Raport Științific și Tehnic

Identificare:

Contract de finanțare nr. 124/01/04.2020
Contractor: Universitatea de Vest din Timișoara (UVT)
Cod: ERANET-CHISTERA3-DIPET
Etapa 4: 1 Dec 2021 - 30 Noi 2022
Nume etapă: Demonstrarea functionalitatii

Autori ai raportului și membrii ai proiectului:

Lect. Dr. Gabriel Iuhasz (Co-PI)
Dr. Silviu Panica
Lect. Dr. Marian Neagul
Asist.cercetare Alexandru Munteanu

Director proiect:

Prof. Dr. Dana Petcu (PI)

Pagini web:

https://dipet.hpc.uvt.ro/
https://dipet.eeecs.qub.ac.uk/

Contents

Obiectivele etapei 4

Proiectul DiPET investighează maparea dinamică și transparentă a aplicațiilor de procesare a fluxurilor de date în medii de calcul de tip Fog și Edge și folosește calculul cu trans-precizie cu scopul final de a îmbunătăți aspectele operaționale in termeni de utilizare a resurselor și consum de energie și de a îmbunătății experiența utilizatorului. Această abordare va crea noi oportunități de îmbunătățire a mai multor aspecte ale sistemelor de tip Fog și Edge, precum creșterea eficienței energetice și a performanței acestor sistemelor (se vizează o îmbunătățire a performanței de 2 ori prin trans-precizie și o îmbunătățire a energiei similare). Mai mult, DiPET își propune să simplifice radical programarea, configurarea și desfășurarea procesării fluxului, permițând programatorilor să declare restricții de plasare la resursele dinamice, controlul de planificare și trans-precizie și / sau aproximarea, toate într-o singură specificație concisă. Planificatorul va fi unul în premieră, întrucât în prezent nu există un planificator bun pentru procesarea fluxurilor de date în medii de tip Fog, cu atât mai puțin conștient de trans-precisizie, și va oferi transparență de funcționare deopotrivă pentru dezvoltatorii de aplicații și utilizatori. Sistemul DiPET va fi partajat comunității sub licențe open source și distribuit ca software reutilizabil, prin pachete software instalabile. În plus, vor fi dezvoltate cazuri de utilizare, alese dintre domeniile de aplicație relevante, centrate pe utilizator: detectarea intruziunilor în rețea și analiza video distribuită.

Obiectivele principale a etapei 4 sunt:

- finalizarea extinderii platformei EDE pe baza rezultatelor obținute din experimente.
- demonstrarea functionalitătii unor metode/operatori peste care se pot optimiza folosind metode transprecise.

Contributii previzionate se refera la WP1 (operatori streaming cu transprecizie, cu adaptare de precizie numerică, cu alegere pe baze algoritmice, mecanisme pentru acordarea algoritmilor si analiza lor), WP2 (biblioteca de operatori de streaming cu transprecizie), WP3 (monitorizarea aplicației de procesare a fluxului, modelarea performanței fluxului de procesare cu transprecizie, maparea dinamică a resurselor conștienete de transprecizie, integrare), WP4 (instrumente pentru testare, evaluare experimentală și performanță).

Rezumatul Etapei 4

În timpul etapei 4, accentul principal a fost pe munca depusă în WP1, unde echipa UVT este investigatorul principal. Acest pachet de lucru are sarcina de a crea/implementa algoritmi de streaming cu precizie reglabilă.

Remantim că în timpul primei etape, ne-am concentrat pe definirea calității, precum și a valorilor de performanță nefuncționale pentru transprecizie în contextul dispozitivelor Edge/Fog. Pentru a doua etapă ne-am concentrat pe implementarea/extinderea motorului de detectare a evenimentelor (EDE) creat inițial pentru proiectul de cercetare H2020 ASPIDE. Au fost adăugate și testate mecanisme de optimizare transparente, cu toate rezultatele pertinente incluse în ultimul raport de referință. In etapa 3 s-a lucrat cu privire la extinderea

metodelor ML pentru a include datele Graph, iar plăcile Nvidia Jetson Nano adăugate infrastructurii UVT în ultima perioadă de raportare au fost instrumentate pentru a fi utilizate în experimente transparente. Astfel, în etapa 4, cerințele de calcul și de putere ale modelelor antrenate anterior au foste testate atât pentru antrenament, cât și pentru inferență.

