podataka kako je definisano u [8 i 11]. Cilj izbora modela procesa za otkrivanje znanja i rudarenje podataka je bio, da se izabere takav model, koji se potvrdio na praktičnim projektima rudarenja podataka i otkrivanja znanja. Takav model procesa je pomoć pri projektiranju izvedbi otkrivanja znanja, koji daje detaljni opis procedure u svakom koraku, od specifikacije problema, do implementacije rezultata. Izabrani model prikazan na slici 1, detaljno je opisan u [12].



Slika 1: Model procesa rudarenja podataka i otkrivanja znanja u šest koraka

Važni dijelovi procesa su iterativni i interaktivni aspekti. Petlje povratne veze su potrebne pošto bilo koje promjene i odluke urađene u jednom od koraka mogu rezultirati u promjenama u kasnijim koracima. U procesu otkrivanja znanja u bazama podataka (KDD), rudarenje podataka je ključni korak, koji vodi u otkrivanje znanja.

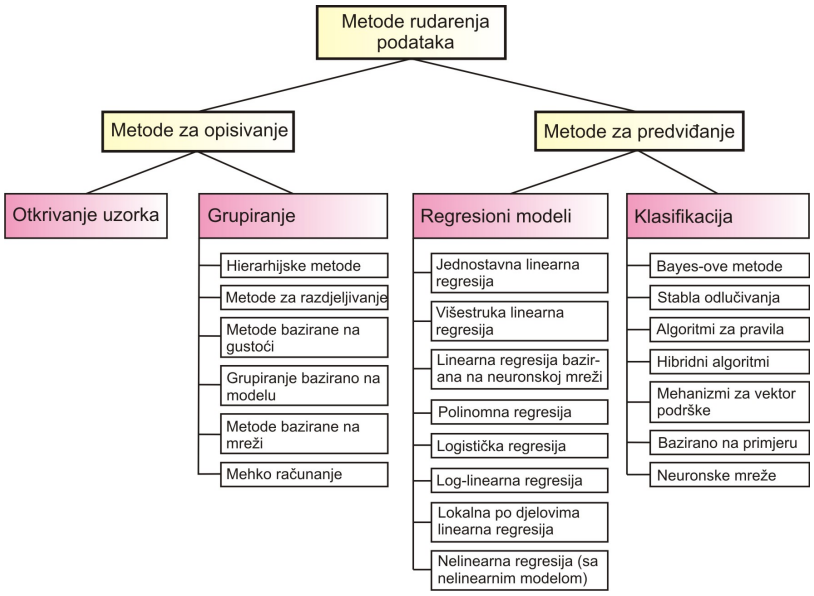
Rudarenje podataka (engl. data mining (DM)) se definira kao korak u procesu otkrivanja znanja u bazama podataka (KDD), koji se sastoji od analize podataka i algoritama za otkrivanje, koji pod prihvatljivim ograničenjima računarske učinkovitosti generira određene uzorke (ili modele) nad podacima [6]. Rudarenje podataka označava mješavinu koncepata i algoritama iz mašinskog učenja, statistike, vještačke inteligencije, i upravljanja podataka [13]. Sa pojavom DM, istraživači i praktičari počinju primjenu ove tehnologije na podacima da otkriju skrivene odnose ili uzorke.

U [14] su prikazane različite metode, koje se upotrebljavaju za izvršavanje funkcija rudarenja podataka. Ove funkcije se ostvaruju različitim metodama koje mogu biti kategorizirane kao: klasične ili moderne statističke tehnike, vještačke neuronske mreže, mehanizmi vektora podrške, algoritmi drveta za odlučivanje i indukcija pravila, pravila asocijacije, rasuđivanje bazirano na

slučaju, sistemi fuzzy logike i grubi skupovi, sekvencijalni uzorci, genetski algoritmi, evolucijsko programiranje, metode vizualizacije, itd. U svakoj od ovih grupa postoji više poznatih algoritama. Tako, na primjer, postoji više od stotinu implementacija algoritama stabala za odlučivanje.

