ANALIZA MOŽNOSTI ZAZNAVANJA PODOBNOSTI MED UPORABNIKI

Božidara Cvetković, Mitja Luštrek
Department of Intelligent Systems
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: {boza.cvetkovic, mitja.lustrek}@ijs.si

POVZETEK

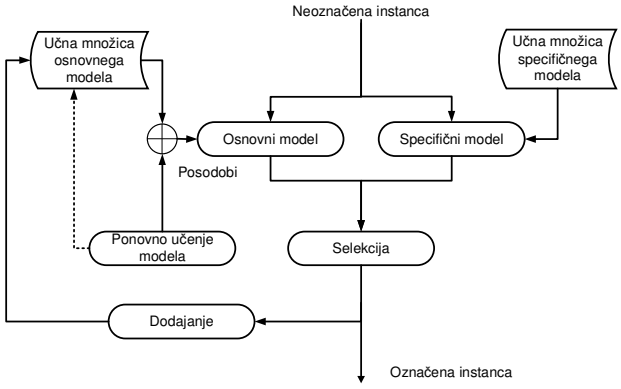
Prispevek predstavlja preliminarne rezultate analize možnosti zaznavanja podobnosti med uporabniki. Cilj analize je izbrati najboljši pristop, ki bo uporabljen v metodi za prilagajanje modela uporabniku MCAT.

1 UVOD IN SORODNO DELO

V aplikacijah, kjer se uporabljajo modeli strojnega učenja za napovedovanje človeškega obnašanja, se pogosto dogaja, da točnost delovanja v realnem okolju ni primerljiva točnosti delovanja v laboratorijskem okolju. Razlog je tako omejena količina učnih podatkov, kot tudi fizična razlika ter razlika v navadah med ljudmi. Fizične razlike se kažejo bodisi v drugačnosti izvajanja akcij v primeru problema prepoznavanja aktivnosti ali v drugačnem metabolnem sistemu v primeru problema ocene porabe energije.
Točnost modela za določenega uporabnika lahko zvišamo na dva načina:

- označimo dodatne učne podatke specifične za novega uporabnika in uporabimo nadzorovano učenje za nov model ali
- uporabimo katero od metod, ki nenadzorovano ali pol-nadzorovano prilagodijo model trenutnemu uporabniku.

Najboljše izboljšanje dobimo z označevanje dodatnih podatkov. Vendar je ta proces časovno zelo zahteven, duhamoren in drag, tako za označevalca kot za uporabnika. Velikokrat se zgodi, da je samo označevanje podatkov v ciljnem okolju onemogočeno, bodisi zaradi samega klasifikacijskega problema (označevanje padcev je lahko nevarno) ali pa zato, ker nam manjkajo dodatne naprave, ki niso mobilne in jih lahko uporabljamo izključno v laboratoriju (poraba človeške energije iz izdihanega zraka). V tem primeru se izkažejo rešitve, ki uporabljajo pol-nadzorovano učenje, bolj primerne. Metode pol-nadzorovanega učenja označijo neoznačene podatke in glede na določeno pravilo izberejo ali zavržejo trenutni podatek za dodajanje v učno množico. Nad učno množico, ki vsebuje nove podatke, se nato uporabi nadzorovan algoritem za strojno učene za pridobitev novega, prilagojenega modela.
Metode pol-nadzorovanega učenja lahko kategoriziramo na več načinov. Glede na število klasifikatorjev, glede na število dimenzij (ortogonalnost atributnih vektorjev), glede na način prilagajanja in glede na to ali se uporablja omejena ali neomejena količina neoznačenih podatkov. Najbolj osnovna metoda je samo-učenje (self-training [1]), ki uporablja en klasifikator za označevanje podatkov in ročno nastavljen prag za odločitev o izbiri podatka za dodajanje v učno množico. Prag je po navadi nastavljen tako, da mora biti zaupanje v napoved 100%. Nadgradnja metode z enim klasifikatorjem je dodajanje več klasifikatorjev, ki so naučeni z različnimi algoritmi in za dodajanje uporabljajo večinski glas (Democratic co-learning [2]) ali več klasifikatorjev z istim učnim algoritmom in več dimenzijami (Co-training [3]). Pomanjkljivost prvega je v ročno nastavljenem pragu (100% zaupanje v napoved), problem drugega pa kompleksnost delitve prostora na dva ortogonalna dela ali dimenziji. Več o metodah pol-nadzorovanega učenja pišemo v našem preteklem delu, kjer smo prilagajali klasifikator za prepoznavanje aktivnosti novemu uporabniku. Pokazali smo, da lahko z mehanizmom za prilagajanje novemu uporabniku (MCAT - Multi-Classifier Adaptive Training [4]) in omejeno količino na novo označenih podatkov (3 aktivnosti po 30 sekund) zvišamo prepoznavanje aktivnosti za približno 12 odstotnih točk. Ogrodje MCAT metode je okvirno predstavljena na sliki 1.