Rezultatele obținute în experimentele din etapele anterioare au fost publicate în articolul din revista Computer Communications (Q2 in AIS.JCR.oct2022) publicat în etapa 4 in regim open-access, iar cele din etapa 4 sunt expuse in lucrarea submisa la Data in Brief (indexată în WoS, revistă tip ESCI).

Demonstrarea funcționalității prototipurile software care au fost propuse în cadrul proiectului s-a realizat prin abordarea unor probleme reale enunțate de compania locală ETA-2U[1]. Lucrarea prezentată in conferința AINA (clasificată in CORE tip B și desfășurată în ultimele zile ale proiectului) a fost îmbunătățită față de versiunea din etapa 3 împreună cu un co-autor de la compania menționată. Acestă lucrare conține rezultate experimentale pentru detectarea și analiza ciclurilor din datele serii de timp. Deși nu este legată de munca efectuată în etapele anterioare pentru studiul de caz GuifiNet, lucrarea se bazează pe un scenariu de tip Edge/Fog: analiza ciclului de producție este una dintre condițiile prealabile pentru întreținerea predictivă în sistemele ciber-fizice.

Metodele și operatorii de preprocesare utilizați pentru detectarea ciclului au fost integrate în Motorul de detectare a evenimentelor (EDE), care a necesitat extinderea componentei de asimilare a datelor pentru a suporta surse de date suplimentare în serie de timp sub forma suportului InfluxDB[2]. Sunt acceptate ambele limbaje de interogare Flux și InfluxQL mai noi.

Gradul de atingere a rezultatelor estimate

Obiectivul științific și tehnic al etapei a fost îndeplinit în totalitate, cum vom prezenta în cele ce urmează.

Descriere științifică și tehnică

Detectarea și analiza ciclurilor

În ultimii ani, o gamă largă de tehnologii au condus la o digitizare puternică a practicilor industriale, rezultând în colectarea diferitelor date de producție. În domeniul industrial, termenul Industrie 4.0 a fost folosit pentru a descrie integrarea dintre sistemele fizice și digitale ale unui mediu de producție. Accesul la acest tip de date permite îmbunătățirea mai multor domenii cheie, de la deciziile de afaceri până la planificarea producției prin Cyber-Physical Systems (CPS) concepute în acest scop.

Una dintre cele mai importante probleme din industria 4.0 este cea a întreținerii predictive (PdM) și a detectării precoce a anomaliilor (AD). Întreținerea în sine poate reprezenta

[1] https://www.eta2u.ro/
[2] https://www.influxdata.com/

aproximativ 16–20% din toate costurile operaționale din producție. PdM își propune să îmbunătățească procesul normal de întreținere, deoarece aceste operațiuni pot fi planificate cu mult timp în avans pe baza unei estimări a duratei de viață utilă rămase a sistemului. Acest lucru duce, la rândul său, la creșterea timpului de funcționare și a disponibilității echipamentelor. O condiție prealabilă pentru PdM este capacitatea de a detecta orice tipare sau comportamente care nu sunt conforme cu așteptările. Pe scurt, avem nevoie de o formă de AD care, la rândul său, poate fi utilizată pentru identificarea pieselor defecte.

Cu toate acestea, există o problemă atunci când se aplică metode AD bazate pe Machine Learning (ML), deoarece acestea sunt bazate pe cunoștințe și pot necesita efort semnificativ atât pentru înțelegerea datelor, cât și pentru formarea modelului predictiv. În plus, aceste metode sunt adesea utilizate pe date multivariate în serie de timp, crescând și mai mult complexitatea conductei de procesare/analiza.

