Pregled nekaterih metod na področju analize časovnih vrst

Žiga Stržinar[1,2], Boštjan Pregelj[1], Igor Škrjanc[2]

[1]*Odsek za sisteme in vodenje, Institut "Jožef Stefan", Ljubljana*
[2]*Fakulteta za elektrotehniko, Univerza v Ljubljani*
E-pošta: ziga.strzinar@ijs.si

Overview of some methods in the field of time series analysis

Time series analysis is gaining importance in recent years, as more and more temporal data are being collected. This paper presents the challenges faced in the field. Three analysis steps are further explored: time series representation, time series distance measures and time series clustering. For each step, relevant methods are presented and evaluated. This paper aims to provide an insight into the field and expose some common concepts, challenges and solutions. The methods chosen here, are a basis for further work in the monitoring of dynamic processes.

1 Uvod

Vedno večja razširjenost senzorjev, pojav interneta stvari (*ang. Internet of Things* - IoT), dostopnost orodij za hrambo velikih količin podatkov, pojav nosljive elektronike, digitalizacija poslovnih procesov in drugi tehnološki dosežki v zadnjih letih so omogočili izgradnjo velikih podatkovnih zbirk časovno označenih podatkov - časovnih vrst (ang. *time series*). Področje analize, obdelave in odkrivanja znanj v teh podatkih zato pridobiva na pomenu na različnih področjih življenja: zdravstvo (nosljiva elektronika, elektrokardiogrami, itd.), finančna industrija (nihanja cen delnic, tečajev in drugih finančnih instrumentov), transport (podatki o prometu), proizvodna dejavnost, itd.

Namen tega prispevka je predstaviti pristope primerne za analizo časovnih vrst. Podani so temeljni problemi s katerimi se srečujemo, predstavljene so nekatere pogosto uporabljene metode.

2 Analiza časovnih vrst

Področje analize časovnih vrst je zelo široko, saj dandanes časovne vrste izvirajo iz zelo različnih virov, problemi, ki se rešujejo pa so zelo raznoliki. Literatura ([1], [8]) izpostavlja več namenov analize časovnih vrst. Pogosta naloga je iskanje podobnosti (ang. *similarity search*) in pa ujemanje sekvenc (ang. *subsequence matching*). Tipičen cilj analize časovnih vrst je tudi zaznavanje odstopanj (ang. *anomaly detection*), kjer se znotraj časovne vrste iščejo nepričakovani dogodki, odstopanja od običajnega delovanja. Na področju sta pomembna postopka rojenje (ang. *clustering*) in klasifikacija (ang. *classification*). Pri prvem je želja zbirko časovnih vrst urediti v roje (tudi gruče, ang. *clusters*). Pri klasifikaciji pa se novo časovno vrsto razvršča v enega od vnaprej določenih razredov. Daljše časovne vrste je včasih potrebno segmentirati v krajše segmnete, ki so bolj primerni za nadaljnjo obdelavo. Pri segmentaciji je ključno vprašanje kako določiti prlomne točke.

Pri delu se ukvarjamo s problemom zaznavanja dogodkov na proizvodnih linijah. Na voljo so časovno označene meritve z opazovanega sistema, želja pa je iz teh meritev razpoznati različne faze delovanja ali korake v proizvodnem procesu. Za doseganje tega cilja je potrebno poznavanje metod analize časovnih vrst predstavljeno v tem prispevku.

V naslednjih poglavjih so predstavljene pogoste metode za doseganje omenjenih ciljev. V poglavju 3 so predstavljene metode za predstavitev časovnih vrst. Poglavje 4 se osredotoča na mere razdalje primerne za delo s časovnimi vrstami, poglavje 5 pa obravnava metode za rojenje.

3 Metode za predstavitev časovnih vrst

Pri obdelavi časovnih vrst je opravka z velikimi količinami podatkov (t.i. *big data*) zato je bil predlagan pristop, ki surove podatke (izvirno časovno vrsto) najprej preslika v aproksimirano predstavitev. Komprimirano predstavitev je možno shraniti v pomnilnik. V drugem koraku je dan problem okvirno rešen, podan je en ali več predlogov za končno rešitev - kandidatov. Te kandidate se v zadnjem koraku preveri z dostopom do izvirnih podatkov. Celoten pristop temelji na ustreznosti prvega koraka - dobri predstavitvi podatkov. V tem poglavju so zato predstavljeni uveljavljeni in obetavni pristopi za predstavitev časovnih vrst.