U djelima [2,4,15,16,17] prikazane su mnoge metode rudarenja podataka, koje se koriste za različite svrhe i postizanje različitih ciljeva. Mnogi autori su pokušali klasificirati metode za rudarenje podataka, ali ne postoji jedna taksonomija metoda rudarenja podataka koja bi nam olakšala razumijevanje raznolikosti metoda i njihovih međusobnih povezanosti.

Metode za opisivanje su namjenjene interpretaciji podataka za lakše razumjevanje podataka među njima. Rezultat je prikaz uzoraka u podacima u vizualnom obliku. Pri metodama za predviđanje, ili prediktivnim metodama, cilj je izgradnja modela ponašanja, koji je sposoban predviđati vrijednosti jedne ili više varijabli s obzirom na promjenjeno početno stanje. Prikaz znanja je u obliku jednadžbi, koje je kasnije moguće jednostavno upotrijebiti. Neke metode za predviđanje također mogu pomoći u dobivanju razumjevanja podataka [4].



Slika 2: Klasifikacija metoda za rudarenje podataka i otkrivanje znanja [12]

2.1. Opisne metode rudarenja podataka

Opisni pristupi se dijele u dvije kategorije [17]: (1) identificiranje interesantnih uzoraka u podacima i (2) grupisanje podataka u značajne grupe. Algoritmi za otkrivanje uzoraka u veoma velikim skupovima podataka, kao što su podaci o prodaji, bili su ključne uspješne priče o istraživanju rudarenja podataka.

Najefikasniji algoritmi za traženje ovih uzoraka (tj. relacija) su „FP-growt“ [18] i „LPMiner“ [19]. Pored navedenih algoritama često se upotrebljavaju još algoritmi „Apriori“, „Eclot“ i „Realim“. Uzorci otkriveni iz podataka procesa mogu dati uvid u odnos između različitih značajki, i mogu se također koristiti za otkrivanje pravila asocijacije.

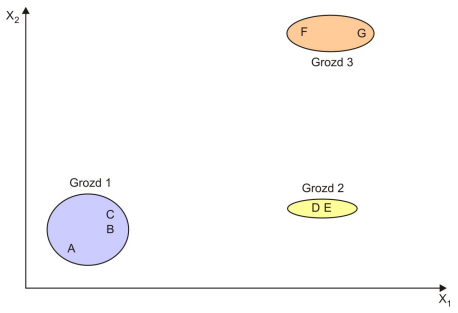
Druga opisna metoda rudarenja podataka je grupisanje, koja je nenadgledana klasifikacija uzoraka (zapažanja, proizvodnih podataka, ili vektora značajki) u grozdove [15]. S drugim rječima rečeno, grupisanje je proces identificiranja konačnog skupa kategorija ili grozdova da opiše podatke. Podaci koji pripadaju istim grupama izražavaju zajedničke karakteristike, dok ovo nije slučaj za

podatke u različitim grupama. Na osnovu [20] postojeće metode za grupisanje mogu se podijeliti na šest kategorija, kako je prikazano na slici 3.

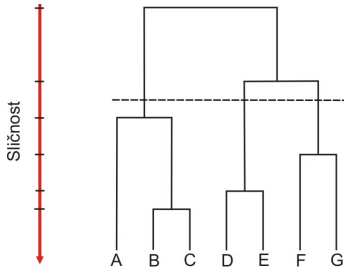
Hijerarhijske metode rudarenja podataka grade grozdove rekurzivno dijeleći instance na način odozgo prema dole ili odozdo prema gore. Hijerarhijske metode možemo podjeliti još na aglomerativno hijerarhijsko grupiranje i razdvajajuće hijerarhijsko grupiranje. Primjer upotrebe hijerarhijskog algoritma za grupiranje na dvo-dimenzionalnom skupu podataka ilustriran je na slici 4.

