Seminarski rad

Izbor atributa (feature selection)

Milica Milutinović 2143

Sadržaj:

1. Uvod

Razvoj mašinskog učenja i savremenih sistema zasnovanih na podacima doveo je do sve veće dostupnosti skupova podataka koji sadrže veliki broj atributa. U mnogim realnim primenama, kao što su biomedicina, finansijska analiza, obrada prirodnog jezika i sistemi za preporuku, podaci mogu imati desetine, stotine ili čak hiljade ulaznih promenljivih. Iako se intuitivno može pretpostaviti da veći broj atributa doprinosi boljem kvalitetu modela, u praksi se često pokazuje da prisustvo velikog broja atributa ne mora da vodi ka boljim rezultatima.

Jedan od ključnih izazova u radu sa visokodimenzionalnim podacima jeste činjenica da svi atributi ne nose jednaku količinu informacija. Pored relevantnih atributa koji imaju direktan uticaj na ciljnu promenljivu, skupovi podataka često sadrže i irelevantne ili redundantne atribute. Irelevantni atributi ne doprinose procesu učenja, dok redundantni atributi sadrže informacije koje su već predstavljene kroz druge promenljive. Prisustvo ovakvih atributa može negativno uticati na proces treniranja modela, dovesti do pojave preprilagođavanja (eng. *overfitting*), povećati računske zahteve i otežati interpretaciju dobijenih rezultata.

Izbor atributa (eng. *feature selection*) predstavlja jedan od ključnih koraka u procesu izgradnje modela mašinskog učenja, čiji je cilj identifikacija i zadržavanje samo onih atributa koji su najznačajniji za rešavanje posmatranog problema. Za razliku od tehnika ekstrakcije atributa, koje transformišu originalne podatke u novi prostor karakteristika, izbor atributa zadržava originalno značenje promenljivih, čime se omogućava bolja interpretabilnost modela. Ovo je posebno važno u oblastima gde je razumevanje odluka modela jednako važno kao i sama tačnost predikcije, kao što su medicina ili finansije.

Primenom odgovarajućih tehnika izbora atributa moguće je smanjiti dimenzionalnost podataka, poboljšati performanse modela i skratiti vreme treniranja, uz istovremeno očuvanje ili čak unapređenje kvaliteta predikcije. Pored toga, jednostavniji modeli sa manjim brojem atributa često imaju bolju sposobnost generalizacije na nove podatke, što ih čini pouzdanijim u realnim uslovima primene. Iz tog razloga, izbor atributa se smatra ne samo tehnikom za optimizaciju performansi, već i važnim alatom za unapređenje robusnosti i pouzdanosti modela mašinskog učenja.

Cilj ovog seminarskog rada jeste da se detaljno analizira problem izbora atributa, prikažu različite tehnike koje se koriste za njegovo rešavanje, kao i da se razmotre njihovi osnovni principi, prednosti i ograničenja. Poseban akcenat biće stavljen na razumevanje uslova pod kojima se pojedine metode primenjuju, kao i na davanje preporuka za njihovu praktičnu upotrebu u različitim scenarijima mašinskog učenja.

2. Identifikacija problema

Jedan od osnovnih izazova u savremenim primenama mašinskog učenja jeste rad sa skupovima podataka koji sadrže veliki broj atributa. Takvi skupovi podataka nazivaju se visokodimenzionalnim i karakteristični su za mnoge oblasti, uključujući biomedicinu, obradu prirodnog jezika, analizu finansijskih tržišta i sisteme za preporuku. U ovim domenima broj atributa često je uporediv ili čak znatno veći od broja dostupnih uzoraka, što predstavlja ozbiljan problem za većinu algoritama mašinskog učenja.

Visokodimenzionalnost podataka dovodi do pojave poznate kao „prokletstvo dimenzionalnosti" (eng. *curse of dimensionality*). Ovaj pojam, koji je prvi put uveo Richard Bellman 1961. godine, opisuje skup problema koji nastaju usled eksponencijalnog rasta prostora podataka sa povećanjem broja atributa. Na primer, u visokodimenzionalnom prostoru, skoro sve tačke postaju približno jednako udaljene jedna od druge, što čini koncepte poput „blizine" ili „sličnosti" manje značajnim. Posledica toga je da mnogi algoritmi, naročito oni koji se oslanjaju na mere udaljenosti, kao što su algoritam k-najbližih suseda (k-NN) ili metode zasnovane na klasterovanju, postaju manje efikasni i manje pouzdani.

Pored problema sa generalizacijom, visoka dimenzionalnost podataka ima i značajne praktične posledice. Povećanje broja atributa direktno utiče na računsku složenost algoritama za treniranje modela. Mnogi algoritmi imaju vremensku i prostornu složenost koja eksponencijalno ili polinomijalno raste sa brojem atributa, što može učiniti proces učenja neizvodljivim ili praktično neupotrebljivim za skupove podataka sa nekoliko hiljada atributa. Ovo je posebno izraženo kod metoda kao što su stabla odlučivanja, mašine potpornih vektora (SVM) ili neuronske mreže sa velikim brojem parametara.

Model koji je preprilagođen pokazuje visoke performanse na skupu za treniranje, ali slabo generalizuje na nove, neviđene podatke. Veliki broj irelevantnih atributa dodatno povećava rizik od ove pojave, jer model dobija više slobode da se prilagodi specifičnostima podataka za treniranje, uključujući i one karakteristike koje predstavljaju samo šum ili slučajne fluktuacije. Umesto da model uči relevantne obrasce koji su prisutni u podacima, on može pokušavati da prilagodi slučajne varijacije, što direktno vodi ka smanjenju pouzdanosti u realnim uslovima primene.

Dodatni problem koji proizlazi iz velikog broja atributa jeste otežana interpretabilnost modela. U mnogim primenama, posebno u medicini, finansijama ili pravnim sistemima, nije dovoljno samo dobiti tačnu predikciju – potrebno je i razumeti na osnovu kojih faktora je model doneo odluku. Model sa stotinama ili hiljadama atributa postaje praktično nemoguće interpretirati, što može biti prepreka za njegovu primenu u praksi, bez obzira na postignute performanse.

3. Osnovni pojmovi i koncepti

Pre detaljne analize tehnika izbora atributa, neophodno je precizno definisati osnovne pojmove i koncepte koji čine teorijsku osnovu ove oblasti. Razumevanje ovih koncepata ključno je za pravilnu primenu metoda izbora atributa i interpretaciju dobijenih rezultata.

3.1 Atributi i prostor atributa

Atribut (eng. *feature*) predstavlja merljivu karakteristiku ili svojstvo objekta koji se posmatra. U kontekstu mašinskog učenja, atributi čine ulazne promenljive na osnovu kojih model uči da predviđa ciljnu promenljivu. Na primer, u zadatku predikcije cene nekretnine, atributi mogu biti površina stana, broj soba, lokacija, sprat, godina izgradnje i slično. Svaki atribut može biti različitog tipa: numerički (kontinualan ili diskretan), kategorički (nominalan ili ordinalan), binarni ili tekstualni.

Prostor atributa (eng. *feature space*) predstavlja matematički prostor definisan svim dostupnim atributima. Ako skup podataka sadrži n atributa, svaki uzorak može se predstaviti kao tačka u n-dimenzionalnom prostoru. Dimenzionalnost ovog prostora direktno odgovara broju atributa, pa se često koristi termin „dimenzionalnost podataka" kada se govori o broju atributa.

3.2 Relevantnost, irelevantnost i redundantnost atributa

Relevantnost atributa odnosi se na stepen u kojem određeni atribut doprinosi predikciji ciljne promenljive. Relevantan atribut sadrži informacije koje su korisne za razlikovanje različitih klasa ili za preciznu predikciju vrednosti izlaza. Formalno, atribut X je relevantan za ciljnu promenljivu Y ako postoji zavisnost između njih, odnosno ako važi $P(Y|X) \neq P(Y)$.

