

Laiko eilutes transformuojančių į vaizdus, siekiant taikyti CNN klasifikavimo uždaviniui spręsti, metodų tyrimas

Rugilė Vasaitytė

Vilniaus universitetas,
Matematikos ir informatikos fakultetas
Informatikos institutas
Didlaukio g. 47, Vilnius
rugile.vasaityte@mif.stud.vu.lt

Santrauka. Tyrimas nagrinėja laiko eilučių transformavimą į vaizdus, siekiant taikyti konvoliucinius neuroninius tinklus (angl. *Convolutional Neural Networks*, CNN) klasifikavimui. Apžvelgiami trys metodai: Gramo kampinis sumavimo laukas (angl. *Gramian Angular Summation Field*, GASF), Gramo kampinis skirtumų laukas (angl. *Gramian Angular Difference Field*, GADF) bei Markovo perėjimo laukai (angl. *Markov Transition Field*, MTF). CNN modelis pritaikytas šešiams skirtingiems duomenų rinkiniams, transformuotiems minėtais metodais. Rezultatai rodo, kad metodų efektyvumas priklauso nuo duomenų pobūdžio, tačiau GADF dažniau pasiekė geresnius rezultatus. Pateikiamos rekomendacijos, kaip pasirinkti tinkamiausią transformaciją klasifikavimo tikslumui gerinti.

Raktiniai žodžiai: laiko eilutės, giliojo mokymosi metodai, konvoliuciniai neuroniniai tinklai, vaizdų klasifikacija, laiko eilučių transformacija, GASF metodas, GADF metodas, MTF metodas.

1 Įvadas

Laiko eilutė – tai duomenų taškų seka, išdėstyta nuosekliai pagal laiką. Laiko eilučių klasifikavimas (angl. *Time Series Classification*, TSC) yra užduotis, kurios tikslas yra priskirti laiko eilutei klasę, t. y. kintamąjį iš aibės, nusakantį, kuriai grupei ar kategorijai priklauso nagrinėjama laiko eilutė, pavyzdžiui, normalus ar anomalinis, teigiamas ar neigiamas pavyzdys. TSC praktikoje gali būti pritaikomas įvairiose srityse – diagnozuojant širdies ligas [1] ar kibernetiniame saugume [2]. Todėl veiksmingų ir efektyvių algoritmų kūrimas laiko eilučių klasifikavimui yra aktyvi tyrimų sritis. Štai [3] ir [4] straipsniuose atskleidžiama, kad giliojo mokymosi metodai, kuriuos naudojant automatiškai išgaunami sudėtingi duomenų požymiai, dažnai pranoksta

klasikinius statistinius ir mašininio mokymosi metodus sprendžiant laiko eilučių klasifikavimo uždavinius. Šis pranašumas ryškus dirbant su daug pavyzdžių ir požymių turinčiais duomenų rinkiniais, kai požymių inžinerija tampa nepraktiška.

Visgi, norint taikyti giliojo mokymosi metodus, pavyzdžiui, konvoliucinius neuroninius tinklus, apdoroti ir klasifikuoti laiko eilutes, susiduriama su realių duomenų rinkinių heterogeniškumu, t. y. laiko eilutės gali būti skirtingo ilgio, turėti skirtingą struktūrą, triukšmingumo lygį. Tuomet laiko eilutes reikia tinkamai apdoroti. Siekiant šio tikslo, galima laiko eilutes transformuoti į vaizdus – taip užtikrinama, kad visi įvesties duomenys turėtų vienodą formatą. Egzistuoja keli tokie transformuojantys metodai, pavyzdžiui, Gramo kampinis sumavimo laukas, Gramo kampinis skirtumų laukas bei Markovo perėjimo laukai [5]. Šie metodai laiko eilutes transformuoja į fiksuoto dydžio dvimačius vaizdus, kuriuos galima analizuoti naudojant giliojo mokymosi metodus.

