Zaznavanje anomalij v proizvodnem procesu preko obdelave kompleksnih dogodkov

[1]Tadej Krivec, [1]Dejan Gradišar, [1]Miha Glavan, [2]Gašper Mušič
[1]Institut Jožef Stefan, Jamova 39, Ljubljana
[2]Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, Ljubljana
tadej.krivec@ijs.si, dejan.gradisar@ijs.si, miha.glavan@ijs.si, gasper.music@fe.uni-lj.si

Fault detection in production plant using complex event processing

One of the main objectives for every production plant is to minimize their down time. At the same time they have to produce high-quality products with the fastest rate possible. Currently installed supporting IT solutions generate more and more data, which remains stored without effective use. There is a high demand for solutions that can process and draw useful information from this data. This article presents a solution for fault detection in production line using data pre-processing methods PCA (Principal Component Analysis) and DPCA (Dynamic Principal Component Analysis). Statistical analysis with hypothesis testing using Hotelling and SPE (Squared Predicted Error) test statistics is presented. Because of the large amount and diverse properties of the data, fault detection was realized using complex event processing which enables static queries on streams of data. Complex event processing uses a powerful constructor (windowing), which enables distributed and fast processing of data. Off-line fault detection algorithm was realized in Python programming language. On-line detection was developed on the Microsoft StreamInsight platform. The result was visualized on the dashboard realized with Microsoft Power BI. The solution was demonstrated on a Tennessee Eastman process simulation model.

Kratek pregled prispevka

V današnji dobi je pomembno, da proizvodnja teče neprekinjeno, hkrati pa zagotavlja produkcijo kakovostnih izdelkov v najkrajšem možnem času. Zaradi hitre rasti količine podatkov, ki se zajamejo v proizvodnji, se je povečala tudi potreba po razvoju rešitev, ki lahko take podatke obdelajo. Predstavljena je rešitev za zaznavanje anomalij v proizvodnem procesu z metodama preobdelave podatkov PCA (ang. Pricipal Component Analysis) in DPCA (ang. Dynamic Pricipal Component Analysis) ter s sklepanjem na podlagi hipotez s statistikama Hotelling in SPE (ang. Squared Predicted Error). Zaradi velike količine in raznovrstnosti podatkov je bilo zaznavanje anomalij realizirano z obdelavo kompleksnih dogodkov, ki definira statične poizvedbe (ang. query) na dinamičnih podatkih (ang. data stream). Močen element obdelave kompleksnih dogodkov so časovna okna, ki omogočajo porazdeljen in hiter način obdelave podatkov. Nesprotni del algoritma zaznavanja anomalij je bil realiziran v programskem jeziku Python, sprotni pa na platformi Microsoft StreamInsight. Rezultat je prikazan v obliki nadzorne plošče v okolju Microsft Power BI. Rešitev smo demonstrirali na simulacijskem modelu procesa Tennessee Eastman.

1 Uvod

V današnji dobi tehnologije se v proizvodnji zahteva produkcija kvalitetnih izdelkov. Ti morajo biti narejeni s čim manjšimi stroški v čim krajšem času. Podjetja si ne morejo privoščiti okvar na napravah, saj to pomeni, da pride do zastojev v proizvodnji. Zastoji lahko pomenijo precejšnjo izgubo denarja.

Ker so cene senzorjev in shranjevanje podatkov postali cenovno zelo dostopni, se je močno povečala količina podatkov, ki se tipično zajema v proizvodnji. S količino podatkov so zrasle tudi potrebe po podatkovni analizi. Koncept zajema podatkov iz različnih naprav, ki so potem povezane s svetovnim spletom imenujemo internet stvari (ang. internet of things). Zajeti podatki se shranjujejo v podatkovne baze v podjetju ali pa v oblaku. Pametne tovarne shranjujejo širok nabor proizvodnih parametrov, ki jih kasneje obdelajo za spremljanje in optimizacijo proizvodnega procesa.

