

Nova metoda grupiranja podataka: algoritam koji se sam podešava i objašnjava rezultate

Novo istraživanje važan je korak prema razvoju algoritama koji uče iz podataka i samostalno podešavaju postavke, pa se mogu koristiti za označavanje (medicinskih) podataka.

ZAGREB, XY listopada/studenja 2025. - Mnogi algoritmi strojnog učenja rade bolje kad znamo kako pravilno podešiti, a za to trebamo dobro označene primjere. U mnogim područjima, na primjer medicini, označavanje velikih količina podataka je skupo i vremenski zahtjevno. Znanstvenici s Instituta Ruđer Bošković (IRB) razvili su metodu samovođenog (autonomnog) podešavanja hiperparametara za problem grupiranja podataka u potprostorima. Programi koji se sami podešavaju obično razvrstavaju podatke malo lošije, do oko 7% od najbolje ručno podešenih. Dodatno, tim je razvio novu metodu vizualizacije koja jasno pokazuje ključne obrasce u podacima, pa bolje razumijemo utjecaj pojedinih postavki. Zahvaljujući tome, postojeće algoritme grupiranja podataka možemo lakše koristiti za označavanje velikih zbirki podataka, posebno u medicini.

Da je riječ o važnim rezultatima potvrđuje i činjenica da je rad objavljen u otvorenom pristupu u uglednom časopisu *Pattern Recognition*, jednom od najboljih u svom području. Metoda ima velik potencijal primjene ondje gdje nemamo označene podatke, posebno u medicini, a svatko je može isprobati jer je kôd javno dostupan na GitHubu.

Vrlo pojednostavljeno, to bi bilo kao kada imamo veliku hrpu fotki i želimo ih posložiti u grupe (npr. lica, brojevi, predmeti) bez da ti itko unaprijed kaže što je na kojoj fotki. Algoritam je kao “pametni sortirni pomoćnik”, ali radi bolje kad su mu neke „kvake“ dobro namještene, a to su postavke koje obično podešavaju stručnjaci uz označene primjere. Ova nova metoda uči te postavke sama tako da isprobava različite mogućnosti, mjeri koliko dobro je posložila podatke i sužava izbor dok ne nađe najbolje. Usput napravi i “prosječnu sliku” svake grupe, pa je moguće vizualno provjeriti izgleda li rezultat smisleno. Tako dobijemo razvrstavanje koje je skoro jednako dobro kao ono koje bi netko dugo ručno podešavao, konkretno, razlika je oko sedam posto, ali bez sporog i skupog označavanja, što je jako korisno, posebno u medicini.

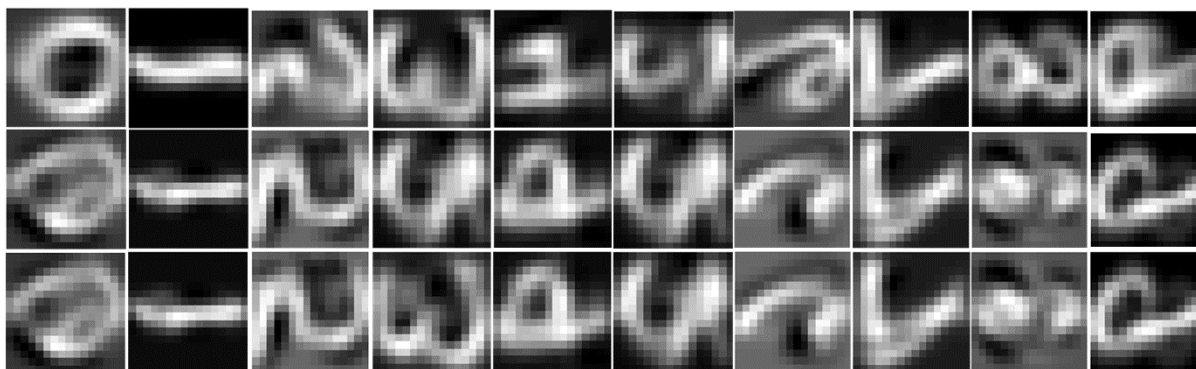
Pametno podešavanje: algoritam sužava izbor postavki dok ne postigne najbolje

Dr. sc. Ivica Kopriva s IRB-a razvio je metodu u kojoj se algoritam sam podešava, bez oslanjanja na označene podatke i to za postojeće linearne i nelinearne algoritme koji grupiraju slične podatke. Većina današnjih rješenja traži da se algoritam podešava na temelju označenih primjera i to iznova za svaki novi skup podataka, što često nije izvedivo jer takvi primjeri nedostaju. Jedan pristup pokušava podešiti algoritam prema unutarnjim pokazateljima kvalitete grupiranja, ali takvi pokazatelji su loš izbor za podatke koji se grupiraju prema modelu unije potprostora (engl. union-of-subspaces). Drugi pristup je meta-učenje, gdje se algoritam oslanja na iskustva sličnih skupova podataka, no u novim područjima tih iskustava nema. Nova metoda zaobilazi ova ograničenja.