Vzorčenje (ang. *sampling*) je najenostavnejši pristop, pri katerem se časovno vrsto namesto z vsemi vzorci predstavi samo z vsakim n-tim vzorcem.

PAA je alternativa vzorčenju. Metoda poznana tudi kot *Piecewise Aggregate Approximation* ([14]) kot nov vzorec vzame povprečno vrednost n zaporednih vzorcev izvirne časovne vrste. Število vzorcev po PPA je enako, kot se dobi z vzorčenjem, vendar je metoda manj občutljiva na šum in osamelce.

PLA (ang. *Piecewise Linear Approximation*), je metoda predstavljena v [13]. Gre za pristop, ki segment nefiksne dolžine opiše s parametri linearne funkcije. Časovno vrsto se najprej segmentira v K ne-enako-dolgih segmentov, končni točki vsakega segmenta sta elementa osnovne časovne vrste. Konec enega segmenta je začetek naslednjega. Algoritem začne s segmenti dolžine 3. Za vsako premico se izračuna normirana napaka po formuli $e_i = \frac{\sum_{m=1}^{j} d_m^2}{j}$, kjer je j število vzorcev v segmentu, i je indeks segmenta. Izračuna se standardna deviacija e_i čez vse segmente in se označi kot B_k (k je število segmentov). Algoritem se nadaljuje tako, da se združita dva sosednja segmenta, število segmentov se s tem zmanjša za 1. Algoritem se ponavlja, dokler ne ostane le en segment. V vsakem koraku se izbere združitev segmentov (ang. *merging*), ki bo rezultirala v najmanjšem B_k v naslednji iteraciji. Na koncu se uporabi tisto število segmentov k, kjer je B_k najmanjši, torej kjer je napaka najbolj konstantna čez vse segmente.

PIP algoritem predstavljen v [6, 9] časovno vrsto predstavi kot zaporedje vizualno pomembnih točk (*Perceptually Important Points*). Za časovno vrsto P algoritem iskanje točk začne z označitvijo robnih točk kot pomembnih. V naslednji iteraciji se poišče tista točka na P, ki ima največjo razdaljo od obstoječih pomembnih točk. V vseh naslednjih iteracijah se kot pomembna označi točka, ki ima največjo razdaljo do dveh sosednjih pomembnih točk. Algoritem se ponavja dokler vse točke na P niso označene kot pomembne. Točke P so po zadnji iteraciji urejene po pomembnosti.

SAX je v zadnjem času pogosta metoda za predstavitev časovne vrste s simboličnim zapisom. SAX je kratica za *Symbolic Aggregate approXimation)* in je predstavljena v [16]. SAX zahteva, da se na časovni vrsti najprej uporabi PAA. Algoritem v nadaljevanju sloni na predpostavki, da ima osnovna časovna vrsta Gaussovo normalno porazdelitev. Glede na to porazdelitev se nato določi nabor prelomnih točk med katerimi je integral funkcije gostote verjetnosti (PDF) enak. Vsakemu intervalu med dvema prelomnima točkama pripada en simbol uporabljene abecede. Časovno vrsto se predstavi z zaporedjem simbolov glede na to v kateri interval 'pade' posamezen segment dobljen s PAA.

Pristrižena predstavitev (ang. *clipped representation*) je pristop za zelo zgoščeno predstavitev časovne vrste, kjer se le ta predstavi kot zaporedje bitov (npr. 10011000), kjer znak 1 pove, da je vrednost večja od povprečja časovne vrste, 0 pa, da je manjša oz. enaka.

Predstavitev skozi rojenje V [8] je naštetih več primerov, kjer avtorji opisujejo izračun značilk oziroma predstavitev časovnih vrst z rojenjem.

Nekateri avtorji segmentirajo časovno vrsto ter vsak segment predstavijo z enostavno obliko (npr. strmina). Segmente se nato roji, s tem se dobi za vsak segment identifikator roja kateremu pripada. Zaporedje identifikatorjev se uporabi kot nova predstavitev časovne vrste.

Glede na dinamično naravo časovnih vrst je obetaven pristop tudi uporaba samorazvijajočih algoritmov rojenja (angl. *evolving clustering* [17]), ki so prilagojeni za spremljanje in analizo toka podatkov, kar je pogosta zahteva pri analizi časovnih vrst.