V nadaljevanju bo predstavljen pristop zaznavanja anomalij v obliki obdelave kompleksnih dogodkov na platformi Microsoft StreamInsight. Platforma je sposobna sprejemati podatke iz različnih virov in realnočasno obdelovati senzorske tokove podatkov. Zaznavanje anomalij je prikazano na orodju za poslovno analitiko Microsoft Power BI [1].

2 Zaznavanje anomalij

V tem poglavju je predstavljen algoritem zaznavanja anomalij. V prvem koraku algoritma se iz podatkov razvijejo le pomembne značilke. V drugem koraku je definiran nesprotni del algoritma, ki rezultira v modelu, ki temelji na obratovanju v normalnem delovanju. V tretjem koraku se definira sprotni del algoritma, ki primerja sprotni vzorec z normalnim delovanjem in zaznava anomalije.

2.1 Obdelava podatkov s PCA in DPCA

Ker je zajetih proizvodnih podatkov lahko zelo veliko, je smiselno v prvem koraku podatke obdelati in izluščiti le vplivne značilke.

Metoda PCA je linearna transformacija, ki zmanjša dimenzije originalnega prostora. Išče smer največje variance in jo projicira v podprostor z enakim ali manjšim številom dimenzij. DPCA je nadgradnja PCA, kjer vhodnemu regresorju dodamo še zakasnjene vektorje značilk. Ti zakasnjeni vektorji značilk predstavljajo dinamični sistem.

Naj ima proces m merjenih značilk in število meritev N. Potem je vhodni regresor definiran z matriko $X \in \mathcal{R}^{N\times m}$. Metoda PCA je zelo občutljiva na skaliranje podatkov, zato je bil vhodni vektor standardiziran. Standardiziran regresor je označen kot $Z \in \mathcal{R}^{N\times m}$.

V naslednjem koraku je bila izračunana kovariančna matrika K, ki predstavlja kovarianco med pari merjenih spremenljivk. Elementi matrike $K \in \mathcal{R}^{j\times k}$ so definirani z enačbo (1).

$$\sigma_{jk} = \frac{1}{N-1}\sum_{i=1}^{N}(z_j^i - \mu_j)(z_k^i - \mu_k) \qquad (1)$$

Nad kovariančno matriko se izvede singularni razcep po enačbi (2).

$$K = P\Lambda P^T \qquad (2)$$

Matrika P predstavlja transformacijsko matriko [2]. Vsebuje lastne vektorje kovariančne matrike K. Matrika Λ je diagonalna matrika, ki vsebuje lastne vrednosti kovariančne matrike K. V transformaciji v nov prostor se obdrži zadostno količino lastnih vrednosti oz. lastnih vektorjev, da se ohrani večino razložene variance. To število označimo z l. Enačbi (3a) in (3b) prikazujeta delitev na obdržane (P_{PC}, Λ_{PC}) in preostale (P_{RES}, Λ_{RES}) lastne vektorje in lastne vrednosti [3].

$$\Lambda = \begin{bmatrix} \Lambda_{PC} & 0 \\ 0 & \Lambda_{RES} \end{bmatrix} \qquad (3a)$$

$$P = [P_{PC} \quad P_{RES}] \qquad (3b)$$

Vhodni regresor Z transformiramo v podprostor Z' s transformacijo, ki jo opisuje enačba (4).

$$Z' = ZP_{PC} \tag{4}$$

2.2 Nesprotni del algoritma

V nesprotnem delu algoritma se izračunajo kritične vrednosti testnih statistik (T^2_{KR}, SPE_{KR}) na podlagi podatkov, ki so zajeti v normalnem delovanju procesa. Skupaj s transformacijsko matriko (P), diagonalno matriko lastnih vrednosti (Λ), vektorjema povprečja učne množice ($\bar{\mu}$) in standardnega odklona učne množice (σ) tvorijo model zaznavanja anomalij.

