ALGORITEM LOF KOT METODA V SISTEMU ZA PODPORO ODLOČANJA

Božidara Cvetković[1,2], *Mitja Luštrek*[1]
[1] Odsek za inteligentne sisteme
Institut »Jožef Stefan«
[2] Mednarodna Podiplomska šola Jožefa Stefana
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: {boza.cvetkovic, mitja.lustrek}@ijs.si

POVZETEK

Sistemi za podporo odločanja imajo večinoma zelo kompleksno arhitekturo. Pomemben del takšnega sistema so eden ali več modulov za oceno tveganja, ki temeljijo na ekspertnem znanju ali na veliki količini označenih podatkov. V tem prispevku bomo predstavili nenadzorovnano metodo LOF in jo predlagali za enega v množici modulov v sistemih za podporo odločanja. Algoritem smo spremenili in dopolnili za namen ocenitve tveganja pri pacientih s srčnim popuščanjem. Predstavljene spremembe so (i) računanje razdalje med nominalnimi atributi, (ii) vrednotenje stopnje nenavadnosti dogodka in (iii) ocena tveganja po parametru. Z eksperimentom smo dobili preliminarne rezultate, ki kažejo na potencialno koristnost spremenjenega algoritma.

1 UVOD

Sistemi za podporo odločanja (SPO) so informacijski sistemi namenjeni asistenci ekspertom pri odločitvenih aktivnostih. Glavni cilj je pospešitev samega procesa odločanja. Zaradi preobremenitve vodilnih profesionalcev v posameznih domenah in pomanjkanja časa za podrobno analizo podatkov, zainteresiranost v takšne sisteme raste. Za nas so zanimivi predvsem inteligentni SPO (slo. ISPO, ang. IDSS (Intelligent Desicion Support System)), ki vsebujejo algoritme umetne inteligence za naprednejšo analizo podatkov. Trenutno lahko najdemo veliko SPO sistemov, ki so razviti za delovanje na različnih domenah kot naprimer: (i) poslovna inteligenca, za hitrejše zaznavanj negativnih in pozitivnih trendov in spremljanje prostih resursov [1], (ii) kriminalne preiskave [2], (iii) klinične preiskave in nadzor pacientov [3] itd.

ISPO sistemi so z visokega nivoja gledano sestavljeni iz baz podatkov, modeliranega znanja in uporabniškega vmesnika. Modelirano znanje je skupek modulov, ki lahko vsebujejo inteligentne metode ali pa tudi ne, končna odločitev sistema pa je agregacija odločitev vseh vsebujočih modulov. Metode, ki se pojavljajo v modulih, so v večini primerov metode nadzorovanega učenja, kot so nevronske mreže, genetski algoritmi, odločanje na osnovi pravil itd. Za zaupanja vredno delovnje teh sistemov je potrebno metodam zagotoviti veliko količino označenih podatkov. Zajemanje označenih podatkov je časovno problematično, prav tako tudi zanesljivost označenih podatkov.

V tem prispevku smo raziskali možnost uporabe metode, ki deluje na osnovi neoznačenih podatkov in omogoča ločitev normalnih od nenormalnih dogodkov. Med takmi metodami smo izbrali Local Outlier Factor (LOF) algoritem [4]. Ta algoritem smo izbrali zaradi njegove enostavnosti in razumljivosti. Dodatno smo algoritem LOF tudi razširili, da bi omogočili vse lastnosti, ki jih dober modul SPO ima. In sicer z: (i) računanjem razdalje med nominalnimi atributi, s tem smo razširili metodo tako da, lahko uporablja več vrst atributov, (ii) vrednotenjem stopnje nenavadnosti, da bo sistem vračal tudi oceno kako tvegan je trenuten dogodek in (iii) oceno tveganja po parametru, da lahko uporabniku svetujemo kateri parametri iz množice mnogih so prispevali k nenavadnosti dogodka.

Algoritem je bil razširjen in uporabljen za klinično podporo odločanja in je le eden od modulov, ki bodo delovali znotraj sistema. Preliminarni rezultati kažejo na uporabnost algoritma LOF za oceno tveganja, ki pa so sicer nastali na sintetičnih podatkih.

Nadaljevanje prispevka je razdeljeno v pet poglavij. V poglavju 2 predstavimo uporabljen algoritem, v poglavju 3 predstavimo razširitev algoritma. Sekcija 4 vsebuje eksperiment in rezultate. S sekcijo 5 zaključujemo prispevek.