Slika 1: *Ogrodje metode MCAT*

Ogrodje MCAT pričakuje naslednje klasifikatorje:

- osnovni model: model, ki se je v laboratorijskem okolju izkazal za najboljšega,
- specifični model: model ali množica modelov, ki vsebujejo znanje o specifikah trenutnega uporabnika,

- selekcija: model, ki izbere končno oznako,
- dodajanje: model, ki se odloča, ali je trenutna instanca dovolj kvalitetna za dodajanje v učno množico osnovnega modela.

Cilje trenutne raziskave je uporabiti isto ogrodje na regresijski domeni, bolj specifično za oceno porabe človeške energije. Označevanje podatkov za novega uporabnika je v tem primeru onemogočeno, saj bi uporabnik moral v laboratorij, kjer se nahajajo potrebne naprave (Cosmed k4b2).

Ta prispevek predstavlja analizo pristopov za možnost detekcije podobnosti med uporabniki. Privzeli bi, da pristop z najboljšim delovanjem opiše trenutnega uporabnika zadosti dobro da ga lahko uporabimo kot specifični model v MCAT algoritmu.

2 NABOR PODATKOV

V raziskavi smo uporabili dva nabora podatkov in sicer podatke, ki so uporabljeni kot učna množica splošnega modela in pa nabor podatkov, ki predstavlja bio-impedanco oseb vsebovanih v učni množici splošnega modela.

Učna množica splošnega modela je bila zbrana v kontroliranem laboratorijskem okolju Fakultete za Šport in vsebuje podatke 10 ljudi, ki so izvajali vnaprej določene sklope aktivnosti. Opremljeni so bili s pospeškomeri na prsih in stegnu, prsnim pasom za merjenje srčnega utripa, napravo Senswear, ki meri oddajanje toplote človeka, galvanski odziv kože in telesno temperaturo ter oceni človekovo porabo energije in indirektnim kalorimetrom Cosmed k4b2, ki meri porabo energije na osnovi izdihanega ogljikovega dioksida in porabe kisika. Ta nabor podatkov je bil uporabljen za gradnjo in vrednotenje več regresijskih modelov za oceno porabe energije. Izbran je bil najboljši, ki vsebuje podatke pospeškomerov, blizu telesne temperature in srčnega utripa. Ta model je privzet za splošni model, deluje s povprečno absolutno napako (MAE) 0.55 MET (Metabolic Equivalent of Task).

Učna množica bio-impedance so podatki, pridobljeni iz naprave InBody [1], ki analizira sestavo telesa. Podatki vsebujejo: višino, težo, starost, količino vode v celicah, izven celic, količino proteinov, mineralov, maščobe, maso skeleta, indeks telesne teže, razmerje med pasom in boki in podatke o teži udov. Vsebuje tudi maksimalne in minimalne vrednosti za vsak tip podatkov, kar smo uporabili na normalizacijo in dodali se maksimalen in minimalen srčni utrip uporabnika. Ta je bil umerjen med 15 minutnim ležanjem (minimalen srčni utrip) in po dveh minutah intenzivnega teka (maksimalen srčni utrip). To učno množico smo uporabili za ugotavljanje podobnosti med uporabniki.

3 PRISTOP ZA UGOTAVLJANJE PODOBNOSTI MED UPORABNIKI

Podobnost med uporabniki smo analizirali z uporabo nabora podatkov bio-impedance in testirali na podatkih osebe, katere podobnost smo ugotavljali. Cilj je pridobiti množico oseb, ki so najbolj podobni novemu uporabniku in iz njihovih modelov oceniti porabo energije novega uporabnika. Točnost ocene mora biti višja od splošnega modela, ki je naše izhodišče.

3.1 Razvrščanje v skupine ali gručenje

Za razvrščanje v skupine smo uporabili algoritem k-means iz orodja za strojno učenje Weka [6]. Za idealno število gruč smo uporabili koeficient Silhouette, ki poda mero, kako dobro podatek ustreza trenutni gruči. Koeficient je definiran z naslednjo enačbo.