Slika 4 prikazuje sedam uzoraka označenih s A, B, C, D, E, F, i G, koji oblikuju tri grozda. Hijerarhijski algoritam daje „dendrogram“, koji predstavlja ugnježđeno grupiranje uzoraka i nivoe sličnosti, na kojem se grupacije povezuju. Termin „dendrogram“ znači dijagram drveta (grčka riječ „dendron“ znači „drvo“). Dendrogram odgovara za sedam tačaka na slici 4. Dendogram može biti prekinut na različitim nivoima da bi dobili različito grupiranje podataka.

Algoritmi za grupiranje razdjeljivanjem postižu jednu podjelu podataka umjesto strukture grupiranja, takve kao što dendrogram proizvodi hijerarhijskom tehnikom. Metode razdjeljivanja imaju prednosti u primjenama koje obuhvataju velike skupove podataka za koje konstrukcija dendrograma je računarski ograničavajući faktor. Problem koji prati upotrebu algoritama za razdjeljivanje je izbor broja željenih izlaznih grozdova. Tehnike za razdjeljivanje obično proizvode grozdove optimiziranjem funkcije kriterija definisane ili lokalno (na podskupu uzoraka) ili globalno (definisana preko svih uzoraka). Kombinatorijalna pretraga mogućih označavanja za optimalnu vrijednost kriterija je jasno računarski ograničavajući faktor. U praksi, dakle, algoritam se tipično pokreće više puta sa različitim početnim stanjima, i najbolja konfiguracija dobijena iz svih pokretanja se koristi kao izlazno grupiranje. Algoritmi za razdjeljivanje [15] se dalje mogu podjeliti na algoritme kvadratne greške, čiji je naj reprezentativniji algoritam „k-means“ (k-srednja vrijednost) i razdvajajući algoritam koji proizlazi iz teorije grafova.



Slika 3: Tačke koje pripadaju u tri grozda



Slika 4: Dendrogram dobijen koristeći algoritam „jedna-veza“

Metode bazirane gustoćom pretpostavljaju da tačke koje pripadaju svakom grozdu su izvučene iz specifične distribucije vjerovatnoće. Najpoznatiji algoritmi ove metode su: EM, DBSCAN, AUTOCLAS, SNOB i MCLUST.

Metode grupiranja bazirane na modelu pokušavaju optimizirati pristajanje između datih podataka i nekih matematičkih modela. Drukčije od konvencionalnih grupiranja, koje identificira grupe objekata, metode grupiranja bazirane na modelu također traže karakteristične opise za svaku grupu, gdje svaka grupa pretstavlja koncept ili klasu. Najčešće korištene metode indukcije su stabla za odlučivanje i neuronske mreže [20].

Metode bazirane na mreži razdjeljuju prostor u konačni broj ćelija koje formiraju strukturu mreže na kojoj se izvode sve operacije za grupiranje. Glavna prednost pristupa je njegovo brzo vrijeme obrade.

U metode mehkog računanja, pored neuronskih mreža spadaju još metode grupiranja, koje baziraju na modelu, kao što su mehko grupiranje (engl. fuzzy clustering), simulirano popuštanje (engl. simulated annealing for clustering) i evolucijski pristupi grupiranju. Najčešće korištene evolucijske tehnike su genetski algoritmi (GA).