Taigi, šio tyrimo **tikslas** – įvertinti, kuris laiko eilutes į vaizdus transformuojantis metodas leidžia gauti tikslesnį vaizdų klasifikavimo rezultatą. Siekiant tyrimo tikslo, bus atlikti šie žingsniai: pasirinkti duomenų rinkiniai bus transformuoti į vaizdus naudojant GASF, GADF ir MTF metodus. Vėliau šie vaizdai bus klasifikuojami konvoliuciniu neuroniniu tinklu. Modelio rezultatui bus analizuojami ir statistiškai įvertinti, kad būtų nustatyta, kuris transformacijos metodas užtikrina tiksliausią klasifikaciją.

2 Tyrimo eiga

Toliau aprašytame tyrime buvo siekiama įvertinti laiko eilučių klasifikavimo tikslumą, po to, kai buvo pritaikyti laiko eilutes į vaizdus transformuojantys metodai: GASF, GADF ir MTF. Tyrimo schema pateikta 1 pav. Procesas prasideda nuo duomenų rinkinio atrinkimo, atliekamas jų paruošimas ir padalijimas į mokymo, validavimo ir testavimo aibes. Laiko eilutės transformuojamos į vaizdus, po to mokymo ir validavimo duomenys naudojami neuroninio tinklo mokymui. Galiausiai modelis testuojamas ir renkami klasifikavimo rezultatų stebėjimai, kurie analizuojami statistiniais metodais.

Duomenys ir jų paruošimas

Buvo pasirinkti šeši skirtingų tipų duomenų rinkiniai: „GunPoint“ [6], „ECG200“ [7], „ECG5000“ [8], „Wafer“ [7], „TwoLeadECG“ [8] ir „Earthquakes“ [9]. Kiekvie-

na laiko eilutė rinkinyje turi atitinkamą žymę, kuri naudojama prižiūrimam mokymui. Šie duomenys parinkti taip, kad sudarytų tam tikrą įvairiapusį rinkinį: jie parinkti iš sensorių (SENSOR), elektrokardiogramų (ECG), žmogaus veiklos atpažinimo (HAR) tipo duomenų rinkinių. Be to, eilutės turi skirtingus ilgius, triukšmingumo lygį, mokymui panaudotas skirtingas pavyzdžių kiekius.

Laiko eilutės transformuotos į vaizdus naudojant *Python pyts* biblioteką, taikant numatytuosius parametrus. Transformuotų eilučių pavyzdžius galima matyti 1 pav. Transformuotos laiko eilutės su atitinkamomis klasėmis buvo suskirstytos į mokymo, validavimo ir testavimo aibes santykiu 70:15:15 atitinkamai. Galiausiai kiekvieno vaizdo pikselių reikšmės buvo normalizuotos į $[0, 1]$ intervalą.

1 lentelė. Naudojamų duomenų rinkinių informacija.

