Elektronski fakultet Niš

Induktivno učenje, Primena ID3 algoritma

Student: Jovan Aritonovic Br. Indeksa: 1425

Uvod

Stablo odlučivanja je struktura koja sadrži čvorove i potege (strelice) i izgrađena je od skupa podataka. Svaki čvor se ili koristi za donošenje odluke (poznat kao čvor odluke) ili predstavlja ishod (poznat kao listni čvor). Najčešće se koriste kod problema klasifikacije Cilj: Model koji predviđa izlaznu vrednost na osnovu ulaznih vrednosti nekoliko parametara Ulazne vrednosti, kao i izlazne, mogu biti kategoričkog ili kontinualnog tipa Svaki čvor odgovara ulaznom parametru, svaka grana (koja potiče iz posmatranog čvora) odgovara nekoj vrednosti tog parametra Listovi predstavljaju izlazne parametre u zavisnosti od vrednosti ulaznih parametara na datom putu kroz stablo, od korena do posmatranog lista (to su obeležja klase, tzv. class label)

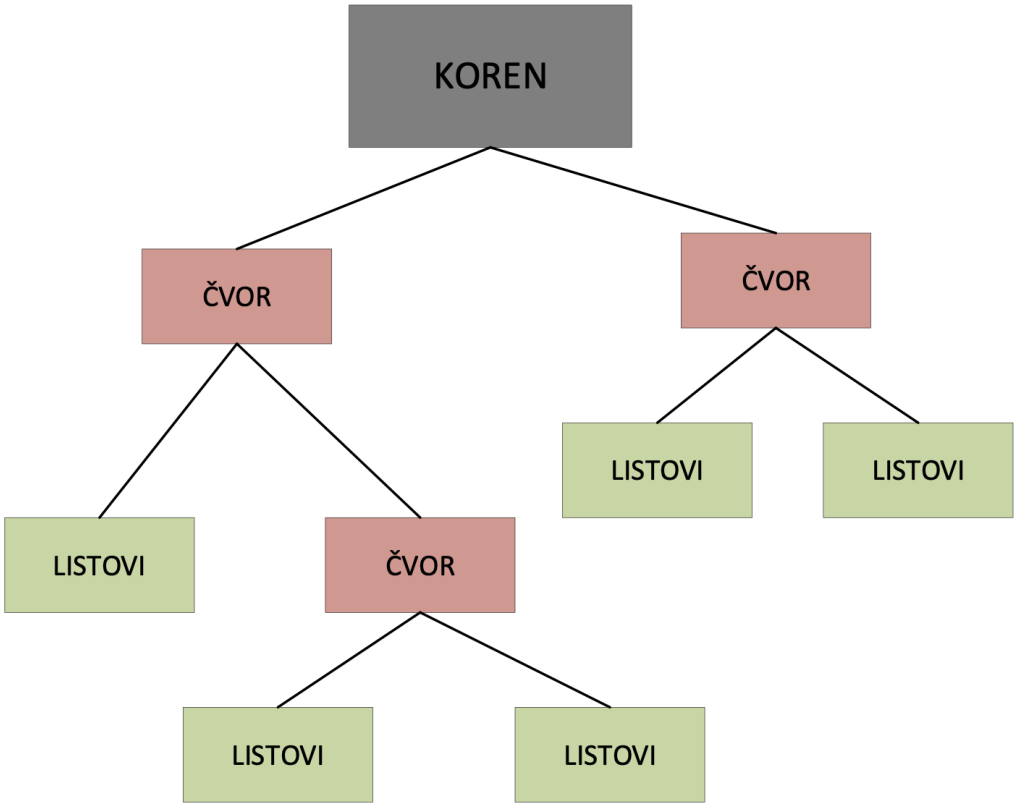
Stabla odlučivanja

Znanje koje se predstavlja u formi stabla odlučivanja ima svojstva hijerarhijske uređenosti, preglednosti i jednostavnosti u učenju, jer svako stablo predstavlja zapravo hijerahijski uređeni skup asocijativnih pravila. Zbog toga se stablo odlučivanja rado primenjuje kao tehnika za odlučivanje. Stabla odlučivanja koja se koriste u otkrivanju zakonitosti u odlučivanju su takozvana induktivna stabla odlučivanja. Sva induktivna stabla odlučivanja se sastoje iz korena, čvorova i listova. Koren je početni čvor po kome se skup podataka iz koga uči stablo. Listovi predstavljaju čvorove odluke tj predstavljaju krajnje čvorove u stablu odlučivanja.