2 ALGORITEM LOF

Algoritem LOF je v prvi vrsti namenjen zaznavanju nenavadnih dogodkov z uporabo gruč in iskanjem izjem na podlagi gostote med primeri v posamezni gruči. Za svoje delovanje ne potrebuje označenih podatkov, oziroma na začetku delovanja se določeni podatki samodejno označijo za normalne in porazdelijo po posameznih gručah. Vsak nadaljnji primer, ki ga algoritem analizira se primerja z normalnimi podatki, ki so že vsebovani v algoritmu. Rezultat algoritma je faktor LOF, ki kaže koliko je primer drugačen od ostalih normalnih primerov.

Recimo da je $N_k(A)$, k število najbližjih sosedov instance A in *k-razdalja(A)* razdalja med A in najbolj oddaljeno instanco v $N_k(A)$. Razdalja dosega (ang. reachability distance) med instancami A in B je definirana kot razdalja

me dvema instancama *d(A,B)*, kjer A ni v množici najbližjih sosedov B. V tem primeru je razdalja enaka *k-razdalja(B):*

$$razdalja\_dosega_k\,(A, B) = max\,(d\,(A, B), k\text{-}razdalja\,(B))$$

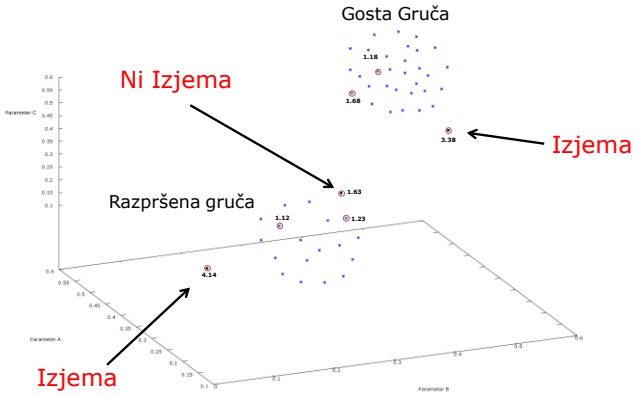
Lokalni doseg instance (lrd) *A* je definiran kot 1 ulomljeno s povprečnim dosegom od *A* do vseh *k* najbližjih sosedov:

$$lrd(A) = \frac{1}{\left(\frac{\sum_{B \in N_k(A)} razdalja\_dosega_k(A,B)}{|N_k(A)|}\right)}$$

Končen lokalni doseg instance A se primerja z lokalnimi dosegi svojih sosedov in to vrne vrednost LOF:

$$LOF_k(A) = \left(\frac{\sum_{B \in N_k(A)} lrd_k(B)}{|N_k(A)|}\right) / lrd(A)$$

Vrednost faktorja LOF okoli ena je primerljiva s sosedi in je privzeta za normalno, LOF z vrednostjo manj kot ena pomeni, da ležijo v zelo gosti gruči. Če je vrednost LOF višja kot ena, pomeni, da je instance izven gruče in potencialno nenavadna. Kolikšna naj bo vrednost faktorja, da je instanca nenavadna, je odvisno od podatkov.
Slika 1 prikazuje vizualno predstavitev dveh različno gostih gruč in detekcijo izjem z algoritmom LOF in njihove vrednosti.



Slika 1: Vizualna predstavitev gruč in detekcija izjem z algoritmom LOF.

Če želimo uporabiti algoritem LOF za oceno tveganja, se mora naučiti na podatkih, ki predstavljajo normalno stanje. Po navadi je to *n* podatkov ob zagonu algoritma. Poleg tega je potrebno določiti kdaj je instanca čudna in za koliko. To bo podrobneje predstavljeno v naslednji sekciji.

3 MODIFIKACIJA ALGORITMA

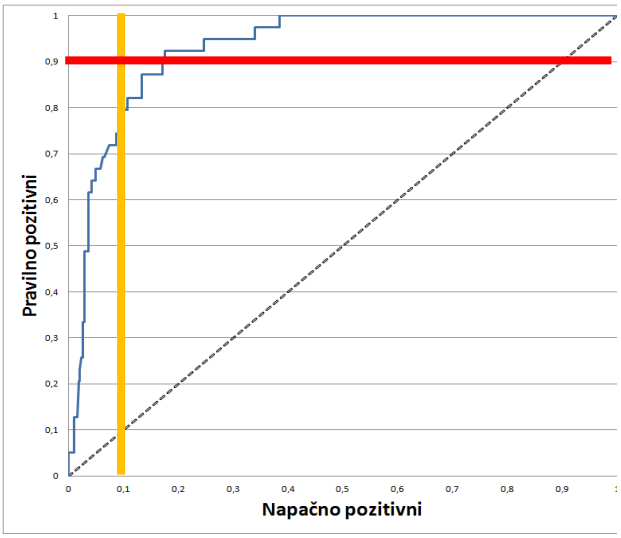
Algoritem LOF po svoji naravi lahko opredeli instanco binarno, normalna ali abnormalna. V našem primeru pa bi želeli LOF vrednost razširiti tako da bi nam povedala tudi stopnjo abnormalnosti. Velikokrat imamo kot atribut nominalno vrednost in v primeru standardnega LOF-a so ti atributi neuporabni. Predstavili bomo metodo, ki opredeli razdaljo med nominalnimi vrednostmi atributov in jih na ta način naredi uporabne.