4 Mere razdalje

V tradicionalnih podatkovnih zbirkah so mere razdalje med vzorci definirane na osnovi točnega ujemanja med dvema vzrocema, na področju analize in obdelave časovnih vrst pa izbira mere še bolj odvisna od cilja ([8, 1]), ki je lahko:

- podobnost v istih časovnih trenutkih, iskanje časovnih vrst, ki se istočasno enako obnašajo?
- podobnost po obliki, neodvisno od absolutne vrednosti časa (premik o časovni osi)
- podobnost po modelu, kjer se iščejo časovne vrste pri katerih v ozadju veljajo enaka pravila

Evklidska razdalja se v analizi časovnih vrst uporablja na dva načina: 1) za direktno primerjavo dveh enako dolgih časovnih vrst (podobnost v istih časovnih trenutkih) 2) za primerjavo transformiranih zapisov časovnih vrst, na primer primerjava parametrov modelov, DFT ali DWT koeficientov.

DTW je na področju analize časovnih vrst pogosto uporabljena mera, podana v [5]. Glavna značilnost je podana že z imenom metode - dinamično časovno popačenje (ang. *Dynamic Time Warping*). Po tej metodi se za dve časovni vrsti Q in P dolžin n in m zgradi matrika n x m, kjer vsak element predstavlja razdaljo med dvema odčitkoma časovnih vrst. Tipično se uporabi evklidska razdalja [8]. Glede na vnaprej določene omejitve (začetna in končna točka, monotonost) se določi optimalna pot - t.i. pot popačenja (ang. *warping path*).

LCSS , razdalja najkrajšega skupnega izseka (ang. *Longest Common Subsequence*) je definirana kot razmerje med najdaljšim skupnim zaporedjem v dveh časovnih vrstah in dolžino celotne časovne vrste [11]. Parameter izračuna razdalje je toleranca – meja pod katero sta dve točki označeni, da se ujemata. Izračun je podoben kot pri DTW.

MVM je obetavna mera, ki obljublja izboljšavo delovanja napram DTW. Prednost pred DTW so boljši rezultati, ko sta primerjani časovni vrsti različno dolgi, metoda namreč dovoljuje preskakovanje ne-dodeljenih 'repov' časovnih vrst.

5 Rojenje

Rojenje je postopek za nenadzorovano analizo množice vzorcev z razdelitvijo v podmnožice - roje, na način, da

so si pripadniki istega roja med seboj bolj podobni, kot so podobni pripadnikom drugih rojev.

V literaturi ([1, 7, 10, 12, 15]) se metode za rojenje delijo v več skupin:

Delitvene metode (ang. *Partitioning methods*) so metode, ki vse vzorce glede na neko mero razdajle razdelijo v posamezne roje. Globalni optimum je težko zagotoviti, običajno se te metode zadovoljijo že z lokalnim minimumom. Med te metode prištevamo k-Means, k-Medoids, Gustafson-Kessel, fuzzy c-means ([7] našteje številne nadgradnje)...

k-Means (težišče roja je povprečje vzorcev), k-Medoids (težišče roja je centralni vzorec) – oba sta primerna za sferične roje in majhne do srednje velike množice. Algoritma sta pogosto uporabljena tudi kot osnova za druge algoritme ali za primerjavo uspešnosti ([11]). Algoritma sta primerna tudi za uporabo z različnimi merami razdalje. Najpogostejša mera je evklidska razdalja, v posebnih primerih, pa so potrebne druge mere, pri časovnih vrstah je denimo pogosta kombinacija k-means in DTW razdalje.

Gustafson-Kessel je adaptivni algoritem mehkega rojenja primeren tudi za ne-sferične roje [15].

Delitvene metode so zaradi enostavnosti in dobrih rezultatov prav tako zelo pogosti na področju analize časovnih vrst. k-Means je pogosta metoda s katero se primerjajo ostale metode.

Hierarhične metode ustvarjajo hierarhično strukturo primernih rojitev. Obstajata dva pristopa - združevanje (ang. *agglomerative*) ter razdiranje, eljenje (ang. *divisive*). Za hierarhično rojenje je bistvena izbira kriterija za razdelitev oziroma združevanje rojev, saj združenega roja tipično ni več mogoče razdeliti, razdeljenega roja pa ne združiti. Metode lahko temeljijo na razdalji, gostoti ali zveznosti. Izpeljanke tega pristopa se uporabljajo tudi za rojenje časovnih vrst [1, 15]. Pogosto navedena prednost teh metod je zmožnost intuitivne vizualizacije poteka rojenja.