2.2.1 Testna statistika Hotelling

Testna statistika Hotelling je multivariatna porazdelitev, proporcionalna porazdelitvi Fisher–Snedecor (F). Je posplošitev statistike Student, ki se uporablja za testiranje multivariatnih podatkov. Kritično vrednost testne statistike Hotelling T_{KR} s stopnjo tveganja $\alpha = 0.01$ in prostostnimi stopnjami l in $N - l$, kjer je N število vhodnih podatkov in l število obdržanih lastnih vrednosti po obdelavi podatkov s PCA, določimo z enačbo (5) [3].

$$T^2_{KR} = \left(\frac{l(N-1)(N+1)}{N^2 - lN}\right) F_{\alpha,l,N-l} \tag{5}$$

2.2.2 Testna statistika SPE

Testna statistika SPE je definirana z izračunom kvadratne predikcije napake. Kritična vrednost testne statistike SPE se izračuna z enačbami (6a), (6b), (6c).

$$SPE_{KR} = \Theta_1 \left(\frac{c_\alpha \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2}\right)^{\frac{1}{h_0}} \tag{6a}$$

$$\Theta_i = \sum_{i=l+1}^{m} \lambda_j^i, i = 1, 2, 3 \tag{6b}$$

$$h_0 = 1 - \frac{2\Theta_1 \Theta_3}{2\Theta_2^2} \tag{6c}$$

Lastno vrednost pri obdelavi z metodo PCA predstavlja λ_j, c_α pa predstavlja spremenljivko standardne porazdelitve, ki pripada zgornjemu $1 - \alpha$ percentilu [3].

2.3 Sprotni del algoritma

Sprotno testno statistiko izračunano na podlagi primerjave vektorja povprečja učne množice in vzorca, ki pride v sistem in ga želimo prepoznati. Vektor živega vzorca je označen z *x,* vektor z pa je njegova standardizirana vrednost. Anomalija se zazna, ko sprotni testni statistiki presegata svoji pripadajoči kritični vrednosti s stopnjo tveganja $\alpha = 0.01$. Sprotna testna statistika Hotelling se izračuna po enačbi (7), sprotna testna statistika SPE pa po enačbi (8) [4].

$$T^2 = z^T P_{PC} \Lambda_{PC}^{-1} P_{PC}^T z \tag{7}$$

$$SPE = e^T e = (1 - PP^T) z \tag{8}$$

3 Rezultati zaznavanja anomalij na arhivskih testnih podatkih

Zaznavanje anomalij je bilo validirano na arhivskih testnih podatkih modela Tenneesee Eastman.

3.1 Arhivski testni podatki

Komponente realnega procesa (kot so npr. kemijska kinetika, procesi in operacijski pogoji) so bile modificirane zaradi zaščite podjetja, sicer pa model temelji na realnem proizvodnem procesu [5]. S simulacijskega modela (Simulink) so bile zajete procesne meritve v normalnem obratovanju ($I_{DV(1)}$). Zajeti so bili tudi podatki pri nastanku 20 različnih anomalij ($I_{DV(2)}, \dots, I_{DV(21)}$), kjer anomalija nastane po 8 urah. Senzorji merijo različne dogodke in stanja na napravah, kot so dovodi v reaktor, pritisk v reaktorju, nivo reaktorja, temperatura hladilne tekočine, nivo tekočin itn.

3.2 Kriterij validiranja

Metoda zaznavanja anomalij s testnima statistikama je bila validirana s kriterijema FDR (ang. Fault Detection Rate) in FAR (ang. False Alarm Rate), ki sta definirana z enačbama (9) in (10).

$$FDR = 100 \times \frac{N(J > J_{th} | f \neq 0)}{N(f \neq 0)} \tag{9}$$

$$FAR = 100 \times \frac{N(J > J_{th}|f=0)}{N(f=0)} \quad (10)$$

$N(J > J_{th}|f \neq 0)$ je število vzorcev, kjer testna statistika J presega mejo kritične vrednosti J_{th} za čase, kjer se je motnja dejansko zgodila. $N(J > J_{th}|f = 0)$ je število vzorcev, kjer testna statistika J presega mejo kritične vrednosti J_{th} za čase kjer se motnja dejansko ni zgodila. $N(f \neq 0)$ in $N(f = 0)$ sta števili vseh vzorcev, kjer se anomalija je oz. ni zgodila [4].