Ne propunem să prezentăm experimente menite să descompună atât PdM, cât și AD în sarcini de bază. Detectarea și identificarea automată a ciclului de producție se află la baza atât PdM, cât și AD în contextul Industriei 4.0. Ne propunem să arătăm cum aceste sarcini pot fi îndeplinite folosind un set de date relativ limitat. Mai mult, arătăm cum tehnicile AD cuplate cu metodele IA explicabile pot fi utilizate pentru identificarea ciclului defectuos și analiza cauzală.

Data

Datele disponibile pentru experimentele noastre constau din mai multe serii temporale. În multe CPS, fiecare dispozitiv este echipat cu mai multe tipuri de senzori, de la senzori de cuplu la microfoane. În cazul nostru avem un singur tip de senzor pentru consumul de energie. Acest lucru la prima vedere poate părea o piedică, totuși, suprainstrumentarea infrastructurii de producție poate fi costisitoare și consumatoare de timp. Instrumentele minime sunt mult mai rentabile. Experimentele detaliate în paragrafele următoare vor arăta că unele probleme pot fi încă gestionate eficient cu surse limitate de date de monitorizare.

În experimentele noastre, am identificat 3 dispozitive dintr-un flux de lucru de producție pentru care nu numai că am cunoscut ciclurile, ci și o imagine de ansamblu clară a pieselor care au fost fabricate într-un interval de timp dat. Figura 1 arată un exemplu de serie temporală pentru un dispozitiv monitorizat. Perioada de timp acoperă 10-28 mai 2022. În acest timp, 6 tipuri unice de piese au fost fabricate pe un dispozitiv cu id 355. Trebuie să rețineți că tipurile de piese sunt fabricate în funcție de nevoi. În intervalul de timp dat, unele piese sunt fabricate de mai multe ori, în timp ce altele sunt fabricate la scară limitată.

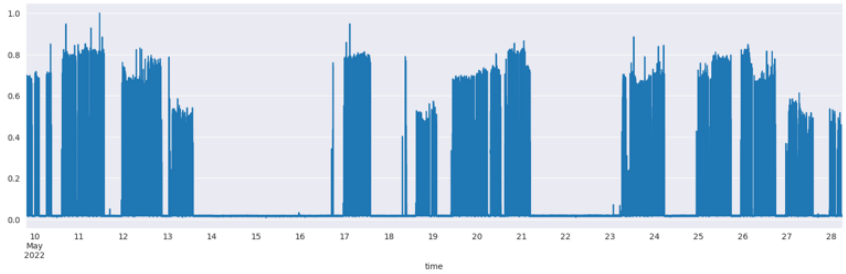


Figura 1 Serii de timp pentru un dispozitiv

Detectarea ciclului

Detectarea ciclului de producție reprezintă o condiție prealabilă pentru alte probleme, cum ar fi PdM, detectarea defecțiunilor mașinii și piesei. Prin detectarea ciclului de producție înțelegem procesul de identificare când o piesă a fost fabricată folosind un anumit dispozitiv într-un flux de lucru de producție.

Primul pas în detectarea acestor tipuri de cicluri este definirea unui model. Scopul este de a număra de câte ori se repetă acest tipar într-o anumită serie de timp și de a marca începutul și sfârșitul fiecărui tipar/ciclu.

Figura 2 prezintă exemple de model definit de utilizator utilizat pentru detectarea ciclului în cazul Dispozitivului 355. Lungimea fiecărui ciclu este importantă mai târziu pentru a remedia alinierea ciclului detectat. În cazul nostru, lungimea ciclurilor pentru dispozitivul 355 este de 88. Ar trebui să reținem că prezentăm experimentele făcute numai pe Dispozitivul 355, rezultatele pentru celelalte 2 dispozitive (Dispozitivele 359 și 369) sunt similare, astfel încât, de dragul conciziei, acestea nu sunt incluse.