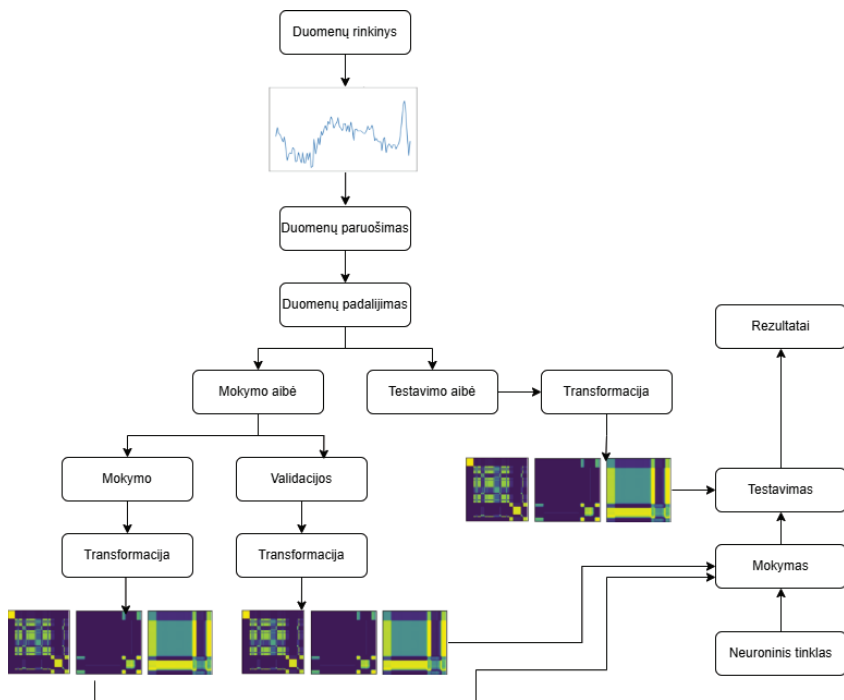
Duomenų rinkinys	Mokymui panaudotų pavyzdžių kiekis	Klasių skaičius
GunPoint	50	2
ECG200	100	2
ECG5000	3500	5
Wafer	1000	2
TwoLeadECG	813	2
Earthquakes	322	2

Modelio architektūra

CNN prasideda įvesties sluoksniu, kuris priima duomenis su nustatytu aukščiu, pločiu ir kanalų skaičiumi. Tuomet naudojami du konvoliucinių sluoksnių blokai, kurie apima konvoliucinį sluoksnį su aktyvacijos funkcija *ReLU*, maksimalų telkimo sluoksnį ir atsitiktinio praretinimo (angl. *dropout*) sluoksnį, kuris naudojamas siekiant sumažinti permokymo riziką. Po šių blokų naudojamas ištiesinimo sluoksnis, kuris duomenis transformuoja į viematį vektorių, atsitiktinio praretinimo sluoksnis ir pilnai jungūs sluoksniai, kurių pirmasis turi 96 neuronus su *ReLU* aktyvacija. Paskutinis sluoksnis yra išėjimo sluoksnis, turintis tiek neuronų, kiek yra klasifikuojamame rinkinyje klasių. Naudotas *Adam* optimizatorius, *sparse categorical crossentropy* nuostolių funkcija ir 20 mokymosi epochų. Stebima vertinimo metrika – klasifikavimo tikslumas testavimo fazėje.

Statistikos

Modelis buvo apmokytas su visais duomenų rinkiniais, kurie buvo transformuoti visais transformuojančiais metodais, kiekvienam eksperimentui buvo fiksuoti rezultatų stebėjimai. Statistiniam vertinimui naudota vienfaktorė ANOVA ir Tukey HSD testas, siekiant nustatyti, ar stebėti skirtumai tarp rezultatų, gautų taikant skirtingus transformuojančius metodus, statistiškai reikšmingi.



1 pav. Eksperimentinės dalies darbo eigos diagrama.

3 Rezultatai

Transformuotų laiko eilučių klasifikavimo rezultatai rodo, kad transformavimo metodo parinkimas daro įtaką klasifikavimo tikslumui, tačiau šis poveikis priklauso nuo konkretaus duomenų rinkinio. Štai klasifikuojant transformuotą „GunPoint“ duomenų rinkinį GASF metodas statistiškai