3.3 Rezultati obdelave podatkov s PCA in DPCA

Pri metodah obdelave podatkov PCA in DPCA je bilo izbranih 9 oziroma 17 največjih lastnih vrednosti oz. lastnih vektorjev. Število zakasnitev pri metodi DPCA je bilo 3. Kritične vrednosti testnih statistik so bile določene na podlagi tveganja $\alpha = 0.01$.

3.4 Rezultat zaznavanja anomalij

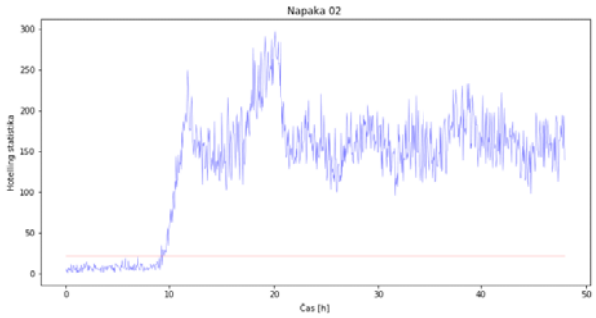
Testna statistika SPE se je v večini primerih izkazala za boljšo. Rezultati FDR in FAR za PCA in DPCA so prikazani v tabelah 1 in 2. Obe metodi odlično prepoznavata anomalije $I_{DV(1)}$, $I_{DV(2)}$, $I_{DV(6)}$, $I_{DV(8)}$, $I_{DV(12)}$, $I_{DV(13)}$, $I_{DV(14)}$, $I_{DV(17)}$ in $I_{DV(18)}$ kot prikazujeta sliki 1 in 2. Z rdečo sta prikazani kritični vrednosti testnih statistik, z modro pa sprotni vrednost testnih statistik. Obe metodi se zelo slabo izkažeta pri zaznavanju anomalij $I_{DV(3)}$, $I_{DV(4)}$, $I_{DV(9)}$, $I_{DV(15)}$ *in* $I_{DV(19)}$. Metodi srednje dobro prepoznavata anomalije $I_{DV(5)}$, $I_{DV(7)}$, $I_{DV(10)}$, $I_{DV(11)}$, $I_{DV(16)}$, $I_{DV(20)}$ in $I_{DV(21)}$. Anomaliji $I_{DV(16)}$, $I_{DV(19)}$ predstavljata klasični napaki, kjer se metodi ne odrežeta najbolje. Prepoznavanje anomalije $I_{DV(16)}$ je prikazano na slikah 3 in 4. Testna statistika Hotelling se ne izkaže vedno za slabšo metodo, zato lahko sklepamo, da bi bilo zaznavanje anomalij najbolje realizirano s kombinacijo obeh testnih statistik. V tem primeru bi motnjo zaznali, ko bi katerakoli od testnih statistik presegala svojo kritično mejo.

Tabela 1: FDR in FAR za razpoznavanje motenj s testno statistiko Hotelling in SPE ter metodo PCA