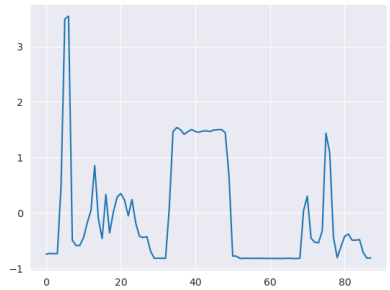


Figura 2 Model de ciclu

Definim o fereastră bazată pe lungimea modelului. Aceste ferestre sunt apoi comparate cu modelul original. Pentru această comparație pot fi utilizate diferite măsuri de distanță, cum ar fi distanțe euclidiene sau Manhattan. Cu toate acestea, aceste măsuri de distanță au mai multe dezavantaje. Pentru experimentele noastre am ales Dynamic Time Warping (DTW) ca măsură de distanță, deoarece permite compararea a două serii temporale care pot varia în viteză sau au lungimi diferite.

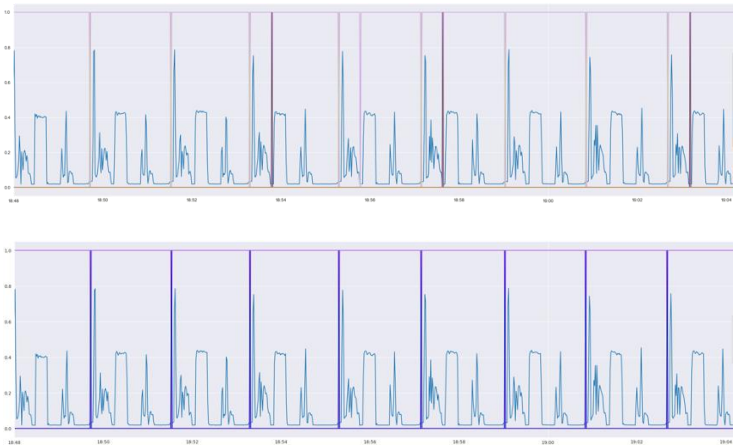


Figura 3 Detectare inițială (sus) detecție finală (jos)

Utilizam Numba[3] pentru implementarea DTW, deoarece aceasta a permis o accelerare considerabilă. Înainte de a calcula DTW, normalizarea scorului Z a fost aplicată după cum urmează:

$$x_z = (x - \mu)/\sigma \quad (1)$$

unde xz este valoarea normalizată, x este originalul, μ reprezintă media datelor și σ abaterea standard. Acest lucru duce la scorul z având o medie zero și o abatere standard de 1. Acest lucru este de dorit în cazul metodelor bazate pe ML, deoarece dacă datele de intrare sunt aproape de zero, modelele tind să convergă mai repede.

Figura 3 (top) arată rezultatele inițiale ale detectării ciclului pentru dispozitivul 355. Putem vedea că DTW a făcut o treabă bună în identificarea ciclurilor de producție, în unele cazuri există o suprapunere între ciclurile identificate. Acest lucru este cauzat de DTW în sine. Există mai multe moduri prin care putem rezolva această problemă. Mai întâi putem defini un prag pentru potrivirile detectate, deoarece ciclurile care se suprapun au invariabil un scor de similaritate mai mic. Aplicarea acestui prag riscă să elimine și ciclurile neobișnuite care pot fi utilizate pentru alte sarcini de analiză, cum ar fi PdM și detectarea defecțiunilor. Văzând că este imposibil ca două cicluri să aibă loc în același timp, definim o euristică care ia în considerare distanța dintre punctul de pornire al ultimelor cicluri și lungimea ciclului nou detectat. Astfel am definit distanța maximă dintre începutul unui ciclu și al altuia ca:

$$\Delta max = \delta p - (\mu - 4 * \sigma)$$

unde Δmax reprezintă distanța maximă, δp reprezintă lungimea modelului în timp ce μ și σ sunt aceleași ca pentru ecuația anterioară. Rezultatele pot fi văzute în Figura 3 (jos) arată rezultatele inițiale ale detectării ciclului pentru dispozitivul 355. Putem vedea că DTW a făcut o treabă bună. Ciclurile suprapuse au fost complet eliminate, în același timp fiind încă capabile să detecteze cicluri cu măsuri de distanță DTW mai mari, păstrând ciclurile potențial anormale.

