Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko

FERI

Analiza vplivov kodirnikov na doseženo okrepitveno nagrado hibridnega algoritma NARM-XCS

TRIINTRIDESETA MEDNARODNA
ELEKTROTEHNIŠKA IN RAČUNALNIŠKA
KONFERENCA ERK 2024

Avtorji:
Damijan Novak (damijan.novak@um.si), Domen Verber (domen.verber@um.si),
Iztok Fister (iztok.fister@um.si) in Iztok Fister ml. (iztok.fister1@um.si)

26.9.2024

Vsebina predavanja

1. Motivacija in namen članka
2. Numerično rudarjenje asociativnih pravil (NARM)
3. Razširjen sistem na osnovi klasifikatorjev (XCS)
4. Hibrid NARM-XCS
5. Kodirniki
   - Binarni kodirnik
   - Kodirnik K-means
   - One-hot kodirnik
6. Eksperiment
7. Rezultati
8. Diskusija in zaključne misli

Motivacija

- Od Industrije 3.0 (temelječe na avtomatizaciji in elektroniki) smo prešli v Industrijo 4.0 (pametne tovarne, internet stvari, kiberfizični sistemi (npr. robotika ter umetna inteligenca), itd.), sedaj pa sledi Industrija 5.0 (poudarek na sodelovanju ljudi in strojev s ciljem izboljšanja človekovega počutja).

- Inteligentne tehnike bodo vedno bolj ključnega pomena za družbo.

- Umetne inteligenca, in njena področja kot je strojno učenje, ter podpodročje okrepitvenega učenja pridobivajo na zaletu (npr., spomnimo samo na članek v reviji Nature podjetja Deep Mind glede množenja matrik).

Okrepitveno učenje predstavlja tip učenja znan pod izrazom *poskus-napaka* (angl. *trial-and-error*), saj mora agent na osnovi izvedbe akcij nad okoljem odkriti katere akcije mu prinašajo največjo nagrado.

Namen

- Potrebno je razvijati tudi klasične algoritme strojnega učenja, ker imajo kljub izrednemu napredku globokih nevronskih mrež, še veliko za ponuditi.

- Razširiti delovanje hibrida NARM-XCS z uporabo še drugih tipov kodirnikov. *Opomba: NARM-XCS je bil predhodno razvit za povezavo informacij pravil NARM z akcijami sistema XCS, ter za prilagajanje obstoječih pravil NARM novim stanjem okolja.*

- Analizirati kakšna je povezava uporabljenih različnih kodirnikov na doseženo nagrado hibridnega algoritma NARM-XCS pri uporabi raznovrstnih zbirk podatkov.

Numerično rudarjenje asociativnih pravil

- *Rudarjenje asociativnih pravil* (angl. *Association Rule Mining*, ARM) je zelo znana tehnika strojnega učenja, ki se uporablja za odkrivanje skritih vzorcev, povezav ali zakonitosti v velikih podatkovnih zbirkah.
- Veliko modernih pristopov ARM bazira na evolucijskih algoritmih in algoritmih roja delcev.
- ARM zgradi množico pravil, ki predstavljajo povezavo med atributi/lastnostmi v zbirki podatkov.
- Sprva usmerjen v uporabo zgolj nad kategoričnimi podatki, a se je nato razširil še nad numerično področje (angl. *Numerical ARM*, NARM).
- Pravilo NARM je definirano kot implikacija:

$$X \Rightarrow Y,$$

kjer je $X \subset F$, $Y \subset F$, in $X \cap Y = \emptyset$.