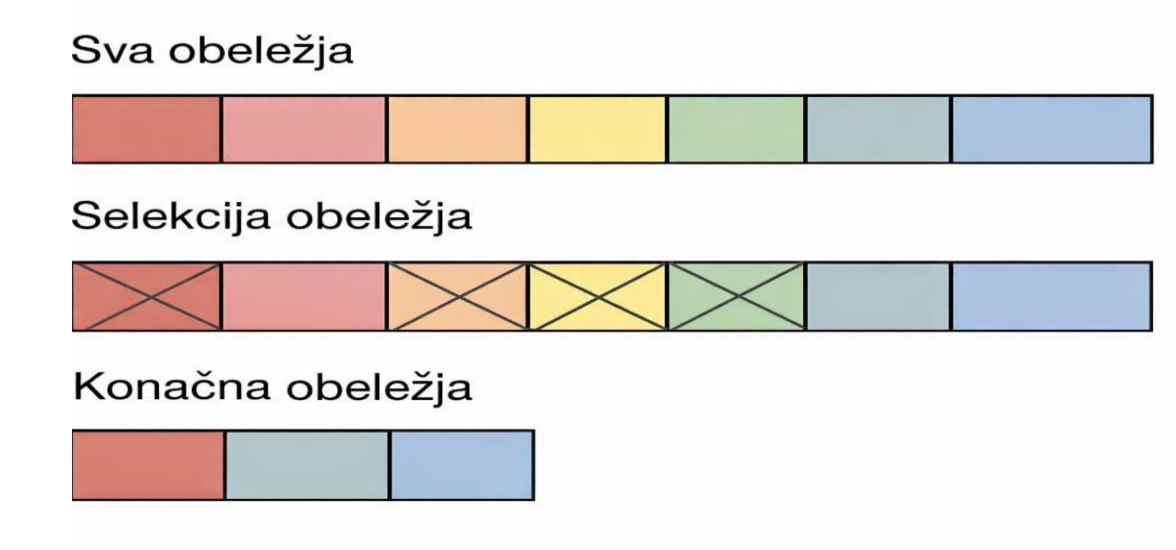
Irelevantni atributi su oni koji ne nose nikakvu informaciju o ciljnoj promenljivoj. Prisustvo irelevantnih atributa u skupu podataka može dovesti do problema jer model može pokušati da pronađe obrasce koji ne postoje, što rezultira preprilagođavanjem. Irelevantan atribut X karakteriše se time da je $P(Y|X) = P(Y)$, što znači da poznavanje vrednosti tog atributa ne menja verovatnoću ciljne promenljive.

Redundantni atributi su oni koji nose informacije koje su već sadržane u drugim atributima. Dva atributa X_1 i X_2 su redundantni ako postoji visoka korelacija između njih, odnosno ako poznavanje jednog atributa omogućava pouzdanu predikciju vrednosti drugog. Na primer, ako skup podataka sadrži i temperaturu u Celzijusima i temperaturu u Farenhajtima, jedan od ova dva atributa je redundantan jer postoji deterministično preslikavanje između njih. Redundantni atributi ne samo da nepotrebno povećavaju dimenzionalnost podataka, već mogu i negativno uticati na performanse određenih algoritama.

3.3 Definicija izbora atributa

Izbor atributa predstavlja proces identifikacije i selekcije podskupa najznačajnijih atributa iz originalnog skupa podataka. Cilj ovog procesa je da se zadrže samo oni atributi koji maksimalno doprinose performansama modela, dok se istovremeno eliminišu irelevantni i redundantni atributi.

Formalno, problem izbora atributa može se definisati na sledeći način: dat je skup podataka sa n atributa $F = \{f_1, f_2, ..., f_n\}$, potrebno je pronaći podskup $S \subseteq F$, gde je $|S| = k < n$, takav da model treniran na atributima iz skupa S ima optimalne performanse prema određenom kriterijumu evaluacije.



Ilustracija feature selection-a

3.4 Razlika između izbora i ekstrakcije atributa

Važno je razlikovati izbor atributa od **ekstrakcije atributa** (eng. *feature extraction*), pošto su ovo dva fundamentalno različita pristupa redukciji dimenzionalnosti podataka.

Izbor atributa zadržava originalne atribute u njihovom izvornom obliku. Rezultat procesa izbora atributa je podskup originalnih atributa, što znači da se značenje i interpretabilnost podataka u potpunosti čuvaju. Ovo je posebno važno u domenima gde je potrebno razumeti koje karakteristike utiču na donošenje odluka, kao što su medicinska dijagnostika ili kreditno skorovanje.

Sa druge strane, ekstrakcija atributa kreira nove atribute transformacijom ili kombinovanjem originalnih atributa. Metode kao što su analiza glavnih komponenti (PCA), linearna diskriminantna analiza (LDA) ili autoenkoderi transformišu originalni prostor atributa u novi prostor niže dimenzionalnosti. Novi atributi su linearne ili nelinearne kombinacije originalnih atributa i često nemaju direktnu fizičku interpretaciju. Iako ekstrakcija atributa može biti veoma efikasna za

redukciju dimenzionalnosti i očuvanje relevantnih informacija, ona nije pogodna kada je interpretabilnost modela prioritet.

3.5 Kriterijumi evaluacije

Izbor atributa zahteva definisanje kriterijuma prema kojem će se ocenjivati kvalitet odabranog podskupa atributa. Postoji nekoliko osnovnih pristupa evaluaciji:

Tačnost predikcije predstavlja najčešće korišćen kriterijum i meri se kroz performanse modela na skupu za validaciju ili testiranje. Za probleme klasifikacije koriste se metrike kao što su tačnost (eng. *accuracy*), preciznost (eng. *precision*), odziv (eng. *recall*) i F1 mera, dok se za regresione probleme koriste srednja kvadratna greška (MSE), srednja apsolutna greška (MAE) ili koeficijent determinacije (R^2).

Računska efikasnost odnosi se na vreme potrebno za treniranje modela i izvršavanje predikcije. Smanjenje broja atributa direktno utiče na smanjenje računske složenosti, što je posebno značajno za primene u realnom vremenu ili za rad sa velikim skupovima podataka.

Interpretabilnost podrazumeva sposobnost razumevanja kako model donosi odluke. Manji broj atributa omogućava lakše razumevanje odnosa između ulaznih promenljivih i ciljne promenljive, što je ključno u domenima gde je potrebno obrazložiti donete odluke.

Stabilnost izbora atributa odnosi se na konzistentnost odabranih atributa pri malim promenama u skupu podataka. Stabilna metoda izbora atributa treba da identifikuje isti ili vrlo sličan podskup atributa čak i kada se trening podaci neznatno promene.

Kombinatorička priroda problema izbora atributa predstavlja značajan izazov, kako broj mogućih kombinacija raste eksponencijalno sa povećanjem broja atributa. Zbog toga su razvijene različite strategije i metode koje omogućavaju pronalaženje prihvatljivih rešenja u razumnom vremenskom okviru.

4. Moguća rešenja - izbor atributa

Zbog problema izbora najznačajnijih atributa, istraživači su razvili različite pristupe koji se razlikuju po načinu na koji evaluiraju kvalitet atributa i po odnosu prema algoritmu mašinskog učenja koji će se koristiti. Sve metode izbora atributa mogu se svrstati u tri osnovne kategorije: filter metode, wrapper metode i embedded metode. Pored ove osnovne podele, postoje i hibridne metode koje kombinuju prednosti različitih pristupa.

4.1 Osnovne kategorije metoda

Filter metode predstavljaju najjednostavniji pristup izboru atributa. One evaluiraju relevantnost svakog atributa nezavisno od algoritma koji će kasnije biti korišćen za učenje modela. Filter metode oslanjaju se na statističke mere ili mere iz teorije informacija kako bi ocenile odnos između svakog atributa i ciljne promenljive. Proces selekcije odvija se pre primene algoritma mašinskog učenja, što znači da se atributi biraju kao korak pretprocesiranja podataka. Ovaj pristup je računski efikasan i skalabilan, što ga čini pogodnim za rad sa velikim skupovima podataka.

