Multidimenzionális idősorok előfeldolgozása és osztályozása deep learning alapú megközelítésekkel - Féléves beszámoló

Személyes adatok:

Név: Giricz Vince
Egyetem: Szegedi Tudományegyetem
Kar: Természettudományi és Informatikai Kar
Szak: Programtervező Informatikus
Neptun kód: FZGCTC

Feladat ismertetése:

A szakdolgozatom célja multidimenzionális idősor adatsorok előfeldolgozása és osztályozása deep learning alapú megközelítésekkel. A kutatás fókuszában olyan adatok állnak, amelyek jellemzői időben változnak, és gyakran több, párhuzamos forrásból származnak (pl. EKG, pulzus, pénzügyi adatok vagy IoT szenzorok adatai). A célom embedding alapú reprezentációk előállítása, amelyek lehetővé teszik a fontos jellemzők automatikus kinyerését. A munka során összehasonlítom a hagyományos gépi tanulási algoritmusok (pl. kNN, Random Forest) és a mélytanuló modellek (CNN, LSTM) hatékonyságát különböző teljesítménymutatók, például az F1-score és az accuracy alapján.

Elvégzett munka (Szeptember - Január):

A félév során az ütemtervnek megfelelően az alábbi feladatokat végeztem el:

2.1. Szakirodalmi kutatás és elméleti alapozás

- Áttekintettem a releváns publikációkat és tudományos cikkeket a multidimenzionális idősorok osztályozásáról.
- Külön figyelmet fordítottam az embedding technikákra, amelyek lehetővé teszik a komplex adatsorok alacsonyabb dimenziós, de információgazdag reprezentációját.

2.2. Adat-előfeldolgozás (Preprocessing)

- Az adatok előkészítése kritikus lépés, különösen az olyan komplex forrásoknál, mint az egészségügyi vagy IoT adatsorok.
- A fejlesztést Python nyelven végeztem.
- A pandas és NumPy könyvtárak segítségével megvalósítottam az adatok tisztítását és normalizálását.
- Elvégeztem az adatok konkatenálását, ami a multidimenzionális jelleg miatt volt szükséges a különböző szenzorforrások összehangolásához.

2.3. Embeddingek előállítása és alapmodell használata

- Az embeddingek generálásához a PyTorch keretrendszert és a MOMENT (egy modern, idősorokra specializált foundation model) könyvtárat használtam fel.
- A NumPy könyvtárat az embedding folyamat során keletkező nagyméretű adatvektorok hatékony kezelésére alkalmaztam.

2.4. Osztályozási kísérletek és klaszterezés

- A kísérletek első fázisában a kinyert reprezentációk minőségét vizsgáltam:
    - Az scikit-learn (sklearn) könyvtár segítségével K-Means klaszterezést hajtottam végre az embeddingeken.
- Ez a lépés lehetővé tette, hogy ellenőrizzem, az embeddingek mennyire jól különítik el az adatok belső szerkezetét, mielőtt rátérnék a komplexebb, felügyelt tanulási modellekre (pl. CNN, LSTM).

Aktuális tevékenység és technológiák:

A féléves beszámoló időpontjában a hangsúly a komplexebb mélytanuló architektúrák felé tolódott. Jelenleg az alábbi területeken végzek kutatást és implementációs kísérleteket:

3.1. Önfelügyelt tanulás (Self-Supervised Learning) és kontrasztív módszerek

- A célom olyan robusztus embeddingek kinyerése, amelyekhez nincs szükség előre felcímkézett adatra. Ezen belül az alábbi technológiákra fókuszálok:
    - Kontrasztív tanulás (Contrastive Learning): Vizsgálom a SimCLR adaptálhatóságát idősoros környezetre, valamint a kifejezetten idősorokhoz tervezett TS2Vec keretrendszert. Ezek segítségével a modell megtanulja megkülönböztetni az adatsorok különböző nézeteit (augmentációit), így stabilabb reprezentációkat hoz létre.
    - Idősor adat-augmentáció: Kutatom a specifikus augmentációs technikákat (pl. jittering, scaling, warping), amelyek elengedhetetlenek a kontrasztív tanulás hatékonyságához és a modellek általánosító képességének javításához.

3.2. Osztályozási stratégiák kevés adat esetén (Few-shot learning)

- Mivel az orvosi vagy IoT adatok címkézése gyakran költséges, vizsgálom a tanult embeddingek hatékonyságát korlátozott tanítóhalmaz mellett:
    - Linear Probing: A rögzített (frozen) embeddingekre épített egyszerű lineáris osztályozó teljesítményének mérése.
    - Fine-tuning: Az előtanított modellek (mint a korábban említett MOMENT vagy a TS2Vec) finomhangolása a specifikus osztályozási feladatra, összevetve a linear probing eredményeivel.

További ütemterv:

A második félév során a kutatási tervemet az alábbi ütemezés szerint kívánom véglegesíteni:

Februártól március közepéig - Modellfejlesztés és finomhangolás:

- A következő félév elején a K-Means eredményeire alapozva fogom finomhangolni a CNN és LSTM modelleket a tényleges osztályozási feladathoz.
- Lezárom a kontrasztív tanulással való kísérleteket, és véglegesítem az adat-augmentációs eljárásokat.
- Összehasonlítom a Linear Probing és Fine-tuning stratégiák hatékonyságát a kinyert embeddingek minőségének tükrében.

Március közepétől április végéig - Kiértékelés és dokumentáció kezdete, és a szakdolgozat lezárása:

- Az elvégzett kísérletek eredményeit részletes elemzésnek vetem alá olyan teljesítménymutatók alapján, mint az accuracy, precision, recall, és az F1-score.
- Megkezdem a szakdolgozat szöveges fejezeteinek a kidolgozását, különös tekintettel az elméleti háttér és a módszertan bemutatására.
- Véglegesítem az eredményeket szemléltető ábrákat és táblázatokat.
- Összegzem a következtetéseket a különféle mélytanulási és önfelügyelt megközelítések előnyeiről és hátrányairól.
- A dolgozat formai és tartalmi ellenőrzését követően benyújtom az egész szakdolgozatot.

Forráskód és szakirodalmi gyűjtemény:

A projekt transzparenciája és nyomonkövethetősége érdekében a fejlesztést szakaszosan feltöltve egy privát GitHub repozitóriumban végzem.

GitHub repozitórium

- GitHub elérhetősége: Gvince04/Szakdolgozat: Multidimenzionális idősorok előfeldolgozása és osztályozása deep learning alapú megközelítésekkel.
- A repozitórium tartalma:
    - Az adat előfeldolgozásához használt Python scriptek, amelyeket Jupyter Notebook formában vannak.

Feldolgozott szakirodalom (válogatás)

A kutatás fázis során az alábbi témaköröket és alapvető publikációkat dolgoztam fel:

- Idősor osztályozás alapjai:
- Foundation Models:
    - AutonLab/MOMENT-1-large · Hugging Face
    - momentfm · PyPI
    - moment-timeseries-foundation-model/moment: MOMENT: A Family of Open Time-series Foundation Models, ICML'24
- Kontrasztív tanulás:
- Mélytanuló architektúrák: