Szökrön Dorottya - Önálló projekt I. (2021/22) - Projektbeszámoló

FEHÉRJE KLASSZIFIKÁCIÓ FUNKCIÓSOSZTÁLYOK ALAPJÁN

A tárgy keretein belül a szakdolgozatomban elkezdett munka folytatását, bővítését tűztük ki célul. Ennek témája a gépi tanulás alkalmazása volt molekuláris biológiai feladatokra, konkrétan fehérjék klasszifikálása az elsődleges szerkezetük alapján. A projektmunka során a fehérjék vizsgálatához a természetesnyelv-feldolgozásban (NLP) is használt módszerek közül alkalmaztam néhányat, és összehasonlítottam a különböző, szekvenciákat feldolgozó modelleket.

A feladatban a szakdolgozatomhoz képest új adathalmazt használtam. Ebben a korábbitól eltérő szempontból vannak osztályozva a fehérjék, amit azért tartok fontosnak kiemelni, mert a korábbi eredményeim alapján arra a megállapításra jutottam, hogy a korábbi modellek jó teljesítményéhez hozzájárult a korábbi adathalmaz osztályainak határozott elkülönülése egymástól. Az adathalmaz forrása egy Kaggle adathalmaz, amely a *Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB)* alapján lett létrehozva. A feladat megoldásához az adathalmazból két mezőt haszáltam fel. Az aminosavszekvenciát, ez a fehérjék elsődleges szerkezete. Valamint a funkciósosztályt, ami funkciós szerepük szerint kategorizálja a fehérjéket. Például a Hydrolase osztály tagjai olyan enzimek fehérjéi, amlyek a vegyületek hidrolízissel történő bontásában vesznek részt.

Az adatbáziban összesen 140911 rekord található, de szükséges volt adattisztítás végezni. Ki kellett szűrni a nem fehérje típusú makromolekulákat, valamint a különböző azonosítóval rendelkező, de azonos szekvenciával rendelkező sorokat.

Az adattisztítás után 87761 rekord maradt, amelyek összesen 3956 osztályhoz tartoztak. Az alapstatisztikák alapján szükséges volt az osztályok számának csökkentése, a kisebb osztályok között meghatározó számban fordultak elő olyanok, melyeknek mérete elhanyagolható volt a nagyobb osztályokéhoz képest. Az általam meghatározott érték alapján a legkisebb osztály mérete megközelítőleg 10%-a lett a legynagyobb osztály méretének, és így összesen 19 osztályt kaptam. A kísérletek egy része az így kapott adathalmazzal történt, de az oversampling-undersampling módszer segítségével megpróbáltam egy kiegyenlített adathalmazt létrehozni, és annak segítségével is végeztem további méréseket.

Kétirányú LSTM modell: Szakdolgozatomban GRU, LSTM és ezek kétirányú változatait használó modelleket hasonlítottam össze. Eredményeim alapján a kétirányú LSTM réteggel rendelkező modell teljesített a legjobban az ilyen típusú, azaz szekvenciális adatokon, ezért a projektfeladathoz kezdésként ezt a modellt használtam fel. Várakozásaimnál

rosszabb teljesítményt produkált, ezért módosítottam a felépítésén és hozzáadtam egy további kétirányú LSTM réteget.

Konvolúciós modell: Konvolúciós hálózatok hagyományosan képfeldogozási feladatokhoz alkalmasak. Működésük legfontosabb jellemzője, hogy a bemenetet részleteiben dolgozzák fel úgynevezett szűrők segítségével. A szűrők mérete és lépésköze meghatározandó paraméter, és ebből adódóan alkalmassá tehetőek szekvenciális adatok vizsgálatára is. Az ilyen hálózatok az 1D-s konvolúciós hálózatok. Kiinduló modellként hagyományos 1D-s konvolúciós rétegeket alkalmaztam, amely három különböző lépésközzel dolgozta fel a bemeneteket. Pontosság szempontjából nem sokkal, de jobb teljesítményt ért el az előző modellhez képest, viszont jelentősen gyorsabban teljesítette ezt az eredményt. Emiatt hatékonyabbnak mondható erre a feladatra.

Hibrid modell: Jelen esetben a hibrid kifejezés arra utal, hogy a modell a bemeneti adatokat két ágon, párhuzamosan dolgozza fel, és a különböző ágakon LSTM valamint 1D-s konvolúciós rétegek találhatóak, majd a kimeneti adatokat összesítve fog megtörténni a klasszifikáció. A nagyobb méret és komplexitás megalapozta a modell teljesítményének javulását a korábbiakhoz képest.