Množica atributov $F = \{A_1, \ldots, A_C; Q_1, \ldots, Q_R\}$

Kategorični atributi — Numerični atributi

Učeči se sistemi na osnovi klasifikatorjev

- *Učeči se sistemi na osnovi klasifikatorjev* (angl. *Learning Classifier Systems*, LCS) so družina algoritmov strojnega učenja, ki se uporabljajo za reševanje problemov klasifikacije, napovedovanja in optimizacije.
- Temeljijo na uporabi populacije klasifikatorjev, ki se zapiše kot množica: $C = \{C_1, C_2, \ldots, C_n\}$. Populacija klasifikatorjev kot celota predstavlja rešitev problema.
- Klasifikator si lahko predstavljamo kot zapis v obliki pogojnega stavka:

*»**ČE** pogoj, **POTEM** akcija«* (angl. *»**IF** condition, **THEN** action«*)

Pogoj je vektor, ki predstavlja stanje okolja, v katerem LCS deluje. Vsak element vektorja pogoja je ustvarjen iz nabora {0, 1, #}.

Razširjen sistem na osnovi klasifikatorjev

- V tem delu uporabljamo *razširjen sistem na osnovi klasifikatorjev* (angl. *eXtended Classifier System*, XCS), saj je ta znan:
  - po visoki prilagodljivosti klasifikatorjev v času delovanja,
  - maksimalno generaliziranih klasifikatorjih, ter, ker se ga da
  - uporabiti za množico različnih raziskovalnih problemov.

- S kombinacijo genetskega algoritma ter okrepitvenega učenja se doseže izboljševanje populacije klasifikatorjev, čemur tudi pravimo učenje klasifikatorjev.

Hibrid NARM-XCS

Hibrid med svojim delovanjem izvrši štiri glavne korake:

1. Priprava podatkov in inicializacija vseh komponent algoritma.
2. Ustvarjanje in procesiranje pravil NARM.
3. Izbira in uporaba kodirnika nad pravili NARM ter nad vrsticami podatkovne zbirke.
4. Vstavljanje kodiranih binarnih pravil v algoritem XCS.

Kodirani klasifikatorji se nato preko evolucijskih mehanizmov nadaljnje izpopolnjujejo s pomočjo okolja (tj., nad učno množico podatkovne zbirke), komponent okrepitvenega učenja in cenitvene funkcije, da se najde optimalna akcija.

Kodirniki

Za namen analize hibrida NARM-XCS smo razširili prejšnje raziskave z novimi kodirniki, ter z razširjenim sistemom začetnih parametrov zagona:

- **Binarni kodirnik**: Za pretvorbo numeričnih podatkov v binarno obliko. Osrednji koncept te metode je razdelitev atributov podatkovne zbirke v naprej določeno število intervalov oz. razdelkov. Razdelki se pretvorijo v nize binarnih števk z določeno dolžino ter se združijo v kodirani seznam.
- **Kodirnik k-means:** Algoritem, ki kodira numerične podatke z uporabo metode k-means (tj., k-srednjih) vrednosti. Za vsak atribut izračuna pripadajočo kategorijo z uporabo modela k-means vrednosti, pretvori to kategorijo v niz binarnih števk ustrezne dolžine, in jih združi v kodirani seznam.
- **One-Hot kodirnik**: Pred začetkom kodiranja se inicializira podatkovna struktura slovarja kategorij, kjer se vsakemu stolpcu v slovarju dodeli seznam edinstvenih vrednosti. Med kodiranjem se nato vsako vrednost v stolpcu pretvori v niz bitov, pri čemer je število bitov določeno glede na število kategorij v stolpcu. Za zmanjšanje dimenzionalnosti končne kodirane predstavitve pri kodirniku One-Hot uporabimo tehniko zgoščevanja (tj., z uporabo zgoščene tabele).

$$\text{številoBitov} = \lceil \log_2(\text{velikost}(\text{slovarKategorij})) \rceil$$

Eksperiment: Podatkovne zbirke

- Izbrane so bile tri podatkovne zbirke iz repozitorija kaggle (vse zbirke izdane v letu 2024):
  1. Vpliv zraka na kakovost zdravja (angl. Air Quality Health Impact, AQHI): 5.811 vrstic
  2. Bolezen raka dojk (angl. Breast Cancer Dataset, BCD): 569 vrstic
  3. Zbirka podatkov, ki se nanaša na statistične podatke iz igre League Of Legends (angl. League of Legends SoloQ matches at 15 minutes 2024, LOL): 24.225 vrstic.