Wrapper metode koriste algoritam mašinskog učenja kao "crnu kutiju" za evaluaciju kvaliteta podskupa atributa. Umesto da direktno mere relevantnost atributa, wrapper metode treniraju model sa različitim kombinacijama atributa i biraju onaj podskup koji daje najbolje performanse. Ovaj pristup uzima u obzir interakcije između atributa i specifičnosti algoritma učenja, što često dovodi do boljih rezultata u odnosu na filter metode. Međutim, wrapper metode su računski zahtevnije jer zahtevaju višestruko treniranje modela.

Embedded metode integrisani su u sam proces učenja algoritma. One biraju atribute tokom treniranja modela, često kroz mehanizme regularizacije ili kroz prirodu algoritma koji automatski dodeljuje različite težine atributima. Ovaj pristup balansira prednosti filter i wrapper metoda, pružajući dobru tačnost uz prihvatljivu računsku složenost.

4. 2 Hibridne metode

Hibridne metode kombinuju elemente različitih pristupa kako bi iskoristile njihove prednosti i ublažile nedostatke. Najčešći pristup je kombinacija filter i wrapper metoda, gde se filter metode koriste za brzu eliminaciju očigledno irelevantnih atributa, nakon čega wrapper metode vrše finiju selekciju iz preostale grupe atributa. Ovakav dvostepeni pristup omogućava da se smanji računska složenost wrapper metoda, jer one rade sa manjim skupom kandidata, uz istovremeno zadržavanje njihove preciznosti.

4.3 Strategije pretrage prostora atributa

Bez obzira na konkretnu kategoriju metoda izbora atributa, proces selekcije podrazumeva pretragu kroz prostor mogućih kombinacija atributa. Budući da broj mogućih podskupova raste eksponencijalno sa brojem atributa, izbor strategije pretrage ima ključan uticaj na izvodljivost i efikasnost algoritma.

Iscrpna pretraga (eng. *exhaustive search*) razmatra sve moguće podskupove atributa i bira onaj koji optimizuje definisani kriterijum kvaliteta. Iako garantuje pronalaženje optimalnog rešenja, ovaj pristup ima eksponencijalnu računsku složenost i praktično je neizvodljiv za skupove podataka sa većim brojem atributa.

Selekcija unapred (eng. *forward selection*) započinje pretragu od praznog skupa atributa i iterativno dodaje jedan po jedan atribut prema određenom kriterijumu. Ova strategija ima znatno manju složenost u poređenju sa iscrpnom pretragom i često se primenjuje kada se očekuje da je relevantan broj atributa relativno mali.

Eliminacija unazad (eng. *backward elimination*) polazi od kompletnog skupa atributa i iterativno uklanja atribute koji najmanje doprinose kvalitetu rešenja. Ovaj pristup je naročito pogodan kada se pretpostavlja da većina atributa nosi korisne informacije.

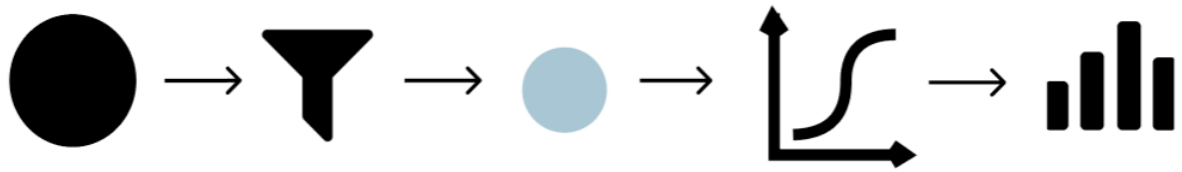
Dvosmerna pretraga (eng. *bidirectional search*) kombinuje selekciju unapred i eliminaciju unazad, omogućavajući istovremeno dodavanje i uklanjanje atributa. Ovakav pristup često omogućava bržu konvergenciju u odnosu na jednosmerne strategije.

Pored determinističkih metoda, koriste se i **metaheurističke strategije pretrage**, kao što su genetski algoritmi, optimizacija rojem čestica (PSO) i simulirano kaljenje. Ove metode koriste probabilističke mehanizme pretrage i ne garantuju pronalaženje globalnog optimuma, ali se često pokazuju efikasnim u radu sa visokodimenzionalnim skupovima podataka.

Izbor strategije pretrage predstavlja kompromis između kvaliteta dobijenog rešenja i potrebnog računskog vremena. Odgovarajuća strategija zavisi od dimenzionalnosti podataka, dostupnih računskih resursa i zahteva konkretnog problema.

5. Filter metode

Filter metode predstavlja najstariji i najčešće korišćen pristup izboru atributa. Osnovna karakteristika ovih metoda jeste da one evaluiraju relevantnost atributa nezavisno od algoritma mašinskog učenja koji će se kasnije primeniti na podatke. Proces selekcije zasniva se isključivo na statističkim svojstvima podataka, što filter metode čini brzim, skalabilnim i pogodnim za rad sa visokodimenzionalnim skupovima podataka.



Ilustracija filter metode

5.1 Princip rada i opšti algoritam

Filter metode funkcionišu kao korak pretprocesiranja podataka, gde se svakom atributu dodeljuje numerička ocena koja odražava njegovu relevantnost za ciljnu promenljivu. Za razliku od wrapper metoda koje koriste algoritam mašinskog učenja za evaluaciju, filter metode se oslanjaju isključivo na statističke mere ili mere zasnovane na teoriji informacija.

Opšti algoritam filter metoda može se formalno predstaviti na sledeći način:

ULAZ:

- $D = \{X, L\}$ // skup podataka za treniranje sa n atributa gde je
- $X = \{f_1, f_2, f_3, ..., f_n\}$ // skup atributa
- L // vektor oznaka (ciljna promenljiva)
- X' // unapred definisan inicijalni podskup atributa ($X' \subset X$ ili $X' = \{\emptyset\}$)
- θ // kriterijum zaustavljanja

IZLAZ:

- X'_{opt} // optimalan podskup atributa

Algoritam:

Početak:
Inicijalizacija:

- $X_{opt} = X'$;
- $\varphi_{opt} = E(X', I_m)$; // evaluacija X' korišćenjem nezavisne mere I_m

Ponavljaj:

- X_γ = *generate*(X); // generisanje podskupa za evaluaciju
- $\varphi = E(X_\gamma, I_m)$; // evaluacija trenutnog podskupa X_γ korišćenjem mere I_m
- **Ako** ($\varphi > \varphi_{opt}$):
    - $\varphi_{opt} = \varphi$;
    - $X'_{opt} = X_\gamma$;

Sve dok kriterijum θ nije dostignut;
Vrati X'_{opt};
Kraj;

Ključni elementi algoritma:

Nezavisna mera I_m - statistička ili informaciona mera koja kvantifikuje relevantnost atributa (npr. Pearsonova korelacija, međusobna informacija, χ^2 statistika, F-statistika). Ova mera ne zavisi od algoritma mašinskog učenja koji će se koristiti.
Funkcija evaluacije $E(\cdot)$ - prima podskup atributa i vraća numeričku ocenu φ kvaliteta tog podskupa.
Funkcija *generate*(·) - generiše kandidate za podskupove atributa (rangiranje, sekvencijalno dodavanje/uklanjanje, metaheuristike).
Kriterijum zaustavljanja θ - definiše završetak algoritma (dostignut broj atributa k, postignuta ocena, evaluirani svi podskupovi, ili prestanak poboljšanja).

Ovakav pristup ima linearnu složenost O(n), gde je n broj atributa, što čini filter metode izuzetno efikasnim čak i za visokodimenzionalne skupove podataka sa hiljadama atributa. Međutim, pošto se svaki atribut ocenjuje nezavisno, ovaj pristup ne može identifikovati interakcije između atributa ili redundantnost. Zbog toga se u praksi često primenjuju dodatni koraci za detekciju i eliminaciju redundantnih atributa.