Da bi se saznalo pravilo po kome je određeni slučaj klasifikovan ili procenjen potrebno je kretati se od korena stabla preko čvorova sve do listova. Postoji mnogo algoritama za za stabla odlučivanja. Pet najčešće korišćenih algoritama. U ovom radu ce biti prikazan i implementiran algoritam ID3.

Algoritam ID3

ID3 algoritam koristi kriterijum informacione dobiti (zasnovan na poznatom obrascu entropije da bi se odlučilo koji atribut treba da se koristi kao čvor pri granjanju stabla.Ovaj proces se radi iterativno dok se ne ispuni neki kriterijum zaustavljanja rasta stabla. Kriterijumi zaustavljanja rasta stabla su: Svi podaci u čvoru pripadaju istoj klasi Najveća informaciona dobit <=0. ID3 algoritam ne radi sa numeričkim atributima i nedostajućim vrednostima.



Entropija predstavlja meru neuređenosti sistema, tj neizvesnosti o tome koju odluku treba doneti. Jedan od načina da se entropija izmeri je i putem sledeće formule koju je definisao Klod Šenon.

$$H(S) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

Gde H(S) predstavlja entropiju skupa slučajeva S, pi verovatnoću da će biti donešena odluka i dok je n broj različitih odluka koje mogu biti donešene. Verovatnoće se mogu izračunati preko formule

$$p_i = \frac{|C_i|}{|S|}$$

Gde Ci je broj slučajeva sa odlukom i, a S ukupan broj slučajeva

Koriscenje algoritma:

1. Izračunajte *Information gain* svake karakteristike.
2. Uzimajući u obzir da svi redovi ne pripadaju istoj klasi, podeliti skup podataka **S** na podskupove koristeći funkciju za koju je dobijanje informacija maksimalno.
3. Napraviti čvor stabla odluka koristeći funkciju sa maksimalnim dobitkom informacija.
4. Ako svi redovi pripadaju istoj klasi, napraviti trenutni čvor kao lisni čvor sa klasom kao oznakom.
5. Ponavljajti za preostale karakteristike dok ne ponestane svih karakteristika ili dok stablo odlučivanja ne sadrži sve lisne čvorove.

Neke glavne **prednosti** ID3 su:

- Razumljiva pravila predviđanja koje se kreiraju iz podataka obuke.
- Gradi kratko stablo odlucivanja za relativno kratko vreme.
- Potrebno je samo da testira dovoljno atributa dok svi podaci ne budu klasifikovani.
- Pronalaženje lisnih čvorova omogućava da se podaci o testu skracuju, smanjujući broj testova.

ID3 može imati neke **nedostatke** u nekim slučajevima, npr.

- Podaci mogu biti previše uklopljeni ili preklasifikovani, ako se testira mali uzorak.
- Samo jedan po jedan atribut se testira za donošenje odluke.

Ulazni i izlazni parametri algoritma

Ulaz

- set za obuku (tabela sa podacima)
- Ovaj ulazni port očekuje PrimerSet. To je izlaz operatora Generate Nominal Data u priloženom Procesu Primera. Ovaj operator ne može da rukuje numeričkim atributima. Izlaz drugih operatora se takođe može koristiti kao ulaz.

Izlaz

- model (stablo odlučivanja)
- Stablo odlučivanja se isporučuje sa ovog izlaznog porta. Ovaj model klasifikacije se sada može primeniti na nevidljive skupove podataka za predviđanje atributa oznake.

- skup primera (tabela sa podacima)
- PrimerSet koji je dat kao ulaz se prosleđuje bez promene na izlaz kroz ovaj port. Ovo se obično koristi za ponovno korišćenje istog skupa primera u drugim operatorima ili za pregled skupa primera u radnom prostoru rezultata.

Implementacija ID3 uz pomoc PySwarms

PySwarms je alatka zasnovana na Python-u za predvidjanje vremena na osnovu postojeceg seta podataka. PySwarms implementira tehnike optimizacije roja sa više čestica na visokom nivou. Kao rezultat toga, teži da bude lak za upotrebu i prilagodljiv. Pomoćni moduli se takođe mogu koristiti da vam pomognu sa konkretnim problemom optimizacije. Primer implementacija ID3 optimizacije dodat je na git repozitorijum.

Literatura

1. https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm
2. https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1
3. https://en.wikipedia.org/wiki/ID3_algorithm
4. https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/id3.html