Predložena metoda samovođenog podešavanja postavki oslanja se na jednostavne mjere uspješnosti grupiranja, poput točnosti ili normalizirane međusobne informacije. Te se mjere računaju na privremenim oznakama koje nastaju pomoću odabranog algoritma, i to između susjednih vrijednosti u unaprijed zadanom rasponu postavki. Pod pretpostavkom da se odabrana dosljedno mijenja (stabilno raste ili pada) moguće se pronaći uži raspon u kojem je rezultat najbolji. Taj se raspon zatim dijeli na nekoliko manjih dijelova i cijeli se postupak ponavlja do zadovoljenja uvjeta konvergencije.

Vizualizacija rezultata grupiranja podataka i interpretabilnost

Predložena metoda samovođenog podešavanja ključnih postavki ponekad može završiti pogrešnim izborom. Zato je razvijen vizualni prikaz rezultata. Za odabranu najbolju postavku, algoritam grupiranja podataka privremeno označi primjere i prema tim oznakama razvrstava podatke u skupine. Za svaku skupinu zatim izračunamo bazu koja razapinje odgovarajući podprostor, iz čega se može prikazati "prosječna" sliku objekta iz te skupine. Ako je označavanje loše, "prosječna" slika izgleda nejasno, što upućuje na to da početni raspon postavki nije bio dobar. U tom slučaju preporučuje se odabrati novi, drugačiji (udaljeniji) raspon vrijednosti i ponovo pokrenuti algoritam za samovođeni odabir postavki. Ovaj je postupak prikazan na slici 1.



Slika 1. Vizualizacija baza potprostora koji reprezentiraju znamenke USPS skupa podataka poredane od 0 do 9. Baze su procijenjene iz 50 slučajno odabranih uzoraka po svakoj skupini (znamenki). Prvi red: particije su dobivene od označenog podataka. Drugi red: particije su dobivene oracle verzijom algoritma (hiperparametri su podešeni koristeći označene podatke). Treći red: particije su dobivene samovođenom verzijom algoritma grupiranja podataka. Točnost grupiranja obje verzije algoritma je približno 85%.

Uz redefiniranje početnog prostora postavki, opisani postupak vizualizacije uspostavlja direktnu vezu između odabrane vrijednosti postavke i kvalitete slike dobivene vizualizacijom procijenjenog potprostora. Pri lošem odabiru postavki pseudo-labele su jako netočne pa uzorci koji čine istu particiju dolaze iz različitih skupina. U tom slučaju vizualizacija baze daje sliku koja ne asocira na odgovarajuću znamenku. Time opisani postupak vizualizacije doprinosi interpretabilnosti rezultata, što je nedostatak većine algoritama grupiranja podataka u potprostorima.

Potvrda metode i kvalitete

Razvijena metoda samovođenog podešavanja postavki demonstrirana je na šest slikovnih skupova podataka (koji predstavljaju objekte, znamenke i lica) pomoću pet linearnih i nelinearnih algoritama grupiranja podataka iz jednog pogleda, te tri skupa podataka pomoću tri linearna i nelinearna algoritma grupiranja podataka iz više pogleda.

Cijelo istraživanje objavljeno je u otvorenom pristupu u renomiranom znanstvenom časopisu *Pattern Recognition* 172 (2026) 112618, <https://doi.org/10.1016/j.patcog.2025.112618>. Čimbenik odjeka časopisa je 7.6 što ga svrstava među vodećih 9 posto u području elektrotehnike, odnosno među vodećih 13 posto u računalnim znanostima i umjetnoj inteligenciji. Procjenjuje se da predložena metoda ima potencijal za primjenu postojećih algoritama grupiranja podataka u potprostorima u novim područjima sa neoznačenim podacima sa naglaskom na medicinske skupove podataka. MATLAB kôd javno je dostupan na: <https://github.com/ikopriva/LFSGSC>.

Projekt je financiran sredstvima **Hrvatske zaklade za znanost** u sklopu natječaja HRZZ-IP-2022-10-6403.

KONTAKT SUGOVORNIKA IZJAVE ZA MEDIJE:

[dr. sc. Ivica Kopriva,](#)

E. Ivica.Kopriva@irb.hr

Za sve upite medijima dostupni smo na adresi pr@irbinstitute.com te broju 01 457 1269.

S poštovanjem,

Ured za odnose s javnošću IRB

Institut Ruđer Bošković

Bijenička cesta 54, Zagreb

E. pr@irbinstitute.com

T. 01 457 1269