5.2. Podela filter metoda

Filter metode se dele na **univarijantne** i **multivarijantne** metode. Univarijantne metode ocenjuju svaki atribut pojedinačno, nezavisno od ostalih atributa, merenjem direktnog odnosa sa ciljnom promenljivom. Ovaj pristup je računski efikasan, ali može dovesti do izbora redundantnih atributa. Multivarijantne metode razmatraju atribute u kontekstu međuzavisnosti i mogu identifikovati redundantnost, ali su računski složenije. U praksi se multivarijantne metode često primenjuju nakon univarijantnih, kao druga faza selekcije. Dodatna podela filter metoda odnosi se na to da li koriste informaciju o ciljnoj promenljivoj. **Supervised filter metode** evaluiraju atribute u odnosu na ciljnu promenljivu (Pearsonova korelacija, ANOVA, Mutual Information, Chi-square). **Unsupervised filter metode** evaluiraju atribute isključivo na osnovu unutrašnje strukture podataka, bez uvida u ciljnu promenljivu (Laplacian Score, uklanjanje po varijansi, korelacioni filter).

5.3. Osnovne filter metode za pretprocesiranje

Pre primene sofisticiranijih tehnika, preporučuje se eliminacija očigledno neupotrebljivih atributa. **Uklanjanje konstantnih atributa** odnosi se na atribute koji imaju istu vrednost za sve uzorke. Za numeričke atribute, konstantnost se identifikuje računanjem varijanse:

$$\mathrm{Var}(X_j) = (1/n) \sum_{i=1}^{n} (x_{ij} - \bar{x_j})^2$$

Ako je $\mathrm{Var}(X_j) = 0$, atribut se eliminiše. **Uklanjanje kvazi-konstantnih atributa** primenjuje se kada jedna vrednost dominira (npr. 98% uzoraka). **Uklanjanje dupliciranih atributa** odnosi se na identifikaciju atributa koji su identični ($X_i = X_j$ za sve uzorke).

Uklanjanje atributa sa visokim procentom nedostajućih vrednosti predstavlja još jedan osnovni pretprocesni korak. Atributi kod kojih nedostaje više od unapred definisanog praga uzoraka (npr. 50%) ne mogu pouzdano doprineti modelu, a imputacija takvih atributa unosi previše veštačke informacije. Preporučuje se njihova eliminacija pre primene sofisticiranijih tehnika selekcije.

5.4. Mere zasnovane na korelaciji

Pearsonov koeficijent korelacije meri linearnu zavisnost između numeričkog atributa X i ciljne promenljive Y:

$$\rho(X, Y) = \mathrm{Cov}(X, Y) / (\sigma_x \cdot \sigma_y)$$

Vrednost se kreće u intervalu [-1, 1], gde $\rho = 1$ označava savršenu pozitivnu, $\rho = -1$ savršenu negativnu, a $\rho = 0$ odsustvo linearne korelacije. Atributi sa $|\rho| > 0.5$ smatraju se relevantnim. Ograničenja uključuju detekciju samo linearnih odnosa, osetljivost na autlajere i pretpostavku normalne raspodele.

Spearmanov koeficijent korelacije ranga predstavlja neparametarsku alternativu koja meri monotonost odnosa:

$$\rho_s = 1 - (6\sum_i d_i^2) / (n(n^2 - 1))$$

gde je d_i razlika između rangova. Spearmanova korelacija je robusnija prema autlajerima i može detektovati monotoničke nelinearne odnose bez pretpostavke o raspodeli podataka.

Detekcija redundantnosti koristi korelaciju za identifikaciju atributa koji su visoko korelisani međusobno. Ako dva atributa imaju $|\rho| > 0.8$, jedan se eliminiše, pri čemu se zadržava onaj sa jačom korelacijom sa ciljnom promenljivom.

Korelacioni filter kao metoda selekcije primenjuje se u dva koraka: najpre se računa korelacija svakog atributa sa ciljnom promenljivom, a zatim se između parova visoko međusobno korelisanih atributa ($|\rho|$ > definisani prag, npr. 0.9) zadržava samo onaj sa većom korelacijom prema cilju. Ovaj pristup direktno adresira problem redundantnosti koji univarijantne metode ne mogu da reše, i posebno je efikasan kao pretprocesni korak pre primene skuplje wrapper ili embedded metode.

5.5. Statistički testovi hipoteza

Hi-kvadrat test (χ^2) koristi se za kategoričke atribute i ciljnu promenljivu:

$$\chi^2 = \sum\sum [(O_{ij} - E_{ij})^2 / E_{ij}]$$

gde je O_{ij} opservovana, a $E_{ij} = (n_i \cdot n_j) / n$ očekivana frekvencija pod nezavisnoš ću. Velika χ^2 vrednost i niska p-vrednost (< 0.05) ukazuju na relevantnost atributa.

ANOVA F-test primenjuje se kada je atribut numerički, a ciljna promenljiva kategorička:

$$F = [\sum_k n_k(\bar{x}_k - \bar{x})^2 / (K - 1)] / [\sum_k \sum_i (x_{ki} - \bar{x}_k)^2 / (n - K)]$$

gde je K broj klasa. Visoka F-statistika ukazuje da atribut dobro diskriminiše između klasa. ANOVA pretpostavlja normalnu raspodelu i homogenost varijansi.

5.6. Mere zasnovane na teoriji informacija

Entropija meri nesigurnost u raspodeli verovatnoća. Za diskretnu promenljivu Y:

$$H(Y) = -\sum_{j=1}^{m} P(y_j) \log_2 P(y_j)$$

Maksimalna entropija označava uniformnu raspodelu (maksimalna nesigurnost), minimalna (H = 0) izvesnost jedne vrednosti.

Međusobna informacija (MI) kvantifikuje zavisnost između promenljivih:

$$I(X; Y) = H(Y) - H(Y|X) = \sum_x \sum_y P(x, y) \log_2[P(x, y) / (P(x)P(y))]$$

Pokazuje za koliko poznavanje atributa X smanjuje nesigurnost o Y. Ako su nezavisni, $I(X; Y) = 0$. Ključna prednost je detekcija bilo kakve statističke zavisnosti, uključujući nelinearne odnose. **Information gain** je ekvivalentan MI i koristi se u stablima odlučivanja:

$$IG(Y, X) = H(Y) - \sum_v [P(X = v) \cdot H(Y|X = v)]$$

Unsupervised filter metode: Laplacian Score

Laplacian Score je unsupervised filter metoda koja rangira atribute prema tome koliko dobro čuvaju lokalnu geometrijsku strukturu podataka. Metoda se zasniva na teoriji grafova: konstruiše se KNN graf susedstva između uzoraka, a Laplasijanova matrica grafa koristi se za merenje koliko svaki atribut reflektuje lokalne odnose između uzoraka.

Za svaki atribut f računa se skor:

$$LS(f) = (f^T Lf) / (f^T Df)$$

gde je L = D - W Laplasijanova matrica, W matrica težina KNN grafa, a D dijagonalna matrica stepena čvorova. **Manji skor znači važniji atribut:** atribut bolje čuva lokalnu strukturu.

Ključna razlika u odnosu na supervised filter metode jeste da Laplacian Score ne koristi ciljnu promenljivu, pa visoko rangiran atribut ne mora biti i diskriminativan prema klasama. Zbog toga se ova metoda koristi kao dopuna supervised pristupima, ili u scenarijima gde ciljna promenljiva nije dostupna.

5.7. Prednosti i nedostaci filter metoda

Prednosti: Računska efikasnost omogućava obradu hiljada atributa u kratkom vremenu jer se svaki atribut evaluira nezavisno i samo jednom. Nezavisnost od algoritma učenja pruža fleksibilnost – odabrani atributi mogu se koristiti sa bilo kojim modelom. Otpornost na preprilagođavanje je značajna jer selekcija ne uključuje trening modela. Jednostavnost implementacije i interpretacije čini filter metode pristupačnim sa jasnom interpretacijom statističkih mera.

Nedostaci: Ignorisanje interakcija između atributa je glavni nedostatak – kombinacija pojedinačno slabih atributa može biti informativna. Nezavisnost od algoritma može biti i mana jer se ne uzimaju specifičnosti modela. Izbor optimalnog broja atributa nije definitivan – filter metode daju rangove bez jasnog odgovora koliko atributa zadržati. Osetljivost na redundantnost se javlja jer metode mogu dodeliti visoke ocene redundantnim atributima bez eliminacije preklapanja.