reikšmingai lenkia tiek GADF, tiek MTF metodus (vidurkio skirtumas 0,066 ir 0,082 atitinkamai). Tarp GADF ir MTF reikšmingo skirtumo nėra. Šiam rinkiniui labiausiai tinkamas transformuojantis metodas yra GASF. Tuo tarpu klasifikuojant „ECG200“ duomenų rinkinį GADF metodas nežymiai lenkia GASF metodą, tačiau ne statistiškai reikšmingai ($p = 0,059$). Tačiau tiek GADF, tiek GASF yra statistiškai reikšmingai pranašesnis už MTF (vidurkių skirtumas 0,047 ir 0,025 atitinkamai). Šiam rinkiniui transformuoti labiau tinka GASF ir GADF. Klasifikuojant „ECG5000“ duomenų rinkinį, GADF metodas pasirodė statistiškai reikšmingai pranašesnis už MTF metodą (vidurkių skirtumas 0,006, $p < 0,001$), taip pat ir GASF metodas statistiškai reikšmingai lenkia MTF (vidurkių skirtumas 0,005, $p < 0,001$). Tačiau tarp GADF ir GASF reikšmingo skirtumo nenustatyta ($p = 0,590$). Aukščiausias klasifikavimo tikslumo vidurkis pasiektas naudojant GADF (0,951), šiek tiek atsilieka GASF (0,950), o mažiausias rezultatas fiksuotas taikant MTF (0,945). Šiam duomenų rinkiniui labiausiai tinkami transformavimo metodai yra GADF ir GASF. Atkreiptinas dėmesys, kad transformuoto „Wafer“ duomenų rinkinio klasifikavimo tikslumui skirtingi metodai (MTF, GASF, GADF) reikšmingos įtakos nedaro. ANOVA testo statistikos $F = 1,006$, $p = 0,374$ rodo, jog metodų poveikis tikslumui yra nežymus. Visgi, didžiausias klasifikavimo vidurkis buvo gautas taikant GASF metodą, mažiausias MTF. Panašus rezultatas fiksuotas ir „Earthquakes“ duomenų rinkinyje – klasifikavimo tikslumui skirtingi metodai (MTF, GASF, GADF) statistiškai reikšmingos įtakos nedaro. Galiausiai, klasifikuojant „TwoLeadECG“ duomenų rinkinį, GADF metodas pasirodė statistiškai reikšmingai pranašesnis tiek už GASF (vidurkių skirtumas 0,018, $p = 0,005$), tiek už MTF (vidurkių skirtumas 0,015, $p = 0,024$). Tuo tarpu tarp GASF ir MTF reikšmingo skirtumo nenustatyta ($p = 0,838$). Aukščiausias klasifikavimo tikslumo vidurkis pasiektas naudojant GADF (0,998), žemesni rezultatai fiksuoti GASF (0,979) ir MTF (0,983) metodais. Šiam duomenų rinkiniui tinkamiausias transformavimo metodas yra GADF.

Apibendrinti rezultatai pateikti 2 lentelėje. Sakytina, kad yra statistiškai reikšmingų skirtumų tarp metodų (MTF, GASF, GADF), kurie daro įtaką klasifikavimo tikslumui, kai transformuojami ir klasifikuojami skirtingi duomenų rinkiniai. GASF ir GADF metodais suklasifikuotų transformuotų laiko eilučių klasifikavimo tikslumas dažniausiai lenkia MTF metodu transformuotų laiko eilučių klasifikavimo tikslumą.

2 lentelė. Vidutiniai klasifikavimo tikslumai naudojant skirtingus metodus.

Duomenų rinkinys	GASF tikslumas	GADF tikslumas	MTF tikslumas	Statistiškai reikšmingas skirtumas pagal ANOVA ir Tukey HSD testus
GunPoint	0,936	0,870	0,854	Taip – GASF > GADF, GASF > MTF
ECG200	0,861	0,883	0,836	Taip – GASF > MTF, GADF > MTF
ECG5000	0,950	0,951	0,944	Taip – GASF > MTF, GADF > MTF
Wafer	0,998	0,997	0,997	Ne
TwoLeadECG	0,980	0,997	0,983	Taip – GADF > MTF, GADF > GASF
Earthquakes	0,8	0,8	0,8	Ne

4 Išvados

Remiantis rezultatais galima teigti, kad laiko eilučių klasifikavimo tikslumas priklauso tiek nuo duomenų rinkinio, tiek nuo pasirinkto transformuojančio laiko eilutės į vaizdus metodo – apibendrintos generalizacijos apie metodus daryti negalima.