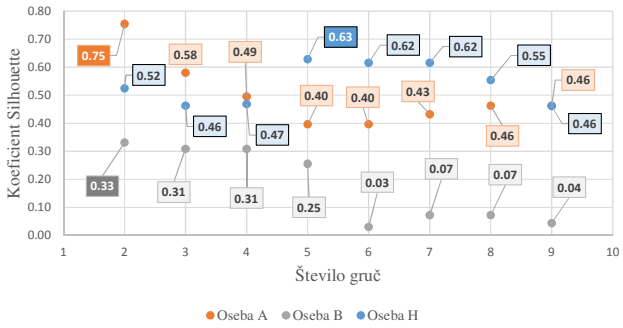
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$-1 \leq s(i) \leq 1$$

Za izračun koeficienta uporabnika *i* uporabimo

- a(i) - povprečna razdalja vseh uporabnikov v gruči
- b(i) - najmanjša razdalja trenutnega uporabnika do sosednje gruče

Ustreznost gruče je definirana z velikostjo koeficienta. Najbolj ustrezna delitev je pri *s(i) = 1*, če je koeficient blizu *0* je na robu dveh gruč in če je *-1* verjetno bolj ustreza drugi gruči. Izračunan koeficient za tri osebe lahko vidimo na Sliki 2. Za osebi A in B je najboljša delitev na dve gruči in za osebo H na 5 gruč.



Slika 2: *Silhouette koeficient ustreznosti delitve.*

S to metodo smo dobili gruče podobnih oseb.

3.2 Meta klasifikacija

Za uteževanje ocen smo poizkusili še meta-klasifikator za vsako osebo posebej. Za meta-klasifikator smo uporabili podatke osmih oseb pri ocenjevanju devete. Za končne evaluacijo smo uporabili deseto osebo.

Začetno množico atributov meta klasifikatorja sestavljajo naslednji atributi:

- evklidske razdalje od trenutne osebe do vseh oseb v gruči,
- trenutna razpoznana aktivnost osebe,
- nivo aktivnosti (nizka, srednja, visoko),
- normaliziran srčni utrip osebe,

Tabela 1: *Rezultati glede na pristop ugotavljanja podobnosti med uporabniki. Pristopi so opisani v sekciji 3.3.*

				Pristopi					
	Splošni model (MAE)	Število gruč	Število oseb v gruči	A	B	C	D	E	F
Oseba A	0.49	2	8	0.53	0.49	0.49	0.49	0.48	0.48
Oseba B	0.69	2	3	0.77	0.69	0.70	0.69	0.73	0.69
Oseba C	0.64	3	4	0.75	0.60	0.61	0.60	0.58	0.59
Oseba D	0.55	4	1	0.93	0.54	0.54	0.54	0.48	0.49
Oseba E	0.44	2	8	0.40	0.47	0.48	0.47	0.44	0.44
Oseba F	0.55	2	8	0.68	0.60	0.60	0.60	0.55	0.55
Oseba G	0.57	2	8	0.50	0.61	0.62	0.61	0.56	0.56
Oseba H	0.46	5	2	0.42	0.51	0.51	0.51	0.46	0.46
Oseba I	0.64	2	8	0.67	0.63	0.63	0.63	0.72	0.63
Oseba J	0.50	6	1	0.65	0.47	0.47	0.47	0.53	0.50
Povprečno	**0.55**			**0.63**	0.56	0.56	0.56	0.55	**0.54**

- povprečna absolutna napaka ocene modela osebe glede na oceno splošnega modela,
- cona srčnega utripa po metodi Zoladz [8],
- procent povprečne absolutne napake ocene modela glede na oceno splošnega modela.

Delovanje meta-klasifikatorja je naslednje. Vsako instanco se oceni z modeli oseb, ki so v gruči, in vsaka ocena je ovrednotena s svojim meta-klasifikatorjem, ki vrne enega od dveh razredov: »da« ali »ne«. Da pomeni, da se ocena uporabi, in ne, da se zavrže. Poleg vsake klasifikacije klasifikator vrne stopnjo zaupanja v svojo napoved. Končna ocena se izračuna glede na število modelov, katerih rezultat je bil »da«:

- število »da« > 1; normalizira se stopnja zaupanja za vsak model, ki je klasificiral »da«. Normalizirane stopnje se uporabijo kot utež trenutne ocene in utežena vsota vseh tvori končno oceno,
- število »da« = 1; stopnja zaupanja je uporabljena kot utež ocene tega modela. Ostanek je uporabljen kot utež ocene splošnega modela. Utežena vsota obeh tvori končno oceno,
- število »da« = 0; uporabi se ocena splošnega modela

Uporabnost atributov smo ovrednotili s kombiniranjem vseh in izločili tiste atribute, ki ne pripomorejo k boljši točnosti izbire in hkrati točnosti ocene. Atributi, ki so ostali v končnem vektorju atributov, so:

- evklidske razdalje od trenutne osebe do vseh oseb v gruči,
- trenutna razpoznana aktivnost osebe,
- nivo aktivnosti (nizka, srednja, visko),
- cona srčnega utripa po Zoladz metodi [8].

3.3 Pristopi

Pristop A: Vsako instanco oceni devet modelov (posamezni model osebe) in končna ocena je povprečje ocen.

Pristop B: Vsako instanco ocenijo modeli oseb, ki so v gruči ,in končna ocena je povprečje ocen.