5.8. Praktična primena filter metoda

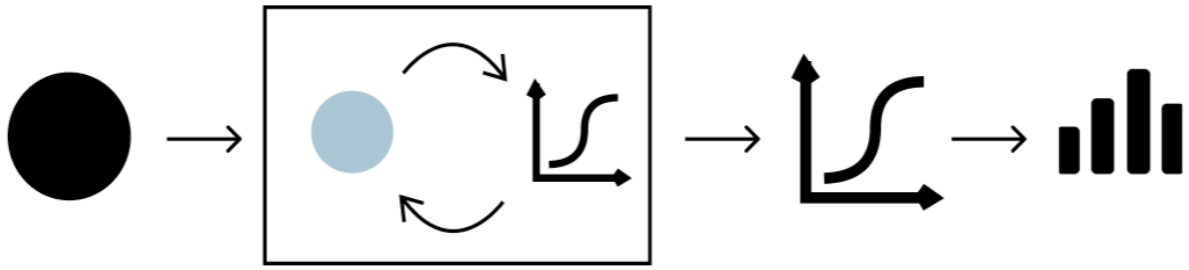
U **bioinformatici**, filter metode se koriste za analizu genskih ekspresija sa 20,000+ gena. Statistički testovi (t-test, ANOVA) omogućavaju brzu redukciju na nekoliko stotina najznačajnijih gena koji razlikuju zdrave od obolelih pacijenata. U **obradi prirodnog jezika**, χ^2 test i međusobna informacija selektuju najinformativnije reči ili n-grame u klasifikaciji tekstova, detekciji spama ili analizi sentimenta, gde rečnik može imati desetine hiljada termina. U **finansijskoj analizi**,

korelacione mere i ANOVA testovi identifikuju ekonomske indikatore koji utiču na tržišta ili predikciju rizika, omogućavajući brzu analizu velikog broja prediktora.

6. Wrapper metode

Wrapper metode predstavljaju model-orijentisan pristup izboru atributa u mašinskom učenju. Za razliku od filter metoda, koje procenjuju relevantnost atributa nezavisno od algoritma učenja, wrapper metode koriste algoritam mašinskog učenja kao sastavni deo procesa selekcije. Relevantnost podskupa atributa procenjuje se na osnovu performansi modela treniranog nad tim podskupom.

Osnovna prednost wrapper metoda ogleda se u njihovoj sposobnosti da uzmu u obzir interakcije između atributa i da pronađu podskup koji je optimalan za konkretan algoritam. Međutim, ova prednost dolazi po cenu značajno veće računske složenosti, zbog čega se wrapper metode najčešće primenjuju na skupovima podataka sa umerenim brojem atributa ili nakon preliminarne redukcije dimenzionalnosti pomoću filter metoda.



Ilustracija wrapper metode

6.1. Princip rada i opšti algoritam

Wrapper metode posmatraju izbor atributa kao problem pretrage kroz prostor svih mogućih podskupova atributa. Svaki kandidat-podskup se evaluira treniranjem mašinskog modela i merenjem njegove performanse pomoću izabrane evaluacione metrike. Algoritam zatim bira podskup koji daje najbolje rezultate prema tom kriterijumu.

Opšti algoritam wrapper metoda može se formalno predstaviti na sledeći način:

ULAZ:

- $D = \{X, L\}$ // skup podataka za treniranje sa n atributa
  - $X = \{f_1, f_2, f_3, \ldots, f_n\}$ // skup atributa
  - L // vektor oznaka (ciljna promenljiva)
- X' // inicijalni podskup atributa ($X' \subset X$ ili $X' = \{\emptyset\}$)
- A // algoritam mašinskog učenja
- θ // kriterijum zaustavljanja

IZLAZ:

- X'_{opt} // optimalan podskup atributa

Algoritam:

Početak:
Inicijalizacija:

- $X_{opt} = X'$
- $\varphi_{opt} = E(X', A)$ // evaluacija podskupa pomoću algoritma A

Ponavljaj:

- $X_\gamma = generate(X)$ // generisanje kandidata za podskup atributa
- $\varphi = E(X_\gamma, A)$ // evaluacija performansi modela treniranog na X_γ
- **Ako** ($\varphi > \varphi_{opt}$):
  - $\varphi_{opt} = \varphi$
  - $X'_{opt} = X_\gamma$

Sve dok kriterijum θ nije dostignut;
Vrati X'_{opt};
Kraj;
Ključni elementi algoritma:
Funkcija evaluacije E(·) meri performanse modela treniranog nad određenim podskupom atributa, koristeći algoritam A i odgovarajuću metriku (npr. tačnost, ROC-AUC, RMSE). Funkcija generate(·) definiše strategiju pretrage prostora atributa (sekvencijalno dodavanje ili uklanjanje, iscrpna pretraga, heurističke metode). Kriterijum zaustavljanja θ određuje kraj procesa (maksimalan broj atributa, prestanak poboljšanja performansi ili ograničenje iteracija).

Za razliku od filter metoda, gde se svaki atribut evaluira samo jednom, wrapper metode treniraju model u svakoj iteraciji, što značajno povećava računsku složenost.

6.2. Podela wrapper metoda

Wrapper metode se najčešće dele prema strategiji pretrage prostora atributa. Najzastupljenije su **sekvencijalne metode**, dok se ređe koriste **iscrpne** i **hibridne** strategije.

Sekvencijalne metode predstavljaju kompromis između kvaliteta rešenja i računske efikasnosti, dok iscrpne metode garantuju optimalno rešenje, ali su praktično neizvodljive za veći broj atributa.

6.3. Forward selekcija

Forward selekcija započinje proces izbora atributa sa praznim skupom. U svakoj iteraciji dodaje se onaj atribut koji, u kombinaciji sa već izabranim atributima, najviše poboljšava performanse modela. Proces se ponavlja dok se ne dostigne kriterijum zaustavljanja. Ova metoda je pogodna kada se pretpostavlja da mali broj atributa ima dominantan uticaj na ciljnu promenljivu. Njena glavna prednost je relativno manja računska složenost u poređenju sa drugim wrapper pristupima. Međutim, jednom dodati atribut ne može biti uklonjen, što može dovesti do suboptimalnog rešenja ako se kasnije pokaže da njegov doprinos opada.

6.4. Backward eliminacija

Backward eliminacija započinje proces sa kompletnim skupom atributa. U svakoj iteraciji uklanja se atribut čije uklanjanje najmanje utiče na performanse modela ili čak dovodi do njihovog poboljšanja. Proces se nastavlja dok dalje uklanjanje atributa ne pogoršava rezultate. Ova metoda je pogodna u situacijama gde se sumnja da skup podataka sadrži veliki broj irelevantnih ili redundantnih atributa. Međutim, početno treniranje modela sa svim atributima čini ovaj pristup računski zahtevnim kod visokodimenzionalnih problema.

6.5. Rekurzivna eliminacija atributa

Rekurzivna eliminacija atributa (eng. Recursive Feature Elimination: RFE) predstavlja strukturisan oblik backward eliminacije koji se odvija u više koraka. Algoritam funkcioniše na sledeći način: model se trenira na celom skupu atributa, atributima se dodeljuju ocene važnosti (npr. apsolutne vrednosti koeficijenata kod linearnih modela), zatim se uklanja unapred definisan broj najmanje važnih atributa (parametar step), a postupak se ponavlja sve dok se ne dostigne željeni broj atributa k.

Ključna prednost RFE u odnosu na običnu backward eliminaciju jeste efikasnost, uklanjanjem više atributa odjednom (step > 1) značajno se smanjuje broj iteracija i ukupno vreme izvršavanja, uz

minimalan gubitak kvaliteta selekcije. RFE zahteva da model poseduje eksplicitnu meru važnosti atributa, kao što su koeficijenti logističke regresije, težine SVM-a ili Gini importance stabala odlučivanja

6.6. Rekurzivna eliminacija atributa sa unakrsnom validacijom (RFECV)

RFECV (eng. Recursive Feature Elimination with Cross-Validation) predstavlja proširenje RFE metode koje automatski određuje optimalan broj atributa k, eliminišući potrebu za ručnim podešavanjem ovog parametra.

Algoritam primenjuje RFE za svaki mogući broj atributa od minimuma do maksimuma, pri čemu se za svaki k performanse modela procenjuju kroz unakrsnu validaciju. Broj atributa koji daje najbolji prosečni CV skor automatski se bira kao optimalan.

Prednosti RFECV u odnosu na RFE su: eliminacija subjektivnog izbora broja atributa i robustnija procena performansi zahvaljujući unakrsnoj validaciji. Nedostatak je povećana računska složenost, jer se model trenira višestruko za svaki kandidatni broj atributa. Zbog toga se u praksi preporučuje kombinovanje sa prethodnom redukcijom dimenzionalnosti filter metodama.

6.7. Iscrpna i proširena pretraga

Iscrpna pretraga razmatra sve moguće kombinacije atributa u unapred definisanom opsegu. Iako garantuje optimalno rešenje, eksponencijalna složenost ograničava njenu primenu na veoma male skupove atributa.

Kako bi se ublažila ograničenja osnovnih sekvencijalnih metoda, razvijene su proširene strategije kao što su LRS (Plus-L, Minus-R) i floating metode. Ovi pristupi omogućavaju povremeno dodavanje i uklanjanje atributa, čime se smanjuje rizik od lokalno optimalnih rešenja, ali uz cenu dodatne složenosti i potrebe za podešavanjem parametara.

6.8. Prednosti i nedostaci wrapper metoda

Prednosti:
Wrapper metode omogućavaju detekciju interakcija između atributa i pronalaženje podskupa koji je optimalan za konkretan algoritam mašinskog učenja. Zbog toga često dovode do boljih prediktivnih performansi u poređenju sa filter metodama.

Nedostaci:
Glavni nedostatak wrapper metoda jeste visoka računska složenost, jer se model trenira u svakoj iteraciji pretrage. Pored toga, postoji povećan rizik od preprilagođavanja, naročito kada se evaluacija ne sprovodi pomoću unakrsne validacije. Wrapper metode su slabo skalabilne na skupove podataka sa velikim brojem atributa i uzoraka.

Poseban izazov predstavljaju visokodimenzionalni skupovi podataka, gde forward selekcija ima kvadratnu složenost $O(n^2)$ po broju atributa, u svakoj iteraciji evaluiraju se sve preostale kombinacije. Iz tog razloga se u praksi wrapper metode gotovo uvek primenjuju nakon preliminarne redukcije dimenzionalnosti filter metodama, a ne direktno na originalnom skupu atributa.

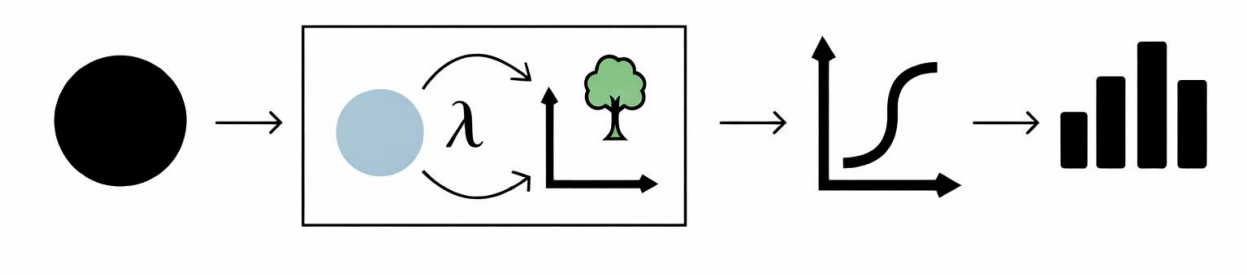
6.9. Praktična primena wrapper metoda

Wrapper metode se najčešće primenjuju kada je broj atributa umeren i kada su performanse modela prioritet. Često se koriste u problemima biomedicine, finansijske analize i inženjerskih sistema, gde je važno identifikovati podskup atributa koji je optimalan za konkretan model. U praksi se gotovo uvek kombinuju sa filter metodama, koje se koriste za inicijalno smanjenje dimenzionalnosti pre primene wrapper pristupa.

7. Embedded metode

Embedded metode predstavljaju treću osnovnu kategoriju tehnika izbora atributa u mašinskom učenju. Njihova ključna karakteristika jeste da se proces selekcije atributa odvija **tokom samog treniranja modela**, kao sastavni deo algoritma učenja. Na taj način embedded metode objedinjuju prednosti filter i wrapper pristupa: efikasnost filter metoda i sposobnost wrapper metoda da uzmu u obzir interakcije između atributa i modela.

Za razliku od filter metoda, koje se oslanjaju isključivo na statističke mere nezavisne od modela, i wrapper metoda, koje zahtevaju višestruko treniranje modela nad različitim podskupovima atributa, embedded metode integrisano optimizuju i strukturu modela i izbor atributa u jednom postupku. Ovo ih čini računski efikasnijim od wrapper metoda i često preciznijim od filter metoda.



Ilustracija embedded metode

7.1. Princip rada i opšti embedded algoritam

Kod embedded metoda, algoritam mašinskog učenja poseduje ugrađen mehanizam za procenu značaja atributa. Tokom procesa treniranja, model automatski favorizuje relevantne atribute, dok se uticaj irelevantnih ili redundantnih atributa smanjuje ili u potpunosti eliminiše.

Opšti algoritam embedded metoda može se formalno opisati na sledeći način:

ULAZ:

- $D = \{X, L\}$ – skup podataka za treniranje
  - $X = \{f_1, f_2, ..., f_n\}$ – skup atributa
  - L – ciljne oznake
- X' – inicijalni podskup atributa ($X' \subseteq X$ ili $X' = \emptyset$)
- θ – kriterijum zaustavljanja

IZLAZ:

- X'_{opt} – optimalan podskup atributa

Algoritam:

1. Inicijalizuj:
   a. $X_{opt} = X'$
   b. $\varphi_{opt} = E(X', I_m)$ – evaluacija korišćenjem nezavisne mere
   c. $\delta_{opt} = E(X', A)$ – evaluacija korišćenjem algoritma učenja A
   d. $C_0 = |X'|$ – kardinalnost početnog skupa
2. Iterativno:
   a. Generiši nove kandidate X^g proširivanjem ili modifikacijom X_{opt}
   b. Izračunaj $\varphi = E(X^g, I_m)$
   c. Ako je $\varphi > \varphi_{opt}$, ažuriraj X_{opt}
   d. Evaluiraj X_{opt} pomoću algoritma A i dobij δ
3. Ako je $\delta > \delta_{opt}$, prihvati novi skup atributa; u suprotnom, prekini postupak.
4. Vrati X'_{opt}.

Za razliku od filter metoda, embedded pristup kombinuje **statističku evaluaciju** i **model-specifičnu evaluaciju**, čime se postiže balans između tačnosti i računarske složenosti.

7.2. Regularizacija kao embedded metoda

Jedan od najznačajnijih primera embedded metoda jeste **regularizacija** u linearnim i logističkim modelima. Regularizacija uvodi kaznenu komponentu u funkciju gubitka sa ciljem da ograniči kompleksnost modela i smanji uticaj irelevantnih atributa.

L1 regularizacija (Lasso)

L1 regularizacija ima posebno značajnu ulogu u izboru atributa jer može dovesti do **potpunog eliminisanja atributa**. Funkcija gubitka kod Lasso regresije definisana je kao:

$$L = MSE + \lambda \sum_{j=1}^{n} |w_j|$$

gde su: *MSE* – srednja kvadratna greška w_j – težina (koeficijent) atributa j, λ – parametar regularizacije.

Povećanjem vrednosti λ, sve veći broj koeficijenata biva sveden na nulu, čime se odgovarajući atributi automatski uklanjaju iz modela. Na taj način Lasso istovremeno vrši **učenje modela i selekciju atributa**.

L2 regularizacija (Ridge)

Za razliku od L1 regularizacije, L2 regularizacija koristi kvadrat koeficijenata:

$$L = MSE + \lambda \sum_{j=1}^{n} w_j^2$$

Iako smanjuje vrednosti koeficijenata, L2 regularizacija ih ne svodi na nulu i stoga **nije pogodna za eksplicitnu selekciju atributa**, već prvenstveno za stabilizaciju modela.