2.2. Metode rudarenja podataka za predviđanje

U [21] rudarenje podataka za predviđanje se definiše kao proces automatskog stvaranja modela za klasifikaciju iz skupa primjera, nazvanog skup za vježbanje, koji pripada skupu klasa. Nakon što je model kreiran, može se koristiti za automatsko predviđanje klase drugih neklasificiranih primjera. Rudarenje podataka za predviđanje se primjenjuje u rangu tehnika, koje traže odnose između specifičnih varijabli (nazvane ciljane varijable) i drugih varijabli u skupu podataka. Metode rudarenja podataka za predviđanje mogu se provoditi za klasifikaciju, predviđanje vrijednosti, opisna pravila itd. Metode za predviđanje općenito uključuju statističke metode, drva za odlučivanje, algoritme za učenje pravila [22] i njihove hibride, vještačke neuronske mreže, i mehanizme za podršku vektora. Statističke metode, koje se upotrebljavaju za klasifikaciju i predviđanje, uključuju Bayes-ove metode i regresione modele [2]. Bayes-ov pristup se bavi obradom vjerojatnosne informacije predstavljene u obliku ranije vjerovatnosti i uvjetne vjerojatnosti. Najpoznatije metode su k-najbliži susjedi, i Bayes-ova mreža opisana u [23]. Model regresije razvija odnos između varijabli na osnovi raspoloživih podataka i iskorištava pretpostavke o statističkoj prirodi takvih podataka i karakteru mogućih grešaka. Zavisno o strukturi modela, modeli regresije mogu se podjeliti u sljedeće kategorije: jednostavna linearna regresija, višestruka linearna regresija, neuronskom mrežom bazirana linearna regresija, polinomna regresija, logistička regresija, log-linearna regresija, lokalna po dijelovima linearna regresija, nelinearna regresija (sa nelinearnim modelom), i neuronskom mrežom bazirana nelinearna regresija.

Drva za odlučivanje, algoritmi za gradnju pravila i njihovi hibridi formiraju drugu kategoriju rudarenja podataka za predviđanje i opisani su u [24]. Drva za odlučivanje su konstruisana na osnovi analize skupa primjera za treniranje za kojeg su poznate oznake klase [25]. Onda se drva za odlučivanje upotrebljavaju za klasifikaciju prethodno neviđenih primjera. Ako se treniraju na visoko kvalitetnim podacima, drva za odlučivanje mogu praviti veoma tačna predviđanja. Primjeri osnovnih i naprednijih verzija drva za odlučivanje su algoritmi: ID3, C4.5, ID5R, i 1RD. Najpoznatiji učenik stabla za odlučivanje je C4.5 [26] (C5.0 je njegova nedavna nadgradnja), koja se široko upotrebljava i također je inkorporirana u druge komercijalne alate za rudarenje podataka (npr. Clementine i Kepler).

U [2] prikazani su algoritmi za učenje pravila, koji su predstavljeni algoritmom DataSqueezer i hibridnim algoritmima CLIP4, koji kombinira najbolje katakteristike drva za odlučivanja i algoritama za pravila.

3. ZAKLJUČAK

Kako je jasno pokazano u pregledu literature, KDD i DM daju djelotvorne i učinkovite metode i alate za učenje, koje se mogu isto tako uspješno primijeniti u proizvodnji. U djelu [12] dat je pristup, koji koristi naučeno znanje iz podataka i upotrijebljava ga za donošenje odluka na sistematičan način. Zasnovano na ovoj spoznaji, uveden je novi koncept proizvodnog sistema, koji ima sposobnost učenja iz historije svoga rada, tj. iz prethodno izvedenih proizvodnih operacija i koji je to znanje također sposoban upotrijebiti na različitim nivoima odlučivanja i operiranja.

4. LITERATURA

[1] Mitra, S., Pal, S.K. & Mitra, P., 2002, Data mining in soft computing framework: A survey, IEEE Transactions on Neural Networks, 13:3-14.

[2] Cios, K.J., Pedrycz, W., Swiniarski R.W. & Kurgan, L.A., 2007, Data Mining, A Knowledge Discovery Approach, New York: Springer.

[3] Symeonidis, A.L. & Mitkas, P.A., 2005, Agent Intelligence through Data Mining, New York: Springer, 11-40.

[4] Maimon, O., Rokach, L., 2005, Introduction to knowledge discovery in databases, In: O. Maimon & L. Rokach, ed. Data mining and knowledge discovery handbook, New York: Springer.

[5] Žapčević, S. & Butala, P., 2013, Adaptive process control based on a self-learning mechanism

in autonomous manufacturing systems, International Journal of Advanced Manufacturing Technology, 66/(9-12):1725-1743.