Pristop C: Vsako instanco ocenijo modeli oseb, ki so v gruči, in končna ocena je utežena vsota glede na evklidsko razdaljo do centroide v gruči.

Pristop D: Vsako instanco ocenijo modeli oseb, ki so v gruči, in končna ocena je utežena vsota glede na evklidsko razdaljo do nove osebe v gruči. Če je v gruči ena oseba je rezultat utežena vsota splošnega modela in modela osebe.

Pristop E: Za oceno so uporabljeni meta klasifikatorji in modeli vseh oseb.

Pristop F: Za oceno so uporabljeni meta klasifikatorji in modeli oseb v gruči.

4 REZULTATI

Rezultati predstavljajo evaluacijo vseh omenjenih pristopov. Cilj je izbrati pristop, ki vrača manjšo ali primerljivo točnost splošnemu modelu. Rezultati so predstavljeni v Tabeli 1 in sicer z povprečno absolutno napako (MAE) definirano z naslednjo enačbo:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|EE_{ocenjena} - EE_{prava}\right|$$

Končna ocena najboljšega pristopa je ocenjena z povprečno absolutno procentualno napako definirano z naslednjo enačbo (MAPE):

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{EE_{ocenjena} - EE_{prava}}{EE_{prava}}\right|$$

V obeh enačbah $EE_{ocenjena}$ predstavlja oceno porabe energije, kot jo vrne regresijski model in EE_{prava} je izmerjena poraba energije.

Točnost splošnega modela je predstavljena v drugem stolpcu Tabele 1. Povprečna napaka modela je 0.55 MET in MAPE modela je 25%. Prvi pristop (pristop A) uporabi povprečno

oceno vseh oseb. Iz rezultata lahko vidimo, da se napaka poveča in da ta pristop ni pravilen, kar je tudi v skladu s hipotezo, da uporabljen model mora biti podoben modelu končne osebe. Pristop B uporabi dodatno znanje o medsebojni podobnosti oseb in za končno oceno uporabi povprečje ocen podobnih oseb (osebe v isti gruči). Rezultat je slabši od splošnega modela, tako v obliki MAPE 26% kot tudi MAE 0.56 MET. Pristop C uporabi za utež napovedi evklidsko razdaljo osebe do centroide. Končna točnost je slabša od splošnega modela in sicer 0.56 MET in 26% v obliki MAPE. Pristop D vrne primerljive rezultate kot pristopa B in C. Pristop E uporabi meta-klasifikator, vendar na vseh osebah. Iz rezultata lahko vidimo, da z vpeljavo meta klasifikatorja dosežemo primerljivo točnost, kot ga dobimo s splošnim modelom. Če uporabimo meta klasifikatorje samo na osebah ki so v gruči, pa pridobimo na točnosti in sicer 0.01 MET v obliki MAE in 3 odstotne točke v obliki MAPE.

5 ZAKLJUČEK

Ta prispevek predstavlja preliminarne rezultate analize pristopov za ugotavljanje podobnosti med uporabniki. Analiza je bila narejena na domeni ocene porabe človeške energije z namenom definirati specifični model za pol-nadzorovano metodo MCAT, katero bomo v prihodnjem delu nadgrajevali.

Pristop, ki vrača najboljšo točnost, uporablja algoritem gručenja za delitev oseb v skupine po podobnosti in meta klasifikatorje posameznih oseb v gruči za končno oceno porabe energije osebe. Z uporabo pristopa za podobnost izboljšamo rezultat najboljšega modela za 3 odstotne točke.

Prihodnje delo zajema razširitev pristopov in uporabo najboljšega pristopa v metodi MCAT.

References

[1] Frinken, V., Bunke, H.: Self-training Strategies for Handwriting Word Recognition. In: Perner P. (eds.) Advances in Data Mining. Applications and Theoretical Aspects. LNCS, vol. 5633, pp. 291--300, 2009.

[2] Zhou, Y., Goldman, S.:Democratic Co-Learning. In: 16th IEEE International Conference on Tools with Artificial Intelligence, pp. 594--602, IEEE press, 2004

[3] Blum, A. and Mitchell, T.: Combining labeled and unlabeled data with co-training. In: 11th annual conference on Computational learning theory, pp. 92--100, 1998.

[4] B. Cvetković, B. Kaluža, M. Luštrek, M. Gams, "Adapting Activity Recognition to a Person with Multi-Classifier Adaptive Training," Journal of Ambient Intelligence and Smart Environments. Accepted for publication, 2014.

[5] InBody, http://www.e-inbody.com/

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations. 11(1), pp. 10–18, 2009.

[7] Silhouette koeficient, http://en.wikipedia.org/wiki/Silhouette_(clustering)

[8] Zoladz, http://en.wikipedia.org/wiki/Heart_rate