Elastic Net

Elastic Net kombinuje L1 i L2 regularizaciju:

$$L = MSE + \lambda_1 \sum |w_j| + \lambda_2 \sum w_j^2$$

Ovaj pristup je naročito koristan kada postoje **jako korelisani atributi**, jer omogućava selekciju uz očuvanje stabilnosti modela.

7.3. Embedded metode zasnovane na stablima odlučivanja

Algoritmi zasnovani na stablima, kao što su **Decision Trees**, **Random Forests** i **Gradient Boosting**, predstavljaju još jednu važnu grupu embedded metoda.

Tokom izgradnje stabla, u svakom čvoru bira se atribut koji najviše smanjuje meru nečistoće (eng. *impurity*), kao što su:

- Gini indeks,
- entropija (information gain),
- varijansa (kod regresije).

Značaj atributa se računa kao **ukupno smanjenje nečistoće** koje taj atribut ostvaruje kroz sve čvorove stabla. Atributi koji se češće koriste i koji dovode do većeg smanjenja nečistoće dobijaju veću vrednost značaja.

Prednost ovog pristupa je što prirodno modeluje **nelinearne odnose** i interakcije između atributa, bez potrebe za dodatnom obradom.

7.3.1. Random Forest importance

Random Forest rešava problem nestabilnosti pojedinačnog stabla agregiranjem mera važnosti kroz veliki broj stabala (ansambl). Svako stablo se gradi na bootstrap uzorku podataka i nasumičnom podskupu atributa, a finalna mera važnosti atributa dobija se usrednjavanjem Gini importance kroz sva stabla ansambla.

Zahvaljujući agregaciji, Random Forest pruža znatno stabilniju i pouzdaniju meru važnosti atributa u poređenju sa pojedinačnim stablom. Posebno je efikasan u situacijama sa velikim brojem atributa i prisutnim šumom, jer nasumični odabir atributa pri izgradnji svakog stabla smanjuje korelaciju između stabala i povećava raznovrsnost ansambla. Atributi koji konzistentno dobijaju visoke ocene važnosti kroz ceo ansambl smatraju se pouzdano relevantnim.

7.3.2. Gradient Boosting i XGBoost importance

Gradient Boosting metode (XGBoost, LightGBM, CatBoost) grade stabla sekvencijalno, pri čemu svako naredno stablo ispravlja greške prethodnog minimizujući gradijent funkcije gubitka. Mera važnosti atributa koja se koristi u ovim metodama naziva se **gain,** ukupno poboljšanje funkcije gubitka koje atribut ostvaruje kroz sve splitove u svim stablima.

Za razliku od Random Foresta koji gradi stabla paralelno i nezavisno, boosting gradi stabla zavisno jedno od drugog, što može dovesti do drugačijeg rangiranja atributa. Atributi koji su korisni za ispravljanje specifičnih grešaka modela mogu dobiti veću važnost nego što bi dobili u Random Forest pristupu. XGBoost dodatno podržava L1 i L2 regularizaciju koeficijenata stabala, što ga čini hibridnom embedded metodom koja kombinuje selekciju zasnovanu na stablima sa regularizacionim pristupom.

7.4. Embedded metode u logističkoj regresiji i SVM-u

Embedded selekcija atributa primenljiva je i u klasifikacionim modelima.

Kod **logističke regresije sa L1 regularizacijom**, funkcija gubitka ima oblik:

$$L = -\log(\text{likelihood}) + \lambda \sum |w_j|$$

Ovaj pristup omogućava automatsku eliminaciju atributa u binarnim i višeklasnim problemima.

Slično tome, **SVM sa L1 penalom** (L1-SVM) omogućava selekciju atributa u visokodimenzionalnim prostorima, iako zahteva veću računsku složenost.

7.5. Prednosti i ograničenja embedded metoda

Prednost jeste i to što embedded metode uzimaju u obzir interakcije između atributa tokom procesa učenja, čime se postiže bolja selekcija od jednostavnih statističkih mera. Dodatno, ovaj pristup smanjuje rizik od preprilagođavanja jer regularizacione tehnike inherentno kontrolišu kompleksnost modela.

Sa druge strane, postoje i izvesna ograničenja. Embedded metode su zavisne od konkretnog algoritma učenja, što znači da različiti algoritmi mogu dati različite skupove odabranih atributa. Kod kompleksnih modela, posebno dubokih neuronskih mreža ili ansambla, interpretacija značaja atributa može biti otežana i manje transparentna. Takođe, uspeh embedded metoda u velikoj meri zavisi od pažljivog podešavanja hiperparametara regularizacije, što zahteva dodatno vreme i ekspertizu.

Dodatno ograničenje embedded metoda zasnovanih na stablima jeste osetljivost mere važnosti na korelisane atribute, kada su dva atributa visoko korelisana, njihova važnost se može rasporediti između njih na nepredvidiv način, što može podceniti značaj oba. U takim situacijama preporučuje se prethodna primena korelacionog filtera.

7.6. Praktična primena embedded metoda

Embedded metode nalaze široku primenu u različitim oblastima mašinskog učenja i analize podataka. U finansijskoj analizi koriste se za selekciju najrelevantnijih ekonomskih indikatora koji utiču na kretanje tržišta ili procenu kreditnog rizika. Bioinformatika predstavlja još jedno važno polje primene, gde se embedded metode koriste za odabir gena koji su značajni za određene bolesti ili biološke procese, što je od ključnog značaja u genomskim istraživanjima. U obradi teksta,

regularizovani linearni modeli omogućavaju efikasnu selekciju relevantnih karakteristika iz velikih korpusa tekstualnih dokumenata. Sistemi preporuke i analiza velikih skupova podataka takođe se oslanjają na embedded metode kako bi izdvojili najznačajnije atribute koji određuju korisničke preferencije ili obrasce ponašanja.

Zahvaljujući povoljnom balansu između tačnosti i efikasnosti, embedded metode se često smatraju najpraktičnijim izborom u realnim sistemima mašinskog učenja.

8. Uslovi primene i poređenje metoda

Izbor odgovarajuće metode selekcije atributa zavisi od nekoliko ključnih faktora: dimenzionalnosti podataka, dostupnih računskih resursa, prirode problema i zahteva za interpretabilnošću modela.

Filter metode su najpogodnije kada je broj atributa veoma veliki (nekoliko hiljada ili više), kada su računski resursi ograničeni, ili kada je potrebna brza preliminarna analiza podataka. One predstavljaju odličan izbor za inicijalno smanjenje dimenzionalnosti pre primene sofisticiranijih tehnika. Filter metode su takođe pogodne kada je potrebna nezavisnost od algoritma učenja, što omogućava fleksibilnost u kasnijoj primeni različitih modela. Međutim, njihova primena je manje efikasna kada postoje složene interakcije između atributa ili kada je optimizacija specifična za konkretan algoritam neophodna za postizanje dobrih rezultata.

Wrapper metode se preporučuju kada je prioritet postizanje maksimalnih performansi modela i kada je broj atributa umeren (obično do nekoliko stotina). One su posebno korisne kada su interakcije između atributa značajne za problem i kada postoji dovoljna računska moć za višestruko treniranje modela. Wrapper pristup je pogodan i u situacijama gde je model već odabran i potrebno je pronaći optimalan podskup atributa baš za taj model. Međutim, ovaj pristup nije preporučljiv za visokodimenzionalne probleme sa hiljadama atributa, za situacije sa ekstremno ograničenim resursima, ili kada je potrebna brza analiza.

Embedded metode predstavljaju srednje rešenje koje balansira prednosti i nedostatke prethodna dva pristupa. Najpogodnije su kada je potreban kompromis između tačnosti i efikasnosti, kada algoritam učenja prirodno podržava mehanizme selekcije (kao što su regularizovani modeli ili stabla odlučivanja), i kada je važno uzeti u obzir interakcije između atributa bez značajnog povećanja računske složenosti. Embedded metode su idealne za primene u realnom vremenu gde je potrebno periodično ažuriranje modela, kao i za probleme srednje dimenzionalnosti gde filter metode mogu biti nedovoljno precizne, a wrapper metode previše skupe.

