Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
TMIT Tanszék

Bindics Boldizsár – Q12CTX

Frank Marcell – UMWAFS

Smuk András – D7S63U

BETEGSÉG-GÉN KÖLCSÖNHATÁS ELŐREJELZÉSE GRÁF NEURÁLIS HÁLÓZATOKKAL

Mélytanulás projekt dokumentáció

Tartalomjegyzék

1 Bevezetés 3

1.1 DisGeNET 3

1.2 Használt segédeszközök 4

2 Tanítás folyamata 4

2.1 Homogén és heterogén csúcshalmaz 5

2.2 Egyéb tervezési döntések 5

3 Modellünk kiértékelése 6

4 Összefoglalás 7

1 Bevezetés

Ez a projekt egy gráf neurális hálózatot (GNN) valósít meg, amely a gén-betegség asszociációk előrejelzésére szolgál. A modell célja, hogy valószínűségeket adjon meg arra vonatkozóan, hogy a jelenleg nem létező (nem behúzott) élek milyen valószínűséggel létezhetnének (behúzhatók lennének) a gráfban. Ez segíthet azonosítani, hogy érdemes-e egy adott gén-betegség kapcsolatot megvizsgálni és esetleg megerősíteni.

A betegségeket közül teljesen véletlenszerűen kiválasztottunk 23 darab „szimpatikusat” és ez alapján építettük fel a gráfunkat.

1.1 DisGeNET

Az adatforrásunk a DisGeNET platform volt, amely gének és genetikai variánsok közötti összefüggéseket tartalmazza. Ezeket az adatokat használják mások is a kutatásokban a betegség-gén összefüggések feltárására, potenciális biomarkereket azonosítanak és terápiás stratégiákat dolgoznak ki.

A letöltött adatbázisból az alábbi attribútumokat tartalmazó fájlt használtuk a modellünk tanítására:

- **geneNcbiID:** A gén azonosítója a National Center for Biotechnology Information adatbázisban.
- **geneDSI:** Ez a mutató (Disease Specificity Index) a gén és a betegség közötti kapcsolatot. Minél magasabb az érték a gén annál kevesebb betegséggel van kapcsolatban, tehát annál specifikusabb. (0,226 - 1)
- **geneDPI:** Ez a mutató a gén pleiotrópiáját (sokféle betegségben való részvételét) méri. Minél magasabb az érték, annál több betegséggel van a gén kapcsolatban. (0,043 - 1)
- **diseaseName:** A betegség neve.
- **score:** Az adott gén és betegség közötti kapcsolat erősségét vagy bizonyítékának megbízhatóságát mutatja. A pontszámot a DisGeNET algoritmusa számolja ki az elérhető adatbázisok, kutatások stb. alapján súlyozva. (0,4 - 0,9)
- **gene_index:** Az adott gén azonosítója a belső adatbázisban.
- **disease_index:** Az adott betegség azonosítója a belső adatbázisban.

1.2 Használt segédeszközök

A feladatot VS Code-ban és Google Collab-ban is implementáltuk. Az implementálás során a GitHub Copilot és a Gemini segítette. Ezeknek a használatát nagyon hasznosnak és hatékonynak éreztük. Előfordult, hogy a kódgenerálásnál hibázott az LLM, ezért használtunk futtatás közbeni visszaellenőrzéseket.

A feladat megoldásához a torchgeometrics, sklearn, seaborn és pandas, a vizualizációhoz matplotlib könyvtárakat használtuk. A feladat megvalósításához két cikket emelnénk ki, mivel ezeket a publikációkat használtuk a hálónk optimalizálásának segítésére:

- Variational Graph Auto-Encoders Thomas N. Kipf, Max Welling (https://arxiv.org/abs/1611.07308)
- node2vec: Scalable Feature Learning for Networks Aditya Grover, Jure Leskovec (https://arxiv.org/abs/1607.00653)

2 Tanítás folyamata

A tanítási folyamat több lépésből állt. Elsőként a gráf felépítése történt meg, ahol a csúcsok a betegségeket és a géneket, az élek pedig az asszociációkat reprezentálták. Az adathalmazt szétosztottuk tanító, validációs és teszt halmazokra, ahol az éleket pozitív (létező asszociációk) vagy negatív (nem létező asszociációk) címkékkel láttuk el. A csúcsokhoz tartozó jellemzők (feature-ök) kinyerése után ezek, valamint az élek címkéi szolgáltak bemenetként a GNN számára.

A tanítás során gráfkonvolúciós rétegeket használtunk, amelyek információt propagáltak és aggregáltak a gráf struktúrájában. A modellt egy olyan veszteségfüggvény minimalizálásával optimalizáltuk, amely a predikált asszociációs pontszámokat hasonlította össze a valós címkékkel. Az optimalizálás során a súlyok visszaterjesztéses algoritmussal (backpropagation) kerültek frissítésre.

A folyamat során több kihívással is szembesültünk, például a gráf konstrukciójánál nem volt egyértelmű, hogy milyen adatstruktúrát lenne érdemes használni, illetve nehézséget okozott a megfelelő gráfkonvolúciós megoldás kiválasztása is.