3.1 Računanje razdalje med nominalnimi atributi

Ustaljena metoda za določanje razdalje med nominalnimi atributi je, da če sta si dva atributa enaka, je razdalja med njima 0, in če sta si različna, je med njima razdalja 1. To sicer velja, vendar če je nominalnih vrednosti atributov več, si želimo da bi razdalja odražala dejansko razdaljo, ki bi bila skladna s pomenom vrednosti.
V našem primeru so nominalne vrednosti rezultat predhodne klasifikacije. Klasifikator je bil naučen na označenih podatkih in ločenih atributih.
Hipoteza je, da lahko iz numeričnih vrednosti atributov, ki se uporabljajo za klasifikacijo določimo razdaljo med nominalnimi vrednostmi. In sicer če imamo *n* razredov, lahko za vsak razred določimo povprečno vrednost vsakega atributa. Nakar lahko z evklidsko razdaljo izračunamo kolikšna je razdalja med posameznimi razredi glede na povprečne numerične vrednosti atributov. V sekciji z eksperimentom bomo pokazali, da na ta način razdalja odraža pomensko razliko med vrednostmi.

3.2 Določanje stopnje nenaormalnosti dogodka in ocena tveganja po parametru

Da bi lahko LOF vračal tudi stopnjo čudnosti smo morali izbrati metodo za določanje pragov faktorja LOF. Že v prejšnjem poglavju smo omenili, da je privzeto, da če je LOF faktor v okolici vrednosti 1 lahko to instanco označimo za normalno. Če želimo razširiti algoritem na več stopenj, v našem primeru tri (nizko, srednje in visoko tveganje), pa je potrebno določiti, kakšne so vrednosti pragov. Tukaj smo uporabili označene podatke.



Slika 2: ROC krivulja in opredeljene meje za pragove za srednje- in visoko tveganje. Rumena premica je za srednje tveganje in rdeča za visoko tveganje.

Za določanje pragov smo uporabili ROC krivuljo in iz nje prebrali odstotek pravilno določenih instanc glede na vrednost postavljenega praga. Na sliki 2 vidimo z modro črto največje AUC področje, ki smo ga dobili, če smo uporabili vrednost ena za število sosedov.
Prag za visoko tveganje je postavljen tako, da pade nad njega 90% primerov označenih kot zelo visoko tveganje. Prag za srednje tveganje je postavljen tako, da lahko 10% primerov označenih kot srednje normalno pade pod prag. Te meje so tudi prikazane na sliki 2, kjer je rumena premica prag za srednje tveganje in rdeča premica za visoko tveganje.
Naslednja modifikacija LOF-a je ocenitev tveganja za vsak parameter posebej. Cilj je, da ko se uporabnika obvesti o tem, da je nek dogodek tvegan, da se mu predlagajo tudi parametri, ki so prispevali k tveganju. Rešitev smo našli v tem, da za vsak parameter posebej izvedemo algoritem LOF. Prav tako je postopek postavljanja pragov identičen kot za pragove tveganja celotnega sistema.

4 EKSPERIMENT IN REZULTATI

Metodo smo poizkusili in evakuirali na sintetičnih podatkih petih ljudi. Cilj je metodo uporabiti na ljudeh s srčnim popuščanjem, vendar takšne podatke bomo dobili v sklopu projekta CHIRON [5] za katerega je bila metoda razvita.