[3] https://numba.pydata.org/

Gruparea ciclurilor

Odată ce am detectat cu succes ciclurile de producție, am vrut să vedem dacă putem detecta câte tipuri de cicluri avem în setul nostru de date. Aici fiecare ciclu de productie reprezinta un tip de piesa produsa. Aplicarea metodelor ML de grupare nesupravegheată pe aceste cicluri are mai multe avantaje. În primul rând, nu necesită date de antrenament pre-etichetate. Poate că acest lucru face metodele nesupravegheate mai ușor de utilizat în CPS.

Când încercăm să grupăm în mod eficient ciclurile de producție detectate, trebuie să luăm în considerare mai multe considerente. În primul rând, pentru că am folosit DTW, nu ar trebui să folosim k-means, deoarece s-ar putea să nu convergă. Media este un estimator cu cel puțin pătrat care minimizează varianța, nu distanța. În schimb, ar trebui să calculăm o matrice de distanță utilizând DTW și să aplicăm aceasta metodelor de grupare ierarhică. În al doilea rând, metodele de grupare care sunt, de asemenea, capabile să detecteze numărul de clustere distincte fără a fi nevoie să le specifice a priori și capacitatea de a face față zgomotului sau a instanțelor de date anormale este, de asemenea, o considerație importantă.

Algoritmii de grupare, cum ar fi DBSCAN și Optics, au fost promițători inițial. Optica este capabilă să facă față diferențelor mari de densități, în timp ce DBSCAN nu poate. În același timp, Optica tinde să marcheze peste 50% din cicluri ca zgomot. S-a descoperit că o metodă de grupare mai adecvată, numită HDBSCAN, produce cele mai bune rezultate. Este de fapt o versiune ierarhică a DBSCAN care nu mai folosește puncte de frontieră, ci doar puncte de bază pentru a defini un cluster.

În cazul HDBSCAN, am folosit DTW pentru a precalcula o matrice de distanță. Apoi am stabilit dimensiunea minimă a clusterului la 30, în afară de aceasta, am obținut cele mai bune rezultate cu parametrii impliciti. Clustererul rezultat a detectat 6 tipuri de ciclu unice pe lângă clusterul de zgomot.

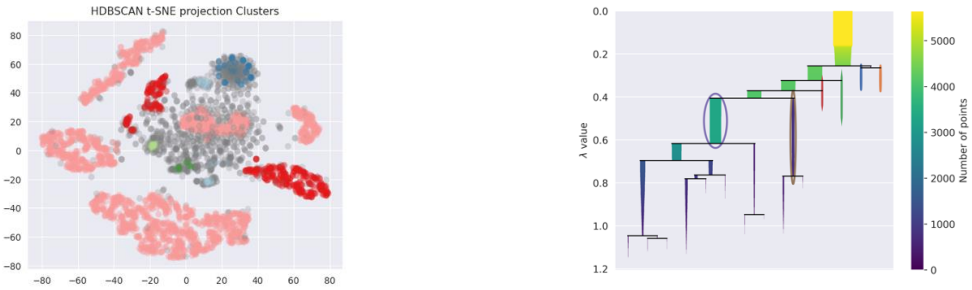


Figura 4 Clustere HDBSCAN

Figura 4 (stânga) arată o reproiectare 2D a clusterelor detectate. Am folosit t-SNE, care este o tehnică de reducere a dimensionalității neliniare care încearcă să păstreze structura locală a datelor originale. Etichetele sunt calculate folosind structura de date originală, nu cea reproiectată, folosim reproiectarea doar în scopuri de vizualizare. Figura 4 (dreapta) prezintă o dendrogramă reprezentând ierarhia clusterului pentru HDBSCAN, unde fiecare ramură reprezintă numărul de puncte dintr-un cluster. Putem vedea de fapt că unele clustere sunt considerabil mai mari decât altele.