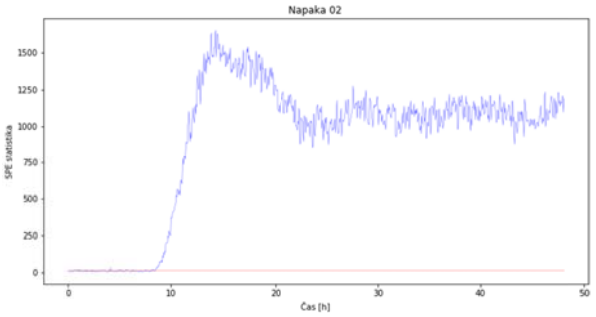
	Hotelling		SPE	
Anomalija	FDR [%]	FAR [%]	FDR [%]	FAR [%]
$I_{DV(1)}$	99.25	0.00	99.75	1.25
$I_{DV(2)}$	96.75	1.25	98.75	2.50
$I_{DV(3)}$	6.25	0.00	5.88	3.12
$I_{DV(4)}$	1.75	1.88	4.12	1.88
$I_{DV(5)}$	25.50	0.00	27.00	5.00
$I_{DV(6)}$	100.00	0.00	100.00	5.00
$I_{DV(7)}$	43.00	0.00	34.50	1.25
$I_{DV(8)}$	96.50	0.00	95.88	3.12
$I_{DV(9)}$	4.25	0.62	5.25	0.62
$I_{DV(10)}$	46.62	11.25	47.00	7.50
$I_{DV(11)}$	19.25	0.62	47.00	2.50
$I_{DV(12)}$	98.88	1.88	94.75	3.12
$I_{DV(13)}$	94.12	0.62	95.25	2.50
$I_{DV(14)}$	98.75	0.62	100.00	1.88
$I_{DV(15)}$	7.88	1.25	6.62	5.00
$I_{DV(16)}$	34.25	0.00	23.88	3.12
$I_{DV(17)}$	83.62	15.00	95.62	5.00
$I_{DV(18)}$	89.25	0.62	90.62	4.38
$I_{DV(19)}$	7.25	1.25	24.12	3.12
$I_{DV(20)}$	32.12	0.00	50.38	3.75
$I_{DV(21)}$	39.62	0.00	52.62	1.88

Tabela 2: FDR in FAR za razpoznavanje motenj s testno statistiko Hotelling in SPE ter metodo DPCA

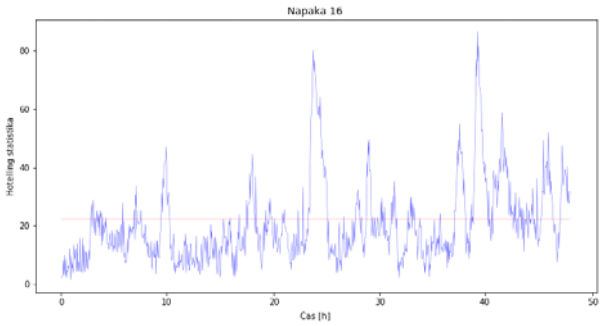
	Hotelling		SPE	
Anomalija	FDR [%]	FAR [%]	FDR [%]	FAR [%]
$I_{DV(1)}$	99.25	1.25	99.75	1.25
$I_{DV(2)}$	98.37	1.88	95.86	1.25
$I_{DV(3)}$	5.52	1.25	7.90	1.25
$I_{DV(4)}$	1.76	2.50	3.76	6.25
$I_{DV(5)}$	27.48	1.25	25.60	1.88
$I_{DV(6)}$	100.00	1.25	99.87	1.88
$I_{DV(7)}$	44.54	0.62	34.88	1.25
$I_{DV(8)}$	97.62	1.25	91.59	6.88
$I_{DV(9)}$	6.02	1.25	6.78	1.88
$I_{DV(10)}$	46.93	11.25	73.40	7.50
$I_{DV(11)}$	47.43	1.88	10.92	1.88
$I_{DV(12)}$	99.25	1.88	88.71	5.00
$I_{DV(13)}$	94.60	1.88	95.23	2.50
$I_{DV(14)}$	100.00	0.62	91.47	0.00
$I_{DV(15)}$	8.28	2.50	14.81	0.62
$I_{DV(16)}$	35.01	0.62	35.76	3.12
$I_{DV(17)}$	94.60	15.00	93.73	6.25
$I_{DV(18)}$	90.21	1.25	90.84	10.00
$I_{DV(19)}$	16.81	0.62	14.93	3.12
$I_{DV(20)}$	39.27	1.25	62.48	1.88
$I_{DV(21)}$	45.29	0.62	64.49	1.25



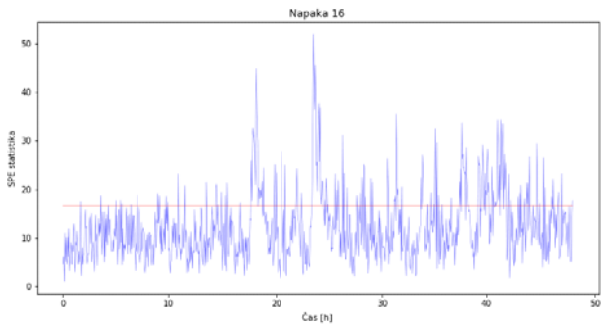
Slika 1: Zaznavanje anomalije $I_{DV(2)}$ s testno statistiko Hotelling.