4.1 Podatki

Eksperiment smo izvedli na podatkih, ki so bili posneti za prepoznavanje aktivnosti in oceno porabe energije [6]. Uporabili smo posnetke petih ljudi, ki so imeli na sebi pospeškomere in prsni trak Zephyr, ki meri srčni utrip. Uporabili smo klasifikatorje za oceno porabe energije in prepoznavanje aktivnosti. Poleg teh dveh podatkov smo uporabili tudi srčni utrip in temperaturo kože. Tako da smo za evaluacijo uporabili le štiri atribute, od tega enega nominalnega. Vsaka oseba je posnela pet ponovitev enega scenarija. Scenariji so podrobneje opisani v prispevku o porabi energije [6].
Za metodo smo generirali deset nenormalnih posnetkov tako da smo za eno osebo zamenjali signale. Torej pri scenariju, kjer je bilo ležanje, smo zamenjali srčni utrip s tistim in scenarija tek. Tako smo dobili sintetične nenormalne podatke za srednje tveganje. Za visoko tveganje smo zamenjali dva signala s signali iz drugih scenarijev.

4.2 Rezultati

Razdalja med nominalnimi atributi

V atributnem vektorju smo od štirih atributov imeli en nominalni atribut in sicer aktivnost. Da bi dobili vrednost, ki realno odraža razdaljo med dvema paroma aktivnosti, smo uporabili metodo opisano v sekciji 3.1. Klasifikator za prepoznavanje aktivnosti smo razdelili po razredih in dobili devet podmnožic vrednosti atributov in sicer za naslednje aktivnosti:

- na vseh štirih,
- klečanje,
- kolesarjenje.
- ležanje,
- tek,
- sedenje,
- stanje,
- tranzicija in
- hoja.

Za vsako aktivnost smo izračunali povprečno vrednost atributov in nato izračunali evklidsko razdalo mode vrednostmi atributov za posamezen par aktivnosti. Rezultati so prikazani v tabeli 1.

	Na vseh štirih	Klečanje	Kolesarjenje	Ležanje	Tek	Sedenje	Stanje	Tranzicija	Hoja
Na vseh štirih	0.00	2.12	3.64	2.51	6.42	3.09	2.78	2.80	4.42
Klečanje		0.00	3.08	2.30	6.06	2.0	2.16	2.67	3.69
Kolesarjenje			0.00	2.80	4.55	2.60	2.62	2.49	2.08
Ležanje				0.00	6.28	2.29	2.28	3.61	3.84
Tek					0.00	6.10	6.05	4.78	4.03
Sedenje						0.00	1.05	3.30	3.25
Stanje							0.00	3.09	3.42
Tranzicija								0.00	2.99
Hoja									0.00

Tabela 1: Razdalje med pari aktvinosti.

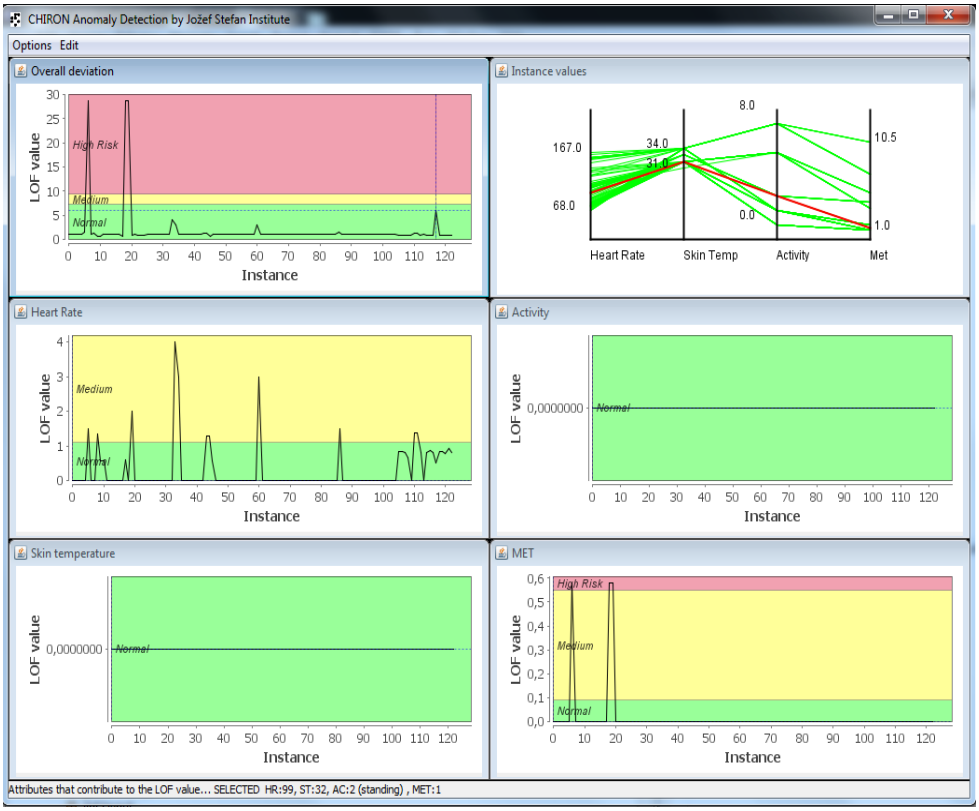
Iz tabele lahko vidimo, da sta si najdlje narazen ležanje in tek, in najbližje skupaj stanje in sedenje. Tranzicija in tek sta si tudi precej blizu. To nakazuje da je metoda računanja razdalje v tem primeru logična in prava izbira. Na ta način lahko dobimo bolj natančen izračun razdalje med instancami, ker imamo med vsakim parom nominalnih atributov unikatno razdaljo.