Direktno **poređenje performansi** tri pristupa pokazuje sledeće karakteristike. Sa aspekta **računske složenosti**, filter metode imaju linearnu složenost O(n), embedded metode zavise od

algoritma ali su generalno efikasne, dok wrapper metode imaju eksponencijalnu složenost u najgorem slučaju. Što se tiče **tačnosti**, wrapper metode obično postižu najbolje rezultate jer direktno optimizuju performanse modela, embedded metode daju vrlo dobre rezultate uz manju složenost, a filter metode mogu biti suboptimalne ali pružaju solidnu osnovu. **Sposobnost detekcije interakcija** je najizraženija kod wrapper metoda koje modeluju sve interakcije, umereno prisutna kod embedded metoda koje implicitno uzimaju neke interakcije, dok filter metode uglavnom ne detektuju interakcije osim kod naprednih multivarijantnih pristupa. **Rizik od preprilagođavanja** je najmanji kod filter metoda jer ne koriste algoritam učenja, umeren kod embedded metoda koje koriste regularizaciju, i najveći kod wrapper metoda ako se ne primenjuje pažljiva validacija.

Poseban slučaj koji utiče na izbor metode selekcije atributa jeste nebalansiranost klasa. Kada jedna klasa značajno dominira nad drugom, standardna metrika tačnosti (accuracy), model koji uvek predviđa većinsku klasu postiže visoku tačnost bez ikakve stvarne diskriminativne moći. U takvim scenarijima preporučuje se korišćenje ROC-AUC metrike, koja meri sposobnost modela da razlikuje klase nezavisno od praga odluke, balanced accuracy koja uzima prosek senzitivnosti po klasama, ili average precision koja je posebno korisna kada je manjinska klasa od primarnog interesa. Izbor evaluacione metrike direktno utiče na rangiranje atributa u supervised filter metodama i na kriterijum optimizacije u wrapper i embedded pristupima.

U praksi se često primenjuje **hibridni pristup** koji kombinuje prednosti različitih metoda. Uobičajena praksa je prethodna primena filter metoda (npr. uklanjanje atributa sa visokim procentom nedostajućih vrednosti, variance threshold i korelacioni filter) kojima se dimenzionalnost redukuje na upravljiv broj kandidata, nakon čega se wrapper metoda primenjuje na preostalom skupu.

9. Dobre i loše strane primene feature selection-a

Primena tehnika selekcije atributa donosi brojne koristi, ali nosi i određene rizike koje je potrebno razumeti i uzeti u obzir.

Među glavnim **prednostima** primene selekcije atributa ističe se smanjenje dimenzionalnosti podataka, što direktno vodi ka bržem treniranju i izvršavanju modela. Ovo je posebno značajno u primenama koje zahtevaju predikcije u realnom vremenu ili obradu velikih količina podataka. Poboljšanje performansi modela često je vidljivo kroz veću tačnost, preciznost i sposobnost generalizacije na nove podatke. Eliminacijom irelevantnih i redundantnih atributa smanjuje se šum u podacima, što omogućava modelu da lakše identifikuje relevantne obrasce. Još jedna važna prednost jeste smanjenje rizika od preprilagođavanja, jer jednostavniji modeli sa manjim brojem atributa imaju manju tendenciju da memorišu specifičnosti trening skupa umesto da uče opšte zakonitosti. Interpretabilnost modela se značajno poboljšava kada je broj atributa manji, što je od

kritičnog značaja u domenima kao što su medicina, finansije i pravo, gde je potrebno razumeti i obrazložiti odluke modela.

Međutim, postoje i **potencijalni nedostaci i rizici**. Gubitak informacija je neizbežan kada se atributi eliminišu, i uvek postoji mogućnost da se uklone atributi koji sadrže korisne, ali ne očigledne informacije. Ovaj rizik je posebno izražen kod filter metoda koje ne uzimaju u obzir interakcije između atributa. Dodatna računska složenost može nastati kada se primenjuju wrapper ili napredne embedded metode, što može produljiti vreme potrebno za razvoj modela. Rizik od pogrešne selekcije postoji naročito kada su skupovi podataka mali, jer statističke mere mogu biti nepouzdane i dovesti do izbora atributa koji dobro funkcionišu na trening podacima ali ne generalizuju. Zavisnost od domena znači da selekcija atributa zahteva razumevanje problema i često uključuje subjektivne odluke o kriterijumima i pragovima. **Stabilnost** selekcije može biti problematična jer male promene u podacima mogu dovesti do značajno različitih skupova odabranih atributa, što otežava reproducibilnost rezultata i povećava nesigurnost u praktičnoj primeni. Stabilnost selekcije atributa odnosi se na konzistentnost odabranih atributa pri malim promenama u skupu podataka ili pri različitim podeljima trening/test skupa. Nestabilna metoda može selektovati potpuno različite skupove atributa pri malim perturbacijama podataka, što otežava interpretaciju i reproducibilnost rezultata. Wrapper metode su generalno manje stabilne od filter metoda zbog osetljivosti na specifičnosti trening skupa, dok embedded metode sa regularizacijom pokazuju umerenu stabilnost. Procena stabilnosti preporučuje se kao dodatni kriterijum evaluacije pored same prediktivne performanse.

Važno je naglasiti da selekcija atributa **nije uvek neophodna niti poželjna**. U situacijama gde svi atributi nose značajne informacije, gde je broj atributa mali u odnosu na broj uzoraka, ili gde algoritmi kao što su regularizovane neuronske mreže ili ansambli stabala sami efikasno upravljaju velikim brojem atributa, eksplicitna selekcija može biti nepotrebna. Takođe, u domenima gde je interpretabilnost manje važna od čiste prediktivne moći, korišćenje svih dostupnih atributa može dati bolje rezultate.

10. Zaključak

Izbor atributa predstavlja fundamentalan korak u procesu izgradnje efikasnih i pouzdanih modela mašinskog učenja. Kao što je prikazano kroz ovaj rad, visokodimenzionalnost podataka dovodi do niza praktičnih i teoretskih izazova, uključujući prokletstvo dimenzionalnosti, povećanu računsku složenost, rizik od preprilagođavanja i otežanu interpretabilnost modela.

Tri osnovne kategorije metoda selekcije atributa – filter, wrapper i embedded – nude različite pristupe rešavanju ovog problema. Filter metode donose brzinu i skalabilnost, wrapper metode obezbeđuju optimalnost za konkretan algoritam, dok embedded metode nude balans između

efikasnosti i tačnosti. Nijedna od ovih metoda nije univerzalno najbolja; izbor zavisi od specifičnosti problema, dostupnih resursa i prioriteta u razvoju modela. Posebno je važno prilagoditi evaluacionu metriku karakteristikama problema, u slučaju nebalansiranih skupova podataka, oslanjanje na metriku tačnosti može dovesti do pogrešnih zaključaka o kvalitetu selekcije.

Praktična primena tehnika selekcije atributa zahteva razumevanje ne samo algoritama, već i prirode podataka i problema koji se rešava. Domensko znanje, pažljiva validacija i kritički pristup rezultatima ključni su za uspešnu primenu. Hibridni pristupi koji kombinuju različite tehnike često daju najbolje rezultate, jer omogućavaju iskorišćavanje prednosti svakog pristupa dok ublažavaju njihova ograničenja.

Selekcija atributa nije samo tehnika za optimizaciju performansi, već predstavlja važan alat za razumevanje podataka i problema. Kroz proces identifikacije najznačajnijih atributa, istraživači i praktičari dobijaju uvid u to koje karakteristike zaista utiču na predikciju, što može voditi ka novim naučnim otkrićima i boljim poslovnim odlukama.

Sa kontinuiranim razvojem mašinskog učenja i rastom količine dostupnih podataka, uloga selekcije atributa će nastaviti da bude od kritičnog značaja. Buduća istraživanja treba da se fokusiraju na razvoj metoda koje su robusnijih, efikasnijih i primenljivijih na ekstremno visokodimenzionalne podatke, kao i na integrisanje domenskog znanja u automatizirane sisteme za selekciju atributa.