

Multidimenzionális idősorok előfeldolgozása és osztályozása deep learning alapú megközelítésekkel

Személyes adatok:

Név: Giricz Vince

Egyetem: Szegedi Tudományegyetem

Kar: Természettudományi és Informatikai Kar

Szak: Programtervező Informatikus

Neptun kód: FZGCTC

Motiváció

Az idősor adatok feldolgozása és elemzése egyre nagyobb jelentőséggel bír a modern informatika számos területén, mint például az egészségügyi adatok - például EKG, GSR, pulzus - pénzügyi idősorok, vagy IoT eszközök által rögzített adatok elemzésében. Az idősorok osztályozása különösen nagy kihívást jelent, mivel az adatok jellemzői időben változnak, gyakran párhuzamosan több forrásból származnak, multidimenzionális szerkezetűek, ezek miatt hagyományos módszerekkel gyakran nehezen elemezhetők.

A deep learning technológiák, különösen az embedding alapú reprezentációk lehetőséget adnak arra, hogy az idősorokból automatikusan kinyerhető jellemzők segítségével pontosabb osztályozás valósuljon meg. A szakdolgozatom célja, hogy ezeket a módszereket kipróbáljam, elemezzem, és összevegyem különféle gépi tanulási modellekkel.

Célok

- Idősorok előfeldolgozása és embedding alapú reprezentációk előállítás.
- Különböző osztályozók kipróbálása: egyszerűbb gépi tanulási algoritmusok (pl. kNN, SVM, Random Forest), és mélytanuló modellek (pl. CNN, LSTM).
- Az embedding alapú megközelítés hatásának vizsgálata az osztályozás teljesítményére.
- Az eredmények kiértékelése különféle mérőszámok (pl. accuracy, precision, recall, F1-score), alapján.
- Következtetések levonása a különféle megközelítések előnyeiről és hátrányairól.

Felhasznált technológiák, szoftverek

- *Python* - fő fejlesztési nyelv
- *Tensorflow / Keras, PyTorch* - neurális hálózatok alkalmazásához
- *scikit-learn* - klasszikus gépi tanulási algoritmusokhoz
- *NumPy, pandas* - adatok kezeléséhez, előfeldolgozáshoz
- *Matplotlib, seaborn* - vizualizációkhoz
- *Jupyter Notebook / Jupyter Lab / PyCharm* - a kísérletek dokumentálására

Ütemezés

Hónap	Feladat
Szeptember	Témakör pontosítása, és tématerv elkészítése.
Október	Releváns szakirodalom áttekintése, adatforrások felkutatása és kiválasztása.
November	Adatok előfeldolgozása: tisztítás, normalizálás, szegmentálás. Alap statisztikai elemzés és vizualizáció.
December	Embedding technikák megvalósítása és első kísérletek egyszerűbb, klasszikus ML osztályozókkal (kNN, SVM, Random Forest).
Január	Mélytanuló modellek (CNN, LSTM) implementálása és tesztelése az embeddingekkel.
Február	Kísérletek kiértékelése, teljesítménymutatók összehasonlítása, eredmények elemzése.
Március	A szakdolgozat szöveges részének megírása.
Április	Eredmények összefoglalása, ábrák és táblázatok véglegesítése, szakdolgozat lezárása.