Slika 2: Zaznavanje anomalije $I_{DV(2)}$ s testno statistiko SPE.



Slika 3: Zaznavanje anomalije $I_{DV(16)}$ s testno statistiko Hotelling.



Slika 4: Zaznavanje anomalije $I_{DV(16)}$ s testno statistiko SPE.

4. Rezultati realnočasnega zaznavanja anomalij na simulacijskem modelu

Nesprotni del zaznavanja anomalij je bil realiziran v programskem jeziku Python. Parametri modela so bili shranjeni v podatkovno bazo SQL Server. Sprotno zaznavanje anomalij je bilo izvedeno na platformi za obdelavo kompleksnih dogodkov StreamInsight v programskem jeziku C#.

Sprotni podatki se pošiljajo s simulacijskega modela Tennessee Eastman [5] na posrednika podatkov RabbitMQ. RabbitMQ doda abstrakcijo med virom podatkov in uporabnikom, ki te podatke sprejme. Tako je mogoče sistem narediti še bolj porazdeljen.

Strežnik Streamnsight bere podatke nesprotnega dela iz baze Microsoft SQL Server. Naročen je na tok podatkov iz RabbitMQ. Tok podatkov ustrezno razdeli na manjša časovna okna in znotraj njih agregira ustrezne značilke za zaznavanje anomalij. Arhitektura sprotnega zaznavanja anomalij je prikazana na sliki 5.

StreamInsight rezultat zaznavanja anomalij pošlje direktno na orodje za poslovno informatiko PowerBI, ki rezultat ustrezno prikaže na nadzorni plošči kot prikazuje slika 6.

4 Zaključek

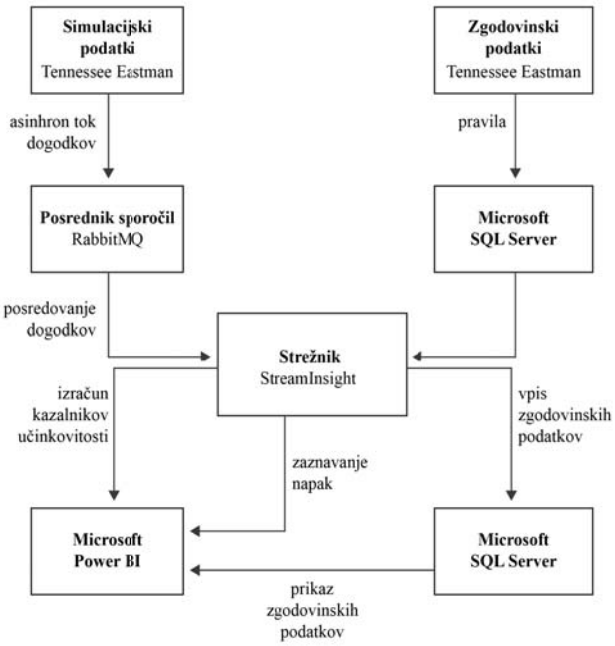
Metodi za obdelavo podatkov PCA in DPCA se v rezultatih občutno ne razlikujeta. Zaradi tega je bila izbrana enostavnejša in računsko manj zahtevna metoda PCA.

Testna statistika SPE v povprečju deluje bolje od testne statistike Hotelling, vendar ne vedno. Zaradi tega je optimalno, da se za zaznavanje anomalij izračuna obe statistiki ter sproži alarm v primeru, da katerakoli presega svojo kritično mejo. Metodi določenih anomalij nista sposobni zaznati. To sledi tudi iz omejitve, da obe metodi analizirata samo trenutno vrednost testne statistike. Izboljšava bi bila zaznavanje anomalije iz več zaporednih vzorcev. Tako bi tudi zmanjšali vpliv šuma na zaznavanje.