Bistvena je izbira primerne mere razdalje, pogosto je uporabljena DTW (poglavje 4). V novejših člankih se pogosto omenja tudi uporaba SAX za predstavitev in povezano mero razdalje.

Hierarhično rojenje časovnih vrst na težave naleti pri dolgih časovnih vrstah in zelo velikih podatkovnih množicah [1].

Metode na osnovi gostote ne temeljijo neposredno na razdalji med vzorci temveč na gostoti. S tem se izognejo težavi ostalih pristopov, ki so osredotočeni na sferične roje. Te metode so primerne za roje arbitrarnih oblik [10, 7]. Te metode povečujejo roj dokler je gostota okolice (ang. *neighborhood*) večja od nekega praga (ang. *threshold*). Metode so primerne za izločevanje šuma in osamelcev [10]. Primera teh metod sta DBSCAN in OPTICS.

Mrežne metode prostor kvantizirajo v mrežo. Vse nadaljnje operacije se dogajajo v novem prostoru, kar pohitri procesiranje [10]. Ta pristop se lahko uporabi v kombinaciji z ostalimi.

Metode z uporabo modelov za vsak roj privzamejo nek model. Pogosta metoda je SOM - Samo-organizirajoče mreže (ang. *Self-organizing maps*). Tako kot mnogi drugi pristopi SOM zahteva enotno dolžino vzorcev. Raziskujejo se tudi pristopi s polinomskimi modeli, modeli Gaussovih mešanic, verigami Markova itd.

Samorazvijajoče metode (ang. *evolving*) imajo posebno lastnost, da se roji prilagajajo (učijo, razvijajo) toku podatkov sproti, ko ti prihajajo. Algoritmi so zasnovani tako, da omogočajo odmik podatkov od običajne delovne točke. Odmik je lahko postopen (ang. *concept drift*) ali hipen (ang. *concept shift*) [18].

Samorazvijajoči algoritmi pridejo v poštev na več področjih: rojenje, iskanje vzorcev, zaznavanje anomalij, diagnostika [18]. Izvedenih je bilo že več konkretnih implementacij samorazvijajočih algoritmov npr. samorazvijajoč krmilnik [2], samorazvijajoča indentifikacija [4, 3], prepoznavanje dogodkov pri vožnji [17], itd.

Stohastična optimizacija Poleg že opisanih skupin metod za rojenje je za rojenje možno uporabiti tudi metode stohastične optimizacije. Gre za metode namenjene optimizaciji poljubne kriterijske funkcije s stohastičnim pristopom, pogosto z mnogokratnimi evalvacijami kriterijske funkcije pri različnih vhodnih parametrih. Glede na cilje rojenja se lahko zasnuje kriterijska funkcija, ki oceni razdelitev v roje. Z uporabo metod stohastične optimizacije je nato možno najti (glede na kriterijsko funkcijo) primerno razdelitev. Med metodami, ki se lahko uporabijo najdemo PSO (ang. *Particle Swarm Optimization*), genetske algoritme in druge.

5.1 Posebnosti rojenja časovnih vrst

Namen rojenja časovnih vrst je enak kot pri običajnem rojenju, razlika je le v vzorcih, ki se jih roji. Pri običajnem rojenju se rojijo vzorci sami, pri časovnih vrstah pa je to lahko ne-praktično. Časovno vrsto lahko sicer obravnavamo kot visokodimenzionalen vzorec (število dimenzij je dolžina časovne vrste). To lahko, odvisno od primera, rezultira v zelo visoko-dimenzionalnem problemu, ki lahko povzroča težave napačno izbranemu algoritmu rojenja. Prav tako tak pristop ne zaobjame specifike časovnih vrst (pomen oblike, amplitudnega zamika itd.) ter težje obravnava časovne vrste različnih dolžin.

Po [15] se algoritmi za rojenje časovnih vrst delijo:

1. Algoritmi, ki prilagodijo klasičen algoritem tako, da uporabijo mero podobnosti/razdalje, ki je primerna za časovne vrste. Ti algoritmi uporabljajo surovo časovno vrsto, 'klasičen' algoritem ter prilagojeno mero.