- Za namen ustvarjanja pravil je bilo za vsako podatkovno zbirko začetno procesiranje pravil NARM izvedeno samo enkrat.

- Podatkovna zbirka se razdeli na učno in testno zbirko (razmerje 80 : 20).

Eksperiment: Okrepitveni del

- Posebnost našega algoritma je, da se lahko poljubno prilagodi na izbran atribut (tj., ne potrebujemo ravno klasifikacijske podatkovne zbirke). Naključno so tako bili izbrani atributi:
  1. AQHI: *PM10*
  2. BCD: *fractalDimensionWorst*
  3. LOL: *blueTeamTurretPlatesDestroyed*

- Posamezne nagrade se vrnejo s pomočjo cenitvene funkcije (spomnimo, XCS je algoritem okrepitvenega učenja):
  - Nagrada je vrednosti **dve** če je algoritem XCS izbral akcijo (tj., atribut), ki smo jo izbrali kot privzeto (tj., atribut, ki je izbran za prilagajanje populacije klasifikatorjev z algoritmom XCS).
  - Drugače se vrne negativna nagrada v vrednosti minus **ena**.

- Skupna nagrada se izračuna tako, da se vsota vseh prejetih nagrad deli s številom vrstic podatkovne zbirke (tj., učne (uz) ali pa testne zbirke (tz)).

Eksperiment: Nastavitve zagona

- Nad vsako podatkovno zbirko se je izvršil zagon vseh treh kodirnikov po petkrat, nato se je izračunala povprečna vrednost nagrad.
- Sistem zagona je šel po sistemu:
  - Binarni kodirnik: Enkrat je bilo število razdelkov nastavljeno na dva (b2) in enkrat na tri (b3).
  - k-means: Enkrat z opcijo k je dva (k2) in drugič z opcijo k je tri (k3).
  - Kodirnik One-hot: Ne potrebuje začetnih nastavitev parametrov.
- Število ponovitev algoritma k-means je bilo nastavljeno na 100 iteracij.
- Pri procesiranju učne in testne zbirke se je eksperiment izvršil enkrat brez vstavljenih pravil v populacijo hibrida NARM-XCS, ter enkrat s sto procenti vstavljenih pravil.
- Algoritma NARM in XCS sta imela vrednosti nastavljene enako kot v originalnem članku ([5]).

[5] D. Novak, D. Verber, I. Fister, I. Fister Jr.: Association Rule Mining as Knowledge Infusion into the Mechanics of an eXtended Classifier System, 2024 IEEE 28th INES, str. 203-208, 2024.

Rezultati

Kodirnik (procent vstavljenih pravil, zbirka podatkov [učna zbirka: uz, testna zbirka: tz])	Nagrada
Binarni kodirnik_b2 (0, uz)	1,778
Binarni kodirnik_b3 (0, uz)	1,738
Kodirnik k-means_b2 (0, uz)	1,812
Kodirnik k-means_b3 (0, uz)	1,826
Kodirnik One-Hot (0, uz)	1,784
Binarni kodirnik_b2 (0, tz)	1,817
Binarni kodirnik_b3 (0, tz)	1,75
Kodirnik k-means_k2 (0, tz)	1,865
Kodirnik k-means_k3 (0, tz)	1,865
Kodirnik One-Hot (0, tz)	1,873
Binarni kodirnik_b2 (100, uz)	1,869
Binarni kodirnik_b3 (100, uz)	1,863
Kodirnik k-means_k2 (100, uz)	1,854
Kodirnik k-means_k3 (100, uz)	0,372
Kodirnik One-Hot (100, uz)	1,798
Binarni kodirnik_b2 (100, tz)	1,893
Binarni kodirnik_b3 (100, tz)	1,896
Kodirnik k-means_k2 (100, tz)	1,872
Kodirnik k-means_k3 (100, tz)	1,672
Kodirnik One-Hot (100, tz)	1,877

Podatki za podatkovno zbirko AQHI.