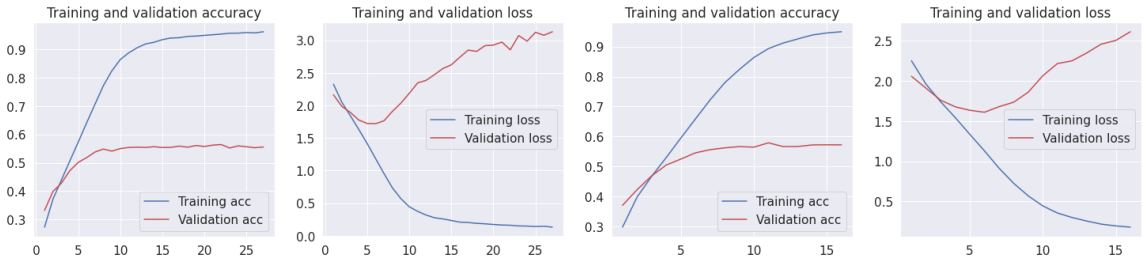
A különböző modellek kezdeti mérései után az adathalmaz kiegyenlítésén dolgoztam az oversampling-undersampling módszerét alkalmazva. Ez azt jelenti, hogy a tanítás során használt adathalmaz alacsony számosságú osztályainak tagjai véletlenszerűen megtöbbszöröződnek vagy a nagy számosságú osztályok tagjai közül véletlenszerűen elhagyásra kerül valamennyi, és így kiegyensúlyozott tanítóadathalmaz áll elő. Ekkor hangsúlyosabbá válhatnak a kisebb osztályok is, tehát elvárható, hogy a modell teljesítménye javuljon.

Összességében egyik modell sem teljesítette az elvárásokat a kiegyensúlyozott adathalmazon, bár egyenletesebb tanulási folyamat volt tapasztalható. Az oversampling elvégzése után a bonyolultabb modellek esetében annyira időigényessé vált a tanítás, hogy használhatatlanná váltak, valamint erős túltanulás volt jellemző mindegyik modell esetében. Az undersampling módszer kis mértékben javított a túltanulás problémáján, de számottevően romlott a modellek pontossága. Megfigyeléseim alapján a nem megfelelő eremédményekhez hozzájárult az adathalmaz minősége.

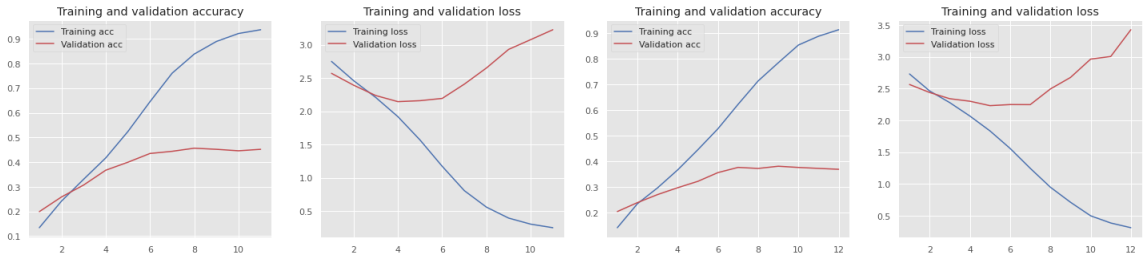
A következő félévekben szeretnék javítani a modellek teljesítményén új architekturális és tanítási módszerek bevezetésével, valamint új modellek, példásul transformer modellek használatával. Mindezek mellett pedig egy jobb minőségű adathalmaz kialakítását is szükségesnek gondolom.

Az alábbiakban szeretnék bemutatni néhány ábrát, amelyek a konvolúciós (CONV1D) és a hibrid (HYBRID) modellek teljesítményét reprezentálják.

Az *1. és 2. ábrán* láthatóak a modellek pontosságának és veszteségfüggvényének értékei az epochok számának előrehaladtával.