2. Algoritmi, ki pretvorijo časovno vrsto v vektor značilk (postopek izluščitve značilk - ang. *feature extraction*) ter uporabijo klasičen algoritem s klasično mero razdalje.
3. Algoritmi, ki kot vektor značilk uporabijo identificirane parametre modela (ang. *model-based approach*).

V [1] avtor predlaga še eno skupino metod za rojenje časovnih vrst:

Več-koračne metode (ang. *multi-step methods*), kjer se časovna vrsta predstavi z več resolucijami, pri vsakem koraku (resoluciji), se izvede rojenje. V [1] je podan pregled in kritična ocena predlaganih večnivojskih metod.

6 Zaključek

V prispevku so predstavljene ključne metode v uporabi na področju analize časovnih vrst. Metode so komentirane glede na primernost za delo s časovnimi vrstami. Obdelane so tri skupine metod: metode za predstavitev časovnih vrst, metode za merjenje razdalj med njimi ter metode za njihovo rojenje.

Povzeto je, zakaj je posebna predstavitev časovnih vrst sploh potrebna. Predstavljenih je osem metod, od najenostajvnejše - z vzorčenjem časovne vrste, do naprednejših algoritmov PLA, PIP in SAX.

V poglavju o merah razdalje je pojasnjena razlika v pojmovanju razdalje med dvema časovnima vrstama in običajnim pojmovanjem razdalje med vektorjema vzorcev, kot ga poznamo na drugih področjih. Predstavljenih je več metod, med njimi pogosto uporabljena DTW.

Področje rojenja časovnih vrst je obravnavano skozi razdelitev na šest skupin metod. Za vsako skupino so opisane glavne značilnosti, uporabnost ter razširjenost pri delu s časovnimi vrstami. Omenjene so posebnosti rojenja časovnih vrst ter omenjeni posebni skupini več-koračnih metod ter samorazvijajočih metod, ki obetata dobre rezultate na tem področju.

Prispevek nudi vpogled v področje ter služi kot osnova za nadaljnje delo na področju. V nadaljevanju avtorji nameravamo tukaj predstavljene metode uporabiti na praktičnem primeru nadzora delovanja dinamičnih procesov.

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.

Literatura

[1] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015.

[2] Goran Andonovski, Bruno Sielly Jales Costa, Sašo Blažič, and Igor Škrjanc. Robust evolving controller for simulated surge tank and for real two-tank plant. *at-Automatisierungstechnik*, 66(9):725–734, 2018.

[3] Goran Andonovski, Gašper Mušič, Saso Blažič, and Igor Škrjanc. On-line evolving cloud-based model identification for production control. *IFAC-PapersOnLine*, 49(5):79–84, 2016.

[4] Goran Andonovski, Gašper Mušič, Sašo Blažič, and Igor Škrjanc. Evolving model identification for process monitoring and prediction of non-linear systems. *Engineering applications of artificial intelligence*, 68:214–221, 2018.

[5] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[6] Fu-Lai Chung, Tak C Fu, Robert Luk, and V Ng. Flexible time series pattern matching based on perceptually important points. 2001.

[7] D. R. Edla, D. Tripathi, V. Kuppili, and R. Cheruku. Survey on clustering techniques. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 696–703, 2018.

[8] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[9] Tak-chung Fu, Fu-lai Chung, Robert Luk, and Chak-man Ng. Representing financial time series based on data point importance. *Engineering Applications of Artificial Intelligence*, 21(2):277–300, 2008.

[10] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[11] Octavian Lucian Hasna and Rodica Potolea. Time series—a taxonomy based survey. In *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 231–238. IEEE, 2017.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning, Second Edition*, volume 1. Springer series in statistics New York, 2009.

[13] Eamonn Keogh. Fast similarity search in the presence of longitudinal scaling in time series databases. In *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, pages 578–584. IEEE, 1997.

[14] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, 3(3):263–286, 2001.

[15] T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[16] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2):107–144, 2007.

[17] Igor Škrjanc, Goran Andonovski, Agapito Ledezma, Oscar Sipele, Jose Antonio Iglesias, and Araceli Sanchis. Evolving cloud-based system for the recognition of drivers' actions. *Expert Systems with Applications*, 99:231–238, 2018.

[18] Igor Škrjanc, Jose Antonio Iglesias, Araceli Sanchis, Daniel Leite, Edwin Lughofer, and Fernando Gomide. Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: a survey. *Information Sciences*, 490:344–368, 2019.