Visgi, sprendžiant laiko eilučių klasifikavimo uždavinį su prižiūrimu CNN mokymusi galima padaryti šias išvadas:

1. Jei tiriant duomenų rinkinį buvo nustatytas statistiškai reikšmingas metodo poveikis klasifikavimo tikslumui (tokie duomenų rinkiniai buvo keturi), tuomet arba GASF, arba GADF buvo pranašesnis už MTF.
2. Tiriant keturis iš šešių duomenų rinkinius nebuvo nustatytas statistiškai reikšmingas skirtumas tarp GASF ir GADF metodais transformuotų laiko eilučių klasifikavimo tikslumo ($p > 0,05$). Natūralu – metodų prigimtis gana panaši. Viena išimtis yra „GunPoint“ duomenų rinkinys: GASF metodas statistiškai reikšmingai pranašesnis už GADF (vidurkių skirtumas 0,066). Šis duomenų rinkinys turėjo mažiausiai mokymo pavyzdžių iš tirtųjų. Kita išimtis yra „TwoLeadECG“: GADF metodas statistiškai reikšmingai pranašesnis už GASF (vidurkių skirtumas 0,018).
3. Kai kurie rinkiniai, pavyzdžiui, „Wafer“ gerai klasifikuojami nepriklausomai nuo transformuojančio metodo. Vidutinis klasifikavimo tikslumas, naudojant visus metodus, viršijo 0,997. Šie duomenys mažai chaotiški, transformuotos laiko eilutės lengvai klasifikuojamos CNN.

4. „Earthquakes“ duomenų rinkinio klasifikavimas pasirodė prastas nepriklausomai nuo taikyto metodo. Vidutinis transformuotų laiko eilučių klasifikavimo tikslumas neviršijo 0,8 tiriant visus metodus. Transformuojančiais metodais gauti vaizdai nepasižymėjo informatyviais vizualiais bruožais. Šiame duomenų rinkinyje yra mažai periodiškumo ir dėsningumų, laiko eilutės labai chaotiškos.

Apibendrinant tai, kas buvo parašyta, tyrimas patvirtina, kad transformavimo metodas turi reikšmingą įtaką klasifikavimo rezultatams. Renkantis konkretų metodą praktikoje, verta atsižvelgti į konkretų duomenų rinkinį, tačiau verta pradėti nuo GADF, kurį taikius dažniau pasiekti aukštesni klasifikavimo rezultatai.

Literatūra

- [1] Jat, T., Patil, N., & Bhat, P. (2024). Detection of heart arrhythmia with electrocardiography. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 13(1), 57.
- [2] Medvedev, V., Budžys, A., & Kurasova, O. (2023). Enhancing keystroke biometric authentication using deep learning techniques. In *Proceedings of the 2023 18th Iberian Conference on Information Systems and Technologies*, 1–6. IEEE.
- [3] Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4), 917–963.
- [4] Mohammadi Foumani, N., Miller, L., Tan, C. W., Webb, G. I., Forestier, G., & Salehi, M. (2024). Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, 56(9), 1–45.
- [5] Wang, Z., & Oates, T. (2015). Imaging time-series to improve classification and imputation. In *Proceedings of the 2015 24th IJCAI International Joint Conference on Artificial Intelligence*, 3939–3945. AAAI.
- [6] Ratanamahatana, C. A., & Keogh, E. (2005). Three myths about dynamic time warping data mining. In *Proceedings of the 2005 5th SIAM International Conference on Data Mining*, 506–510. Society for Industrial and Applied Mathematics.
- [7] Olszewski, R. T. (2001). *Generalized feature extraction for structural pattern recognition in time-series data*. Carnegie Mellon University.
- [8] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, S. H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23), e215–e220.
- [9] Bagnall, A., Davis, L., Hills, J., & Lines, J. (2012). Transformation based ensembles for time series classification. In *Proceedings of the 2012 12th SIAM International Conference on Data Mining*, 307–318. Society for Industrial and Applied Mathematics.