2.1 Homogén és heterogén csúcshalmaz

A projekt során foglalkoztunk mind a homogén, mind a heterogén gráf megközelítéssel. A homogén gráf esetében a betegségek és gének ugyanabba a csúcshalmazba tartoztak, ezért itt bizonyos node feature-ök esetén dummy értékekkel kellett dolgoznunk, valamint a csúcsok indexelése, valamint a negatív élek mintavételezése is nehézséget okozott. Ezzel szemben a heterogén gráf külön csúcstípusokat használt a betegségek és gének számára, megoldva ezzel a különböző node feature-ök problémáját.

Végül a heterogén megközelítés mellett döntöttünk, mivel ez jobban tükrözi a valódi biológiai hálózatok komplexitását, és lehetőséget adott a különböző csúcstípusok explicit modellezésére. Ezáltal a hálózat tanítása során pontosabb predikciókat érhettünk el, és a modell további bővítése akár új tulajdonságokkal, vagy adatpontokkal is egyszerűbb.

2.2 Egyéb tervezési döntések

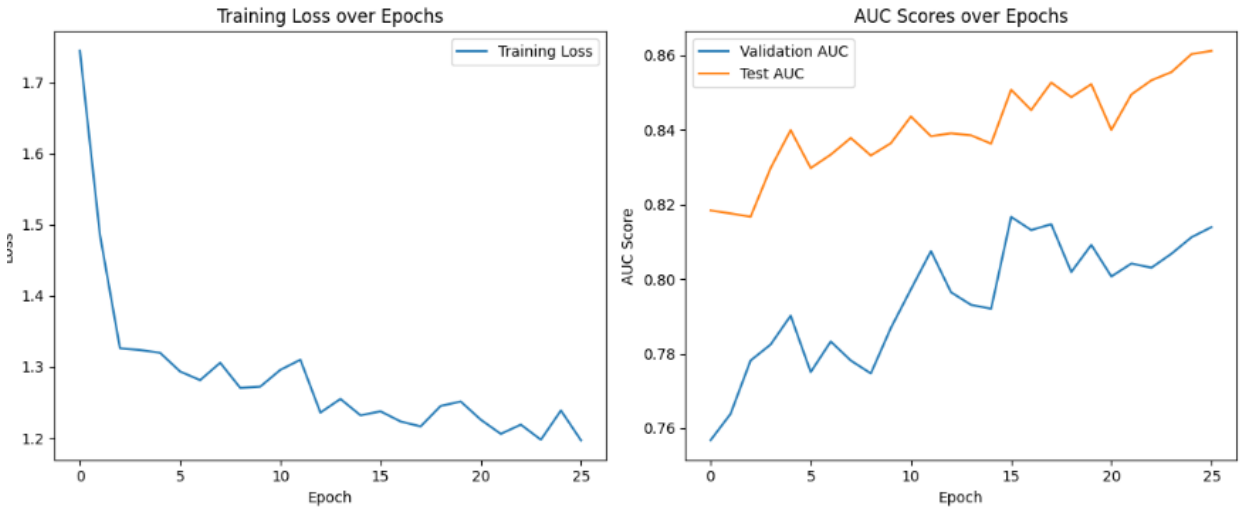
Ehhez a heterogén GNN feladathoz az alábbi eszközöket választottuk, mivel mindegyik hatékonyan támogatja a linkpredikció és a heterogén gráfok kezelésének speciális igényeit:

- **SAGEConv:** A Pytorch Geometrics dokumentációjában található GNN Cheatsheet alapján választott SAGEConv homogén gráfokra tervezett konvolúciós háló (GCN), azonban a torch_geometric.nn.to_hetero() hívás segítégével alkalmas heterogén gráfok egyszerű és gyors kezelésére. Hatékonyan modellezi a gráf lokális struktúráját, amely kulcsfontosságú a linkpredikciós feladatban.
- **torch.optim.Adam:** Az Adam optimalizálót választottuk, mert gyorsan konvergál és hatékonyan kezeli a gradiens alapú optimalizálás során fellépő instabilitásokat. Alacsony konfigurációs igénye miatt jól alkalmazható a GNN-ek komplex modelljeinek tanításához.
- **BCEWithLogitsLoss:** A bináris keresztentrópia logit-alapú változata természetesen illeszkedik a linkpredikciós feladat bináris osztályozási jellegéhez. A logit-alapú megközelítés stabilabb numerikus viselkedést biztosít, különösen sigmoid aktivációval kombinálva.

Ezek az eszközök kombinációban biztosítják, hogy a modell hatékonyan és pontosan tanulja meg a heterogén gráfok bonyolult szerkezetéből származó mintázatokat.

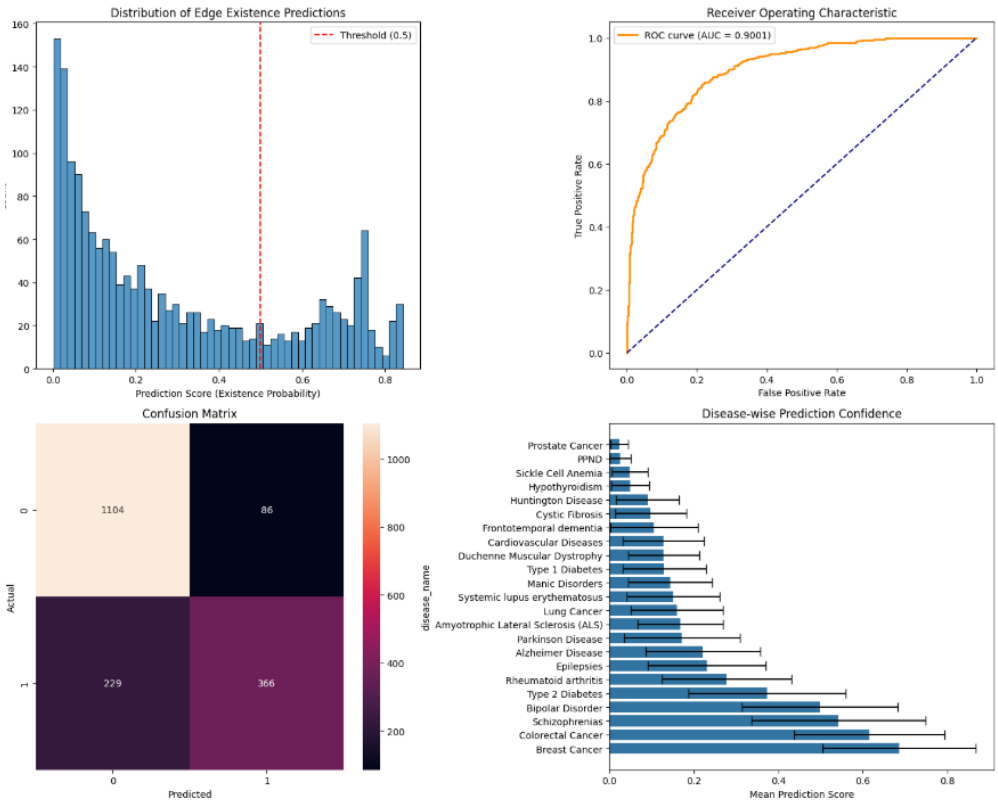
3 Modellünk kiértékelése

Ahogyan az alábbi ábra is mutatja, a tanítási ciklusok során a számított veszteség egyre csökkent, és az ROC görbe alatti területe pedig növekedett, így a grafikonokon látható értékek alapján a modellünk tanítása sikeresnek mondható.



ábra 1 Veszteség és AUC változása a tanítási ciklusokban

Az ábrán a modell teljesítményének különböző aspektusait láthatjuk. Az első grafikon, a "Distribution of Edge Existence Predictions", az él-kapcsolat valószínűségeit mutatja, kiemelve a választott 0.5-ös küszöböt, amely felett a kapcsolatok pozitívnak minősülnek. A második grafikon, a "Receiver Operating Characteristic (ROC)", az osztályozás hatékonyságát illusztrálja, amelynek AUC értéke 0.9001, jelezve, hogy a modell jól teljesít a hamis pozitív és valódi pozitív arányának mérséklésében. A harmadik ábra, a "Confusion Matrix", a modell predikciós teljesítményét ábrázolja a valódi és prediktált címkék összevetésével, ahol jól látható a modell sikeres pozitív és negatív osztályozása. Végül, a "Disease-wise Prediction Confidence" diagram a betegség-specifikus predikciós eredményeket mutatja, rangsorolva a legmagasabb és legalacsonyabb előrejelzési pontszámokkal rendelkező betegségeket, ahol a Prostate Cancer és a PPND a legbiztosabb jóslatok között szerepelnek.



ábra 2 Modellünket jellemző vizualizációk

4 Összefoglalás

A munkánk lényege, hogy egy graph neural network (GNN) modellt fejlesztettünk ki a betegségek és gének közötti kapcsolat előrejelzésére, amely lehetővé teszi számunkra, hogy megértsük, mely gének és betegségek közötti kapcsolatok érdemesek a további vizsgálatokra. Mivel minden kutatás egy alapvető kérdéssel kezdődik – „milyen géneket érdemes vizsgálni egy adott betegség esetében?” – a mi modellünk segít eligazodni a bonyolult biológiai hálózatokban, és értékes információval szolgálhat a kutatók számára a releváns gének kiválasztásában.

Továbbfejlesztésként a jövőben több gént és betegséget is bevonhatunk a modellbe, hogy szélesebb körű predikciókat végezhessünk. Emellett érdemes lenne különböző megközelítéseket összevetni, például más típusú GNN modelleket is alkalmazni, és versenyeztetni őket a legjobb teljesítmény elérése érdekében. A hiperparaméterek további finomhangolása manuálisan vagy automatikusan optimalizált kereséssel tovább növelhetné a modell hatékonyságát, így még pontosabb előrejelzéseket adhatnánk a betegség-gén kapcsolatok terén.