Razlaga klasifikatorjev na podlagi podkonceptov

Nejc Mušič

Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

Mentor: prof. dr. Marko Robnik Šikonja

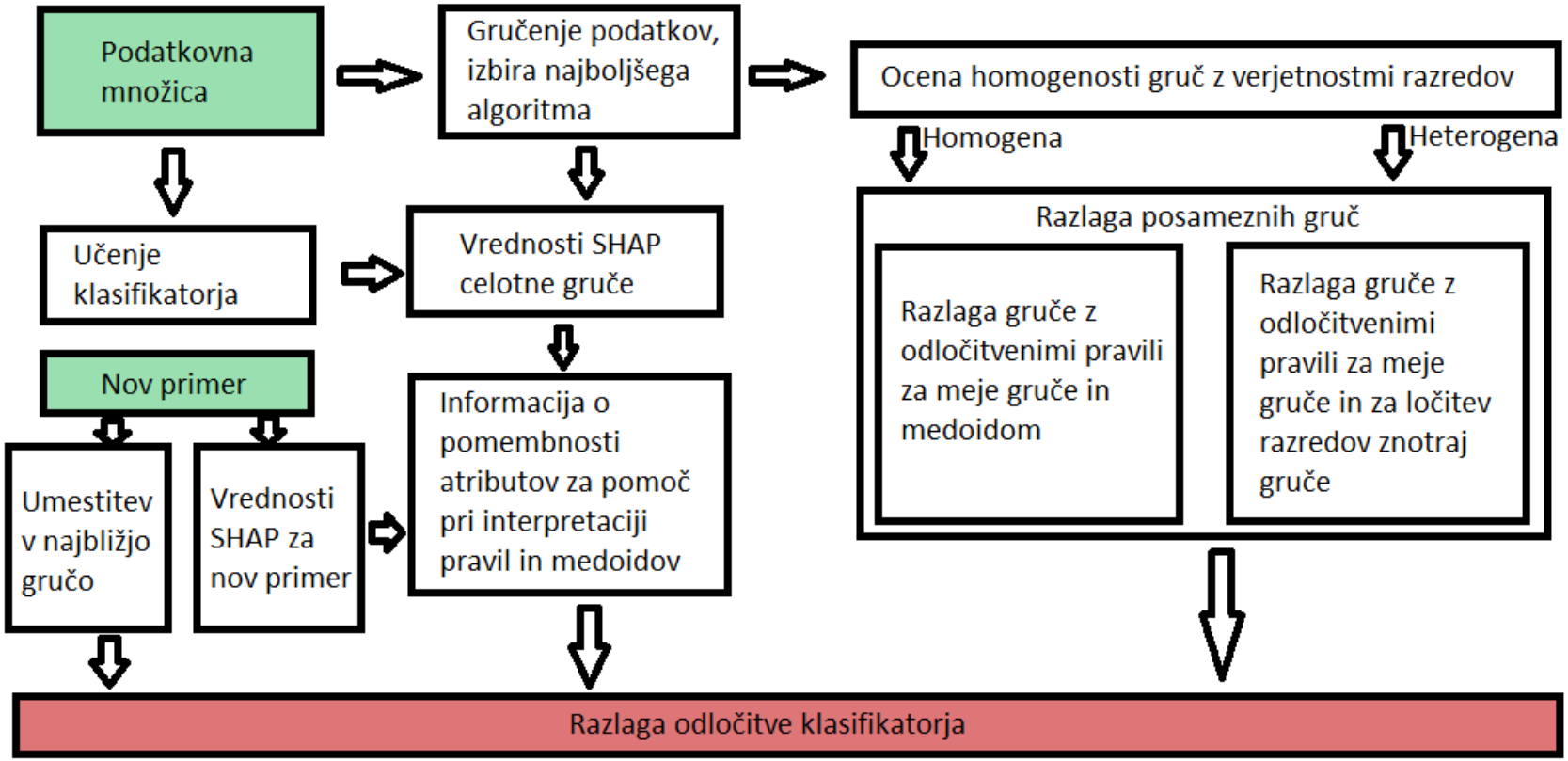
10. september 2023

Pregled predstavitve

Motivacija za interpretacijo klasifikatorjev

- Klasifikatorji imajo dobro točnost pri napovedovanju
- Zaradi kompleksnosti in narave modelov ne razumemo vzrokov za sprejeto odločitev
- Kritične odločitve potrebujejo razlago
- Ljudem razlaga modelov omogoči zaupati v odločitve

Postopek razlage



Algoritmi za gručenje

Primerjali smo:

- MDEC (Multidiversified Ensemble Clustering) se uporablja za gručenje visoko dimenzionalnih prostorov
- K-MEANS dobro gruči podatke sferičnih oblik
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) temelji na osnovi gostote podatkovnih točk v prostoru
- HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) izvaja algoritem DBSCAN pri različnih vrednostih epsilon in najde gručenje, ki zagotavlja najboljšo stabilnost

Hevristike za oceno gručenj

- Uporabimo za oceno gručenja in avtomatski izbor parametrov pri algoritmih gručenja
- Koeficient silhuete oceni, kako dobro so točke znotraj iste gruče povezane in kako dobro so razmejene od drugih gruč
- Indeks DBCV (Density-Based Clustering Validation) temelji na gostoti gruč in je primeren tudi za nepravilne oblike

Klasifikatorja

- XGBoost (Extreme Gradient Boosting) kombinira moč odločitvenih dreves in tehnike gradientnega spusta
- Logistična regresija napoveduje verjetnosti, da bo določen vhodni primer spadal v enega izmed razredov

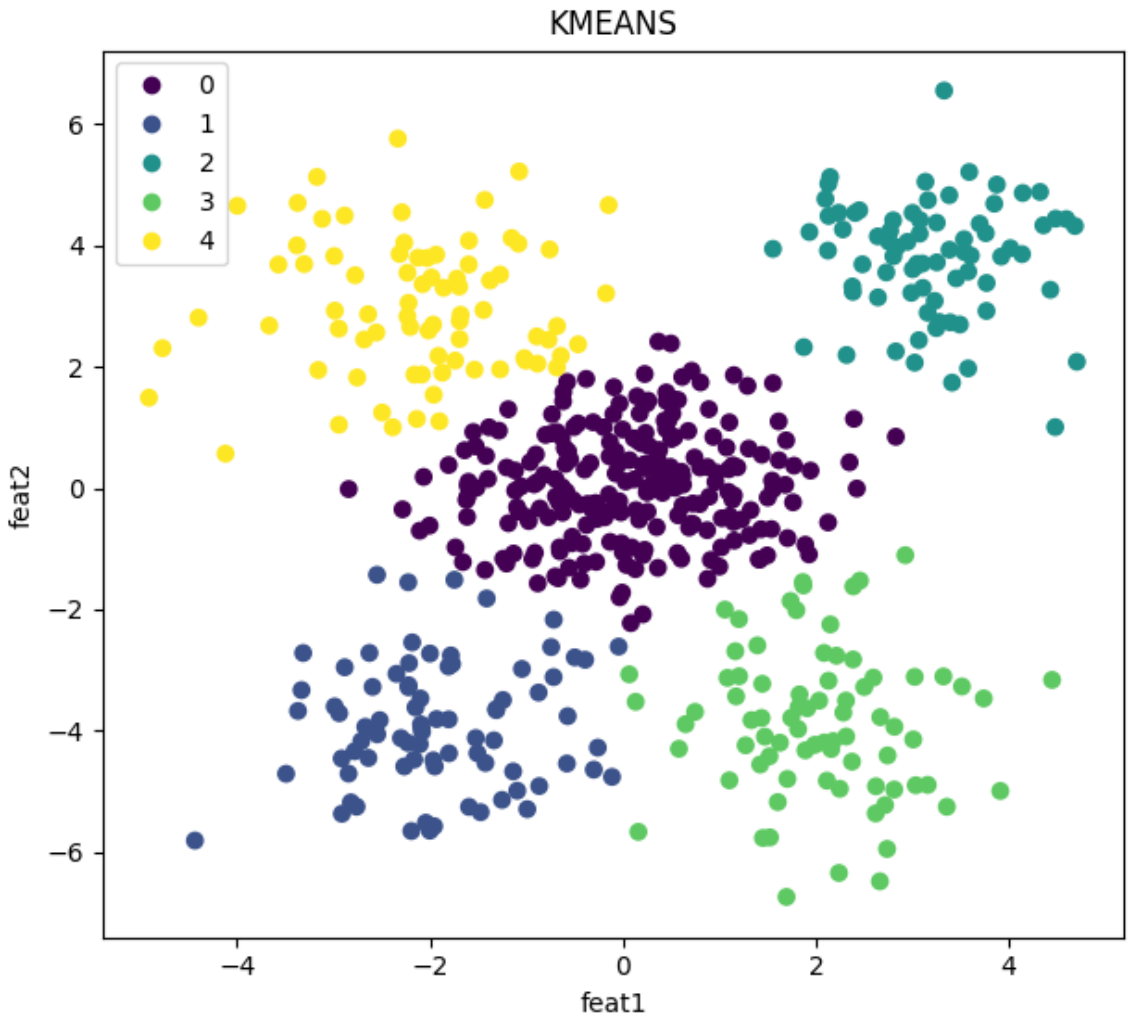
Metode razlage

- Odločitvena pravila uporabimo, saj so lahko razložljiva z obliko "IF conditions THEN response"
- Medoide uporabimo za opis homogenih gruč, saj gre za tipičen primer in ponuja enostaven opis
- Vrednosti SHAP uporabimo pri interpretaciji medoidov in odločitvenih pravil v visokih dimenzijah, saj razložijo pomembnost atributov

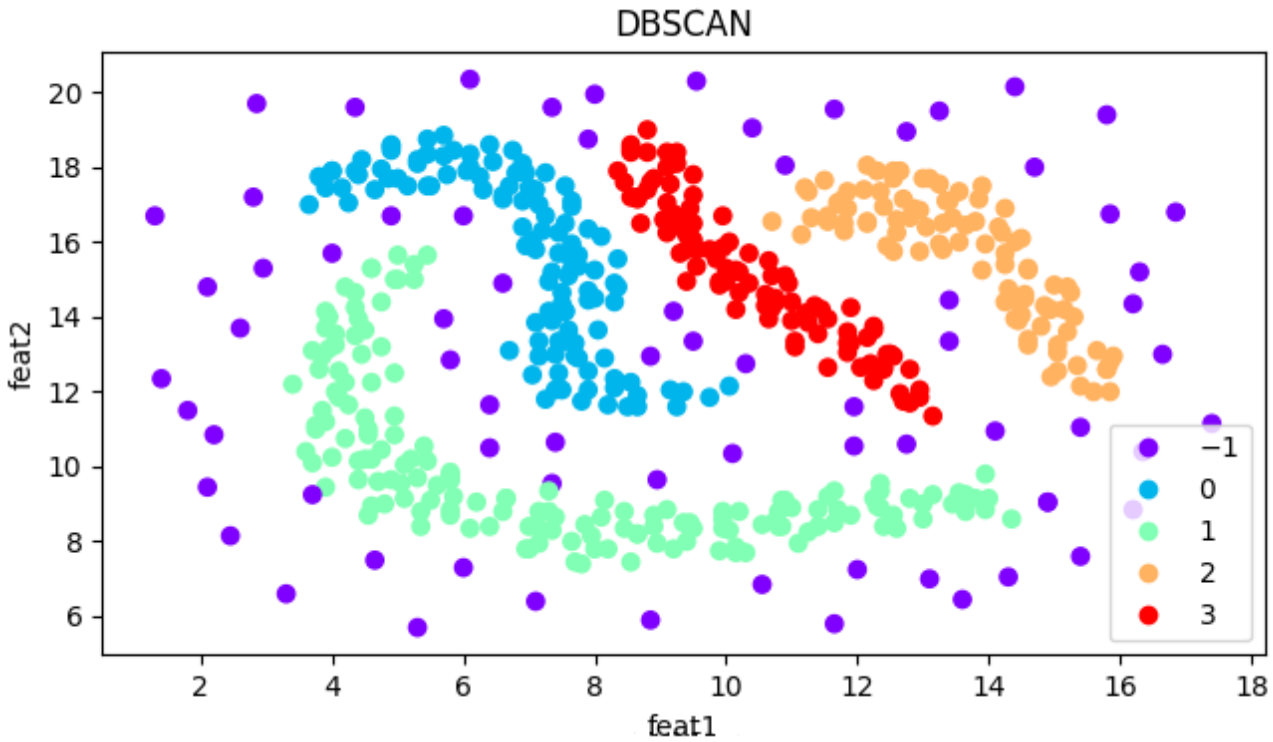
Zmanjšanje dimenzionalnosti

- Tehniko t-SNE uporabimo za vizualizacijo visoko dimenzionalnih podatkov v nižjih dimenzijah
- Nesistematično preizkusimo tudi tehniko UMAP (Uniform Manifold Approximation and Projection) in PCA (Principal Component Analysis), a t-SNE vrne bolje ločene gruče
- Tehniko t-SNE uporabimo tudi za predprocesiranje visoko dimenzionalnih podatkov pri gručenju z algoritmi K-MEANS, DBSCAN in HDBSCAN

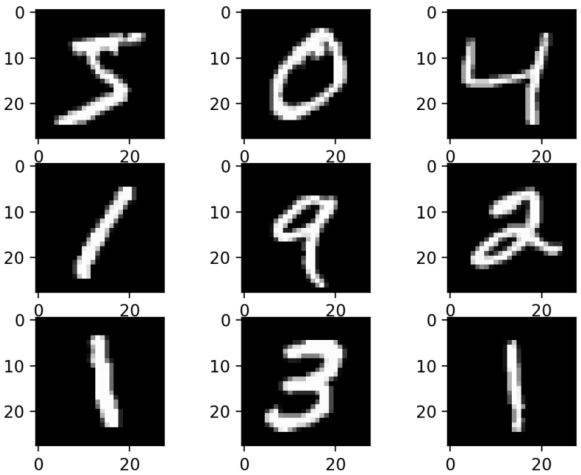
Podatkovna množica Krožne gruče

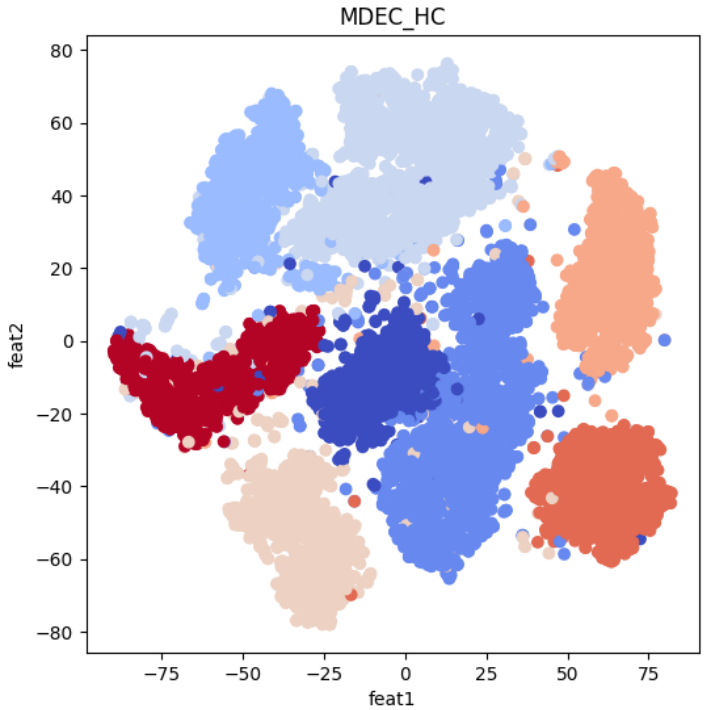


Podatkovna množica Trakovi 4-3



Podatkovna množica MNIST





Pregled rezultatov algoritmov gručenja

Podatkovna množica	Najboljši algoritem	Smiselni rezultati
Krožne gruče	K-MEANS	HDBSCAN
Trakovi 4-3	DBSCAN	HDBSCAN
MNIST	MDEC	HDBSCAN

Razlaga klasifikatorja na podatkovni množici KDD99

- Podatkovna množica KDD99 je široko uporabljena pri razvoju sistemom za detekcijo vdorov in računalniško varnost
- Algoritem HDBSCAN je v kombinaciji s tehniko t-SNE našel sedem skupin (5 homogenih in 2 nehomogeni)
- Dve gruči imata večinske normalne mrežne povezave, ostale imajo večinske škodljive mrežne povezave
- Po izračunu vseh metod razlage odločitev klasifikatorja interpretiramo na podlagi novega primera (normalna povezava)
- Ogledamo si vrednosti SHAP, ki pomagajo pri interpretaciji odločitvenih pravil in medoidov
- Opišemo, kako se medoida normalnih mrežnih povezav razlikujeta od medoidov škodljivih mrežnih povezav
- Podamo dodatno primerjavo med medoidoma, ki pripadata normalnim mrežnim povezavam

Zaključki

- Medoidi so smiselni pri homogenih gručah, v visokih dimenzijah nepregledni in neintuitivni (pomagamo si z vrednostmi SHAP)
- V visokih dimenzijah imamo več podmnožic odločitvenih pravil, ki dobro ločijo gruče. Nekatere značilke v pogojih niso pomembne.
- Struktura podatkov, ki jo pričakuje naša razlaga, je redka pri realnih podatkovnih bazah
- Postopek ni v celoti avtomatiziran in težek za interpretacijo (izbor algoritma gručenja, človeška interpretacija posameznih komponent razlage, visoka dimenzionalnost, težaven prikaz)
- Izboljšave: dodatna razlaga gruč (vizualizacija večih primerov), analiza atributov, avtomatsko preverjanje ali je razlaga aplikativna za podatkovno množico