Kodirnik (procent vstavljenih pravil, zbirka podatkov [učna zbirka: uz, testna zbirka: tz])	Nagrada
Binarni kodirnik_b2 (0, uz)	1,676
Binarni kodirnik_b3 (0, uz)	1,875
Kodirnik k-means_b2 (0, uz)	1,707
Kodirnik k-means_b3 (0, uz)	N/A
Kodirnik One-Hot (0, uz)	1,84
Binarni kodirnik_b2 (0, tz)	1,532
Binarni kodirnik_b3 (0, tz)	1,868
Kodirnik k-means_k2 (0, tz)	1,611
Kodirnik K-means_k3 (0, tz)	N/A
Kodirnik One-Hot (0, tz)	1,795
Binarni kodirnik_b2 (100, uz)	1,739
Binarni kodirnik_b3 (100, uz)	1,854
Kodirnik k-means_k2 (100, uz)	1,761
Kodirnik k-means_k3 (100, uz)	N/A
Kodirnik One-Hot (100, uz)	1,821
Binarni kodirnik_b2 (100, tz)	1,726
Binarni kodirnik_b3 (100, tz)	1,847
Kodirnik k-means_k2 (100, tz)	1,663
Kodirnik k-means_k3 (100, tz)	N/A
Kodirnik One-Hot (100, tz)	1,816

Podatki za podatkovno zbirko BCD.

Kodirnik (procent vstavljenih pravil, zbirka podatkov [učna zbirka: uz, testna zbirka: tz])	Nagrada
Binarni kodirnik_b2 (0, uz)	1,835
Binarni kodirnik_b3 (0, uz)	-0,953
Kodirnik k-means_b2 (0, uz)	1,661
Kodirnik k-means_b3 (0, uz)	N/A
Kodirnik One-Hot (0, uz)	-0,322
Binarni kodirnik_b2 (0, tz)	1,869
Binarni kodirnik_b3 (0, tz)	-0,992
Kodirnik k-means_k2 (0, tz)	1,843
Kodirnik k-means_k3 (0, tz)	N/A
Kodirnik One-Hot (0, tz)	1,314
Binarni kodirnik_b2 (100, uz)	1,85
Binarni kodirnik_b3 (100, uz)	-0,939
Kodirnik k-means_k2 (100, uz)	1,86
Kodirnik k-means_k3 (100, uz)	N/A
Kodirnik One-Hot (100, uz)	0,588
Binarni kodirnik_b2 (100, tz)	1,874
Binarni kodirnik_b3 (100, tz)	-0,991
Kodirnik k-means_k2 (100, tz)	1,867
Kodirnik k-means_k3 (100, tz)	N/A
Kodirnik One-Hot (100, tz)	1,87

Podatki za podatkovno zbirko LOL.

Diskusija

- Rezultati so pokazali izredno odpornost hibridnega algoritma na tip izbranega kodirnika, saj so vsi trije dosegali vrednosti nagrad blizu zgornjim mejam, ki jih lahko zagotovi osnovni algoritem XCS.

- Razlika v nagradah brez dodanih pravil in pri sto procentov vstavljenih pravil NARM v XCS je zaznavna (tj., v prid vstavljanja pravil), vendar je manjšega obsega kot pri dveh zbirkah uporabljenih v originalnem članku.

- One-Hot kodirnik se je pokazal kot izredno dobra izbira (pri čemer je še veliko „manverskega“ prostora glede ne-statične dolžine klasifikatorjev).

V prihodnosti

- Preučiti kakšna je povezava specifične zbirke podatkov (tj., glede števila vrstic zbirke in razporeditve podatkov) z različnimi vrednosti hiperparametrov (npr. velikost populacije XCS).

- Raziskati povezavo vplivov različnih procentov vstavljenih pravil NARM v populacijo XCS algoritma, ter odkriti če je kakšen specifičen vzorec podatkovnih zbirk za katere se hibridni algoritem še posebej izkaže kot izredno dobra izbira.

- Izboljšati cenitveno funkcijo.

- Nadgraditi še ostale komponente algoritma XCS, ter poiskati njihovo sinergijo.

Hvala za vaš čas. Prosim, pridružite se diskusiji.