Ocena tveganja celostno in po parametru

Za oceno tveganja smo uporabili predhodno opisan LOF z vsemi potrebnimi modifikacijami. Uporabili smo štiri atribute srčni utrip, temperaturo kože, aktivnost ter oceno porabljene energije.
Postavljanje pragov je pokazalo, da moramo za vsako osebo postaviti svoj prag. Pragovi po osebi so prikazani v tabeli 2.

	Oseba 1	Oseba 2	Oseba 3	Oseba 4	Oseba 5
Srednje	1.08	7.34	3.49	15.00	4.31
Visoko	1.08	8.00	15.00	20.11	34.00

Tabela 2: Pragovi izraženi v vrednosti LOF.



Slika 3: Prototip LOF-a za oceno tveganja.

Vidimo lahko da so si pragovi med seboj zelo različni, kar nakazuje na to, da so lahko ti pragovi izključno inicialni. V času delovanja jih lahko spremeni ekspert, v našem primeru zdravnik. Pri osebi 1 lahko vidimo, da čiste meje med srednjim in visokim tveganjem ni.

Rezultat te metode je tudi prototip, nenadzorovan modul za končni SPO, ki bo testiran na pacientih s srčnim popuščanjem. Slika 3 prikazuje detekcijo abnormalnega stanja v prototipu. Prototip vsebuje šest panelov. Levo zgoraj je slika ocene tveganja za celoten posnetek enega človeka, na kateri se vidita dve konici ki segata v rdeče oziroma visoko tvegano območje. Spodnje štiri slike kažejo oceno tveganja po vsakem od štirih parametrov, za celoten posnetek. Vidimo lahko, da je parameter za oceno energije nenavaden v primerjavi z ostalimi parametri v instanci in da je ta v dveh zaznanih konicah vzrok za anomalijo.

Zgornja desna slika kaže razporeditev instance po vseh parametrih.

5 ZAKLJUČEK

V prispevku smo predstavili modificiran algoritem LOF za uporabo pri oceni tveganja. Tukaj je predstavljen kot eden od možnih modulov za uporabo v sistemih za podporo odločanju.

Predstavili smo metodo za računanje razdalj med vrednostmi nominalnih atributov. Rezultati izračunanih razdalj med pari aktivnosti odražajo tudi logično pomensko razdaljo med vrednostmi.

Slabost metode LOF je ta, da ko sistem dobi novi vzorec, ki je morda normalen, ta vzorec klasificira kot nenormalnega in s tem se poveča število napačno klasificiranih negativnih primerov. Prednost metode je, da ne potrebujemo označenih podatkov in domenskega eksperta.

Literatura

[1] G. F. Knolmayer, P. Mertens, A. Zeier, J. T. Dickersbach. SAP Systems for Supply Chain Management. Supply Chain Management Based on SAP Systems. pp. 73-159, Springerlink. 2009.

[2] L. Duan. Criminal Investigation DSS Based on Trust Intuition Analysis Model. Advances in Intelligent and Soft Computing. pp. 375-382. Springer. 2012.

[3] M. Krasowski. Clinical decision support of therapeutic drug monitoring of phenytoin: measured versus adjusted phenytoin plasma concentrations. BMC Medical Informatics and Decision Making. 12 (1), pp. 1-11. BioMed Central. 2012.

[4] M. M. Breunig. Quality Driven Database Mining. PhD thesis, University of Munich, 2001.

[5] CHIRON, http://www.chiron-project.eu/

[6] M. Luštrek, B. Cvetković, S. Kozina. Energy expenditure estimation with wearable accelerometers. 2012 IEEE International Symposium on Circuits and Systems (ISCAS). South Korea, 2012.