Din cele 6 clustere detectate, unul conține mai multe puncte de date decât celelalte combinate. În plus, zgomotul este al doilea cel mai reprezentat cluster. Există mai mulți factori care contribuie la acest comportament. Putem vedea din Figura 5 că există perioade lungi în care nu s-a măsurat consumul de energie. Acestea corespund aproximativ weekend-urilor și zilelor libere. De asemenea, unele piese sunt fabricate pentru câteva zile, în timp ce altele doar pentru câteva ore. Unele piese sunt destul de asemănătoare și pot fi doar mici variații față de alte piese. Toate acestea conduc la o distribuție foarte disproporționată a clusterelor.

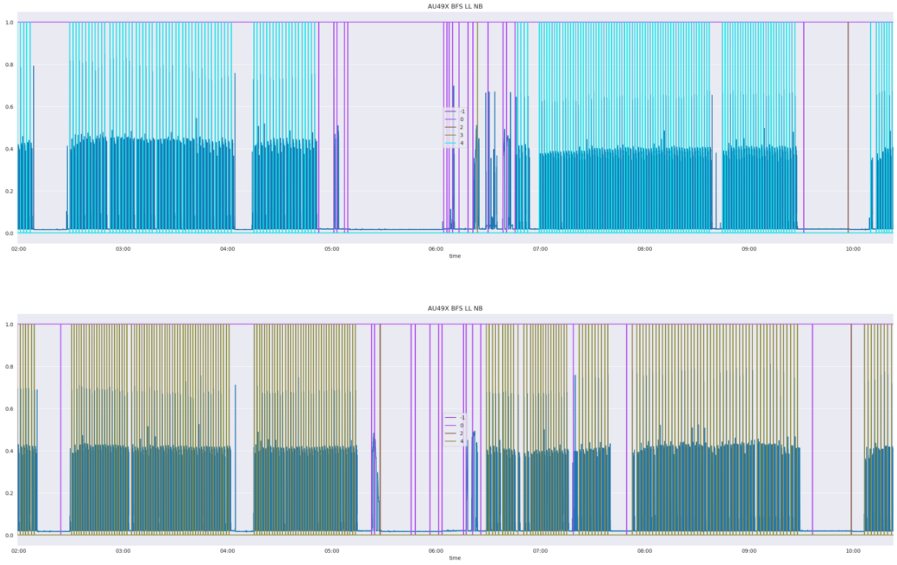


Figura 5 Exemplu de detectare a ciclului de producție

Anomalii

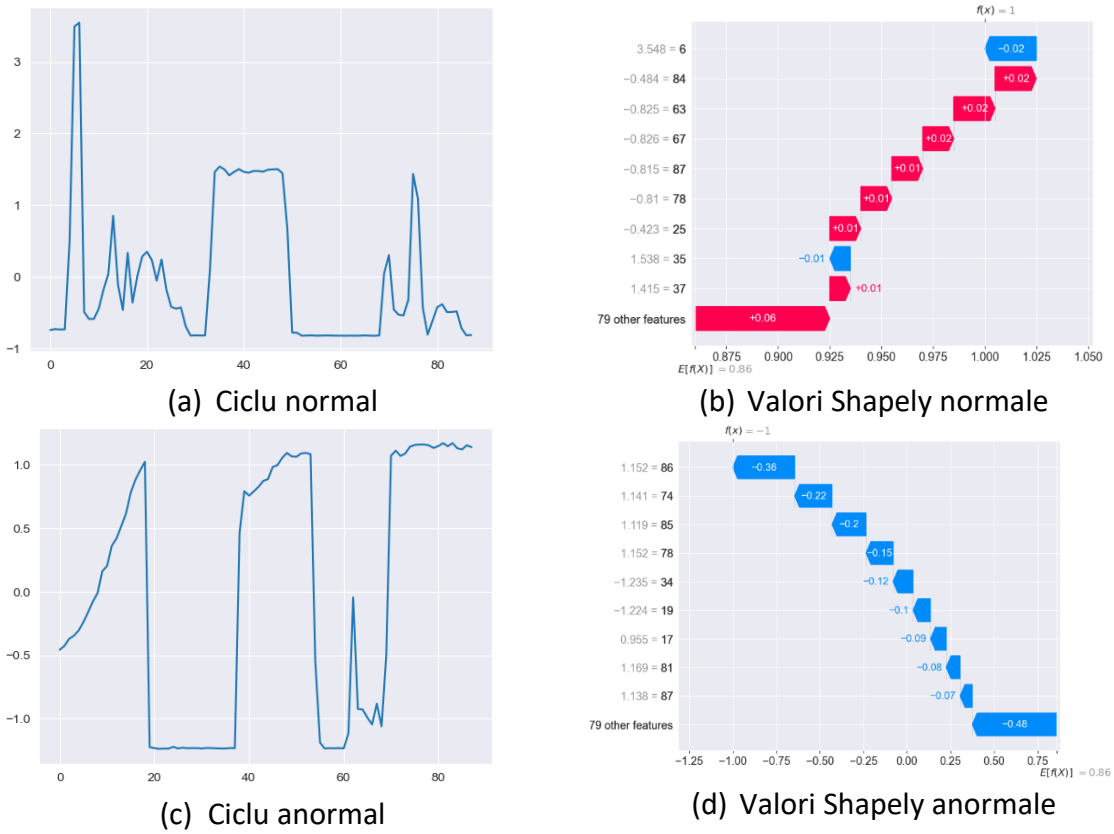
Figura 5 arată două secvențe în care știm că a fost fabricată o anumită piesă. Am putut să identificăm cu succes un tip de piesă adiacentă pentru aceste secvențe. Cu toate acestea, au fost detectate și unele cicluri anormale. Acestea apar de obicei la începutul unor secvențe lungi de cicluri normale sau, rareori, la sfârșit.