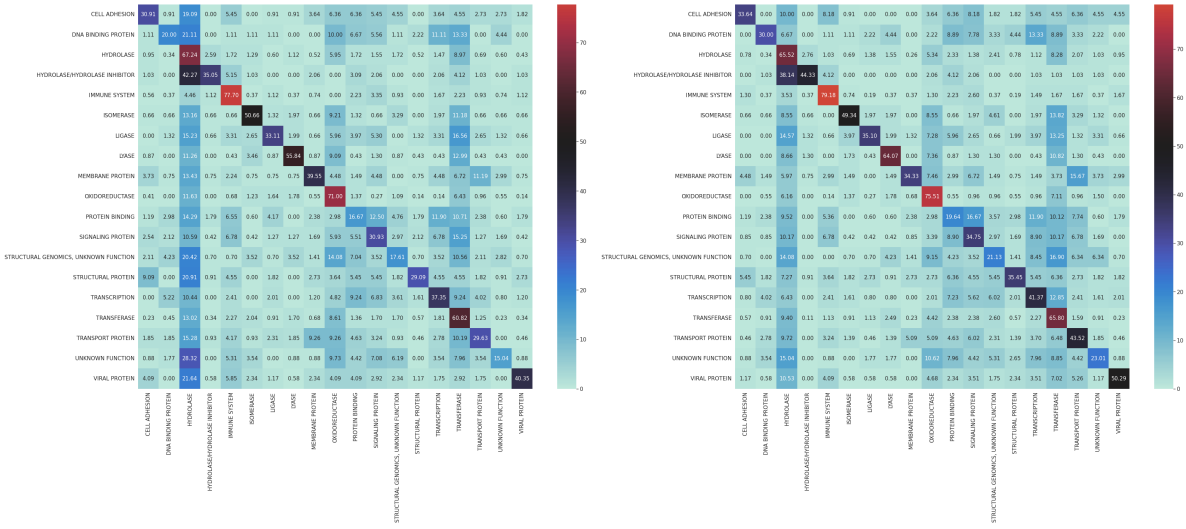
1. ábra. CONV1D (balra) és HYBRID (jobbra) kiegyensúlyozatlan adathalmaz esetében



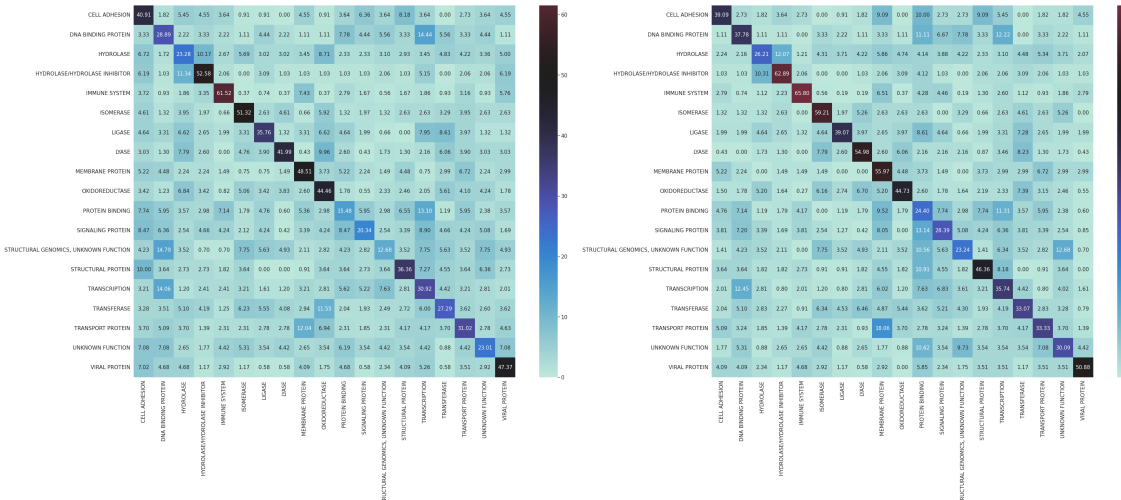
2. ábra. CONV1D (balra) és HYBRID (jobbra) kiegyensúlyozott adathalmaz esetében (undersampling)

	CONV1D	CONV1D	HYBRID	HYBRID
Database	normal	undersampled	normal	undersampled
Train acc (%)	98.23	96.71	97.7	96.12
Train loss	0.06	0.15	0.09	0.17
Val acc (%)	55.58	45.18	57.18	38.23
Val loss	3.13	3.22	2.61	3.41
Test acc (%)	56.88	45.98	57.84	39.42
Test loss	2.99	3.19	2.56	3.37

A *3. és 4. ábrán* láthatóak a modellek tévesztési mátrixai. Az eredmények százalékos arányban vannak megadva.



3. ábra. CONV1D (balra) és HYBRID (jobbra) kiegyensúlyozatlan adathalmaz esetében



4. ábra. CONV1D (balra) és HYBRID (jobbra) kiegyensúlyozott adathalmaz esetében (undersampling)

Források:

[1] https://www.kaggle.com/shahir/protein-data-set

[2] https://www.kaggle.com/srajpara/protein-classification-rnn-lstm-v-1

[3] Jason Brownlee "How to Develop LSTM Models for Time Series Forecasting" (2020) https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/

[4] Ronak Vijay "Protein Sequence Classification - A case study on Pfam dataset to classify protein families." (2019) https://towardsdatascience.com/protein-sequence-classification-99c80d0ad2df

[5 Jason Brownlee "How Do Convolutional Layers Work in Deep Learning Neural Networks?" (2020) https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/