Zaradi zahteve po realnočasnem zaznavanju anomalij, se je sprotni del izvedel na platformi

za obdelavo kompleksnih dogodkov Microsoft StreamInsight. Platforma se je izkazala za zelo uporabno saj je čas razvoja za uporabnike jezika C# zelo kratek.

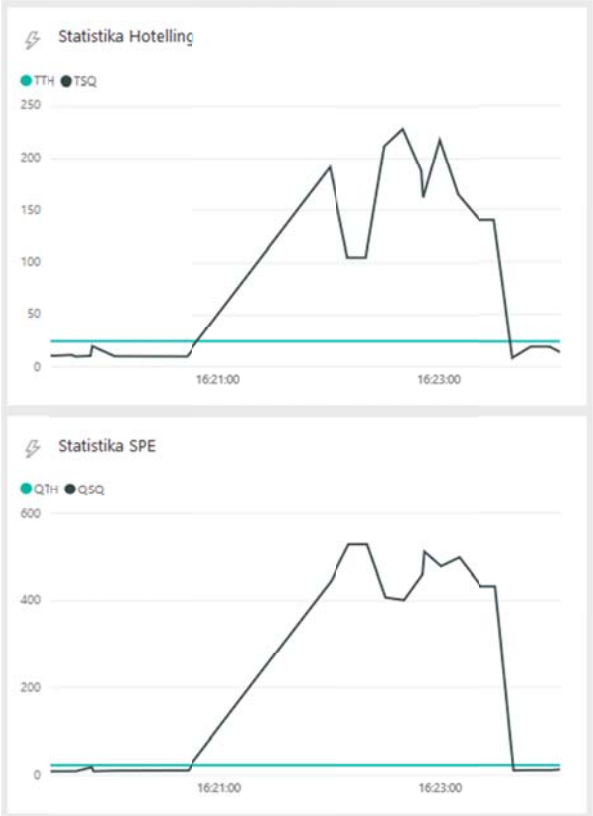
Dobra lastnost platforme je, da omogoča razvoj rešitev na lastni programski opremi (ang. on premise). Slabost platforme pa je, da Microsoft ne razvija več posodobitev. Microsoft pospešeno razvija oblačno storitev Azure Stream Analytics, ki prav tako podpira obdelavo kompleksnih dogodkov, vendar deluje v oblaku (ang. cloud).



Slika 5: Arhitektura sprotnega zaznavanja anomalij.

5 Zahvala

Delo je bilo izvedeno v sklopu programa GOSTOP, ki ga delno financirata Republika Slovenija – Ministrstvo za izobraževanje, znanost in šport ter Evropska Unija – Evropski sklad za regionalni razvoj in v sklopu nacionalnega raziskovalnega programa Sistemi in vodenje, P2-0001.



Slika 6: Nadzorna plošča zaznavanja anomalij na orodju Microsoft PowerBI.

6 Literatura

[1] Tadej Krivec, Dejan Gradišar, Miha Glavan, Gašper Mušič. Obdelava kompleksnih dogodkov pri spremljanju proizvodnega procesa. Revija za fluidno tehniko, avtomatizacijo in mehatroniko Ventil, 25:46-53, 2019.

[2] S. Raschka. Python Machine Learning. Packt Publishing Ltd., 2015.

[3] Shen Yin, Steven X. Ding, Adel Haghani, Haiyang Hao, and Ping Zhang. A comparison study of basic data-driven fault diganosis and proces monitoring methods on the benchmark Tennessee Eastman process. Journal Of Process Control, 22:1567-1581, 2012.

[4] Tiago J. Rato and Marco S. Reis. Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR). Chamometrics and Intelligent Laboratory Systems, 125:101-108, 2013.

[5] J. J. Downs and E. F. Vogel. A Plant-Wide Industrial Process Control Problem. Computers chem. Engng., 17(3):245-255, 1993.