Am decis să aruncăm o privire mai concentrată asupra cazurilor anormale. Pentru aceasta am ales un algoritm specializat de detectare a anomaliilor. Isolation Forest (IF) este un algoritm de detectare a valorii aberante care este construit din mai mulți arbori de izolare. Explorează subspații aleatorii din date, fiecare arbore explorează diferite împărțiri, explorând astfel subspații locale aleatorii. Notarea se face prin calificarea cât de ușor este să găsești un subspațiu local de dimensionalitate scăzută pentru un eveniment izolat dat.

Am antrenat un model IF pe toate ciclurile detectate. Am selectat un factor de contaminare de 0,05 și numărul de estimatori la 100. Am detectat un total de 597 de anomalii. Instanțele anormale sunt destul de izolate din regiunile mai dens, similar cu ceea ce Figura 4 (stânga) afișat pentru HDBSCAN. Scopul nostru cu IF este să identificăm acele cicluri care sunt adevărate anomalii. Sperăm să înțelegem de ce aceste cicluri particulare sunt marcate ca anomalii.

Aceste cazuri anormale au fost în mare parte aliniate cu zgomotul detectat de HDBSCAN. În continuare, am vrut să analizăm în continuare aceste cicluri anormale folosind valori Shapely,

astfel încât să putem vedea ce parte a unui ciclu este cea mai influentă în marcarea acestuia ca anormal. Figura 6 arată o comparație cu normalul. Figura 6 a și ciclu anormal. Figura 6 c împreună cu o diagramă în cascadă care ilustrează clasamentul derivat al valorii Shapley a fiecărui punct din ciclu, Figura 6b și respectiv Figura 6c. Ultimele două cifre arată dacă o caracteristică împinge un ciclu spre a fi o instanță normală (valori pozitive marcate cu roșu) sau un ciclu anormal (valori negative marcate cu albastru). Din aceste cifre putem observa că în cazul ciclurilor anormale punctele 86, 74, 85, 78 sunt cele mai influente. Putem deduce din aceasta că sfârșitul ciclului este semnificativ diferit de unul normal.