[6] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996, From data mining to knowledge discovery in databases, AI Magazine, 17/3:37-54.

[7] Anand, S. & Buchner, A., 1998, Decision Support Using Data mining, Financial Times, London: Pitman Publishers.

[8] Cios, K., Teresinka, A., Konieczna, S., Potocka, J. & Sharma, S., 2000, A knowledge discovery approach to diagnosing myocardial perfusion, IEEE Engineering in Medicine and Biology Magazine, special issue on Medical Data Mining and Knowledge Discovery, 19/4:17-25.

[9] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A., 1998, Discovering Data Mining: From Concept to Implementation, New Jersey: Prentice Hall

[10] Shearer, C., 2000, The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing, 5/4:13-19.

[11] Cios, K.J. & Kurgan, L.A., 2005, Trends in data mining and knowledge discovery, In N.R. Pal, & L.C. Jain, eds. Advanced Technique in Knowledge Discovery and Data Mining, London: Springer, 1-26.

[12] Žapčević, S., 2013, Model samoučečega proizvodnega delovnega sistema, Doktorsko delo, Univerza v Ljubljani, Fakulteta za strojništvo, Ljubljana.

[13] Harding, J.A., Shahbaz, M., Srinivas, S. & Kusiak, A., 2006, Data mining in manufacturing: A review, Journal of Manufacturing Science and Engineering, 128:969-976.

[14] Wang, K., 2006, Data mining in manufacturing: the nature and implications, In K. Wang, G. Kovacs, M. Wozny & M. Fang, eds. International Federation for Information Processing (IFIP), Volume 207, Knowledge Enterprise: Intelligent Strategies In Product Design, Manufacturing, and Management, Boston: Springer, 1-10.

[15] Jain, A.K., Murty, M.N. & Flynn, P.J., 1999, Data clustering: a review, ACM Computing Surveys, 31:264-323.

[16] Lavrač, N., 1999, Selected techniques for data mining in medicine, Artificial Intelligence in Medicine, 16:3-23.

[17] Charaniya, S., Hu, W.S. & Karypis, G., 2008, Mining bioprocess data: opportunities and challenges, Trends in Biotechnology, 26/12:690-699.

[18] Han, J., Pei, J., Yin, Y. & Mao, R., 2004, Mining frequent patterns without candidate generation: a frequent-pattern tree approach, Data Mining and Knowledge Discovery, 8:53-87.

[19] Seno, M. & Karypis, G., 2001, LPMiner: an algorithm for finding frequent itemsets using length-decreasing support constraint, In N. Cercone, T.Y. Lin & X. Wu, eds. Proceedings of the 2001 IEEE International Conference on Data Mining, 29 November – 2 December 2001 San Jose, CA, Washington: IEEE Computer Society, 505-512.

[20] Rokach, L., Maimon, O., 2005, Clustering methods, In: O. Maimon & L. Rokach, ed. Data mining and knowledge discovery handbook, New York: Springer.

[21] Bharatheesh, T.L. & Iyengar S.S., 2004, Predictive data mining for delinquency modeling, In H.R. Arabnia, et al. eds. Proceedings of the 2004 International Conference on Embedded Systems and Applications, 21-24 June 2004, Las Vegas, USA, CSREA Press, 99-105.

[22] Fürnkranz, J., Gamberger, D., Lavrač, N., 2012, Fondations of rule learning, Heidelberg: Springer.

[23] Ben-Gal, I., 2007, Bayesian Networks. In F. Ruggeri, & R. Kennett, eds. Encyclopedia of Statistics in Quality & Reliability, Hoboken, NJ: Wiley & Sons.

[24] Witten, I.H. & Frank, E., 2005, Data mining: Practical Machine Learning Tools and Techniques, 2nd ed. San Francisco: Morgan Kaufman.

[25] Kingsford, C. & Salzberg, S.L., 2008, What are decision trees? Nature Biotechnology, 26:1011-1013.

[26] Quinlan, J.R., 1986, Introduction of decision trees. Machine Learning, 1:81-106