(a) Ciclu normal

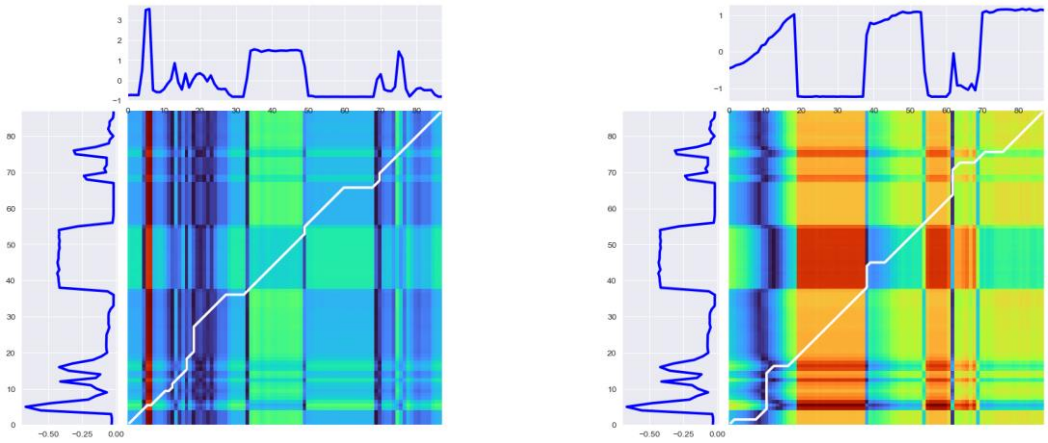
(b) Valori Shapely normale

(c) Ciclu anormal

(d) Valori Shapely anormale

Figura 6 Comparația dintre ciclurile normale și anormale

Figura 7 arată calea optimă de aliniere între două cicluri, linia din matricea costurilor reprezintă calea minimă de la primul până la ultimul punct de ciclu. Figura 7a reprezintă un ciclu normal și Figura 7b un ciclu anormal. Pașii de analiză prezentati în Figura 6 și Figura 7 pot fi utilizați pentru analiza cauzei principale. Putem spune cu un grad ridicat de încredere de ce un anumit ciclu este marcat ca o anomalie și care parte a ciclului este cea mai influentă în această decizie.



(a) Aliniere normală (b) Aliniere anormală

Figura 7 Căi de aliniere în raport cu modelul original

Diseminare

1. Articole

Articolele submise în etapa 3 au fost acceptate și modificate conform recenziilor și publicate în etapa 4:

- Llorenc Cerda-Adalbert, *Gabriel Iuhasz*, Gabriele Gemmi,"Anomaly Detection for Fault Detection in Wireless Community Networks Using Machine Learning", Computer Communications, Volume 202, 2023, Pages 191-203, ISSN 0140-3664, Elsevier (Q2/AIS), https://doi.org/10.1016/j.comcom.2023.02.019
- *Gabriel Iuhasz, Silviu Panica*, Alecsandru Duma, "Cycle Detection and Clustering for Cyber Physical Systems" AINA-2023, International Conference on Advanced Information Networking and Applications (rank B in CORE), Lecture Notes in Networks and Systems, vol 655. Springer, Cham, https://doi.org/10.1007/978-3-031-28694-0_10

În etapa 4 a fost submisă și este în evaluare următoarea:

- Llorenc Cerda-Adalbert, *Gabriel Iuhasz*, „Dataset for Anomaly Detection in a Production Wireless Mesh Community Network", submis la Data in Brief[4] (indexat in catalogul ESCI mentinut de Clarivate)

2. Livrabile

Proiectul DIPET este un proiect european colaborativ. Livrabilul elaborat de către echipa UVT (disponibil în platforma EVOC, ca anexă la raport) este:

[4] https://www.sciencedirect.com/journal/data-in-brief

- Livrabilul D1.3: Final version of transprecise streaming operators and hyper-parameter tuning methodology (rezultat al WP1), 30.03.2023, autor: *Gabriel Iuhasz*

Buget

Fondurile etapei 4 au fost cheltuite exclusiv pe salarii, regie si audit. Nu au fost efectuate deplasari, toate ședințele proiectului european au fost realizate on-line. Prezentarea la conferința AINA din Brazilia a fost de asemenea susținută online.

Concluzii

Obiectivele etapei au fost îndeplinite în totalitate. Experimentele descrise mai sus dovedesc validatea conceptelor introduse în proiect.

PI / UVT

Prof. Dr. Dana Petcu