Az AIDS előrehaladásának felismerése gépi tanulás eszközeivel

Buza Krisztian[1]
[1]*tudományos munkatárs*
[1]Hildesheimi Egyetem
E-mail: buza@ismll.de

Összefoglalás: Az utóbbi három évtizedben a HIV vírus világszerte több mint 25 millió ember halálát okozta. A közelmúltban a kaggle.com internetes portál egy versenyt rendezett, melyen a résztvevőknek olyan (statisztikai adatelemzésen alapuló) modellt kellett készíteniük, amely képes előre jelezni, hogy rövidtávon javulni fog-e egy páciens állapota a kezelés hatására. Ebben a cikkben az általunk készített modellt mutatjuk be.

Kulcsszavak: HIV, AIDS, gépi tanulás, adatbányászat, szekvenciák osztályozása

Abstract: In the last three decades, HIV caused the dead of more than 25 million persons worldwide. Addressing more than 100 researchers and practitioners of machine learning and biomedical technologies, kaggle.com has recently hosted a challenge aiming at developing statistical forecast models for predicting patients' short term progression, i.e. if the patient's status improves or not. In this paper, we describe our model we developed for the challenge.

Keywords: HIV, AIDS, machine learning, data mining, sequence classification

1. Bevezetés

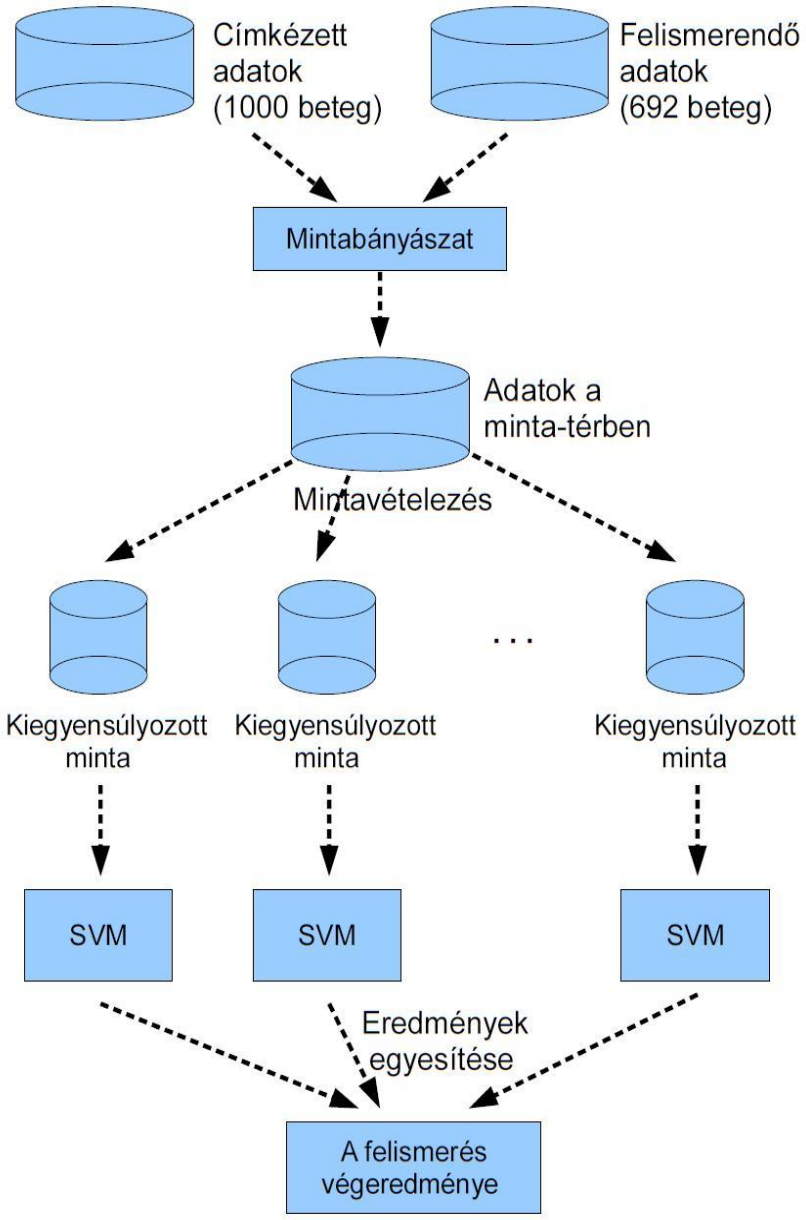
A HIV vírus 1981-es felfedezése óta – az egészségügyi világszervezet statisztikái szerint – több mint 25 millió ember halálát okozta világszerte. [1] (Ez a szám más forrás szerint körülbelül 40 millió [10].) Bár a közelmúltban jelentős előrelépések történtek a betegség megértésével kapcsolatban, és egy védőoltás kifejlesztése is biztató közelségbe került [2], jelenleg az AIDS a legveszedelmesebb kórok egyike. Különösen fontos kérdés a már fertőzött betegek kezelése; ennek oka kettős: egyrészt egy jövőbeli esetleges védőoltás hasznát ők már nem élvezhetik, másrészt megfelelő terápia mellett évekig, vagy akár évtizedekig is értékes életet folytathatnak. A probléma komolyságát mutatja az is, hogy 2008-ban világszerte több, mint 30 millió HIV-fertőzött ember élt [10].
A kaggle.com internetes portál a közelmúltban egy versenyt rendezett [1], melyen a résztvevők feladata az volt, hogy statisztikai (adatbányászati) elemzésen alapuló modellt készítsenek, mely képes előre jelezni, hogy mely páciensek állapota fog rövidtávon javulni. Az előrejelző modellek fejlesztéséhez 1000 páciens adatait lehetett használni. Ezen adatok genetikai szekvenciákat, valamit viral load számot és a CD4+ számot (fehérvérsejtek száma) tartalmaztak. Adott továbbá, hogy az 1000 páciens közül melyek állapota javult, és melyeké nem. Az említett 1000 fő mellett egy további, 692 főből álló csoport adatait (nukleotid szekvenciák, viral load és CD4+ szám) is megismerték a verseny résztvevői. Ezen 692 fős

csoport esetében (ellentétben az előbbi 1000 emberrel) a verseny résztvevői *nem tudták*, hogy a páciens állapota javult-e vagy sem. A feladat annak előrejelzése volt, hogy közülük melyik paciens állapota fog javulni, és melyiké nem. A résztvevők az előrejelzéseiket interneten keresztül küldték be, a verseny szervezői pedig kiértékeltek ezeket az előrejelzéseket. (A verseny szervezői tudták, hogy a 692 beteg állapota hogyan változott, ez alapján ki tudtak számolni, hogy a beküldött eredmény az esetek hány százalékában volt helyes.)
A megmérettetésen az adatbányászat, gépi tanulás és biomedikus technológiák több mint 100 kutatója és gyakorlati alkalmazója vett részt, s összesen több mint 800 modell és modell-változat előrejelzései kerültek kiértékelésre. Ebben a cikkben leírjuk az általunk készített modellt.

2. Az általunk fejlesztett modell

Említettük, 1000 betegről tudjuk, hogy rövidtávon javult-e az állapotuk. Ezen páciensek adatait (amely betegenként két-két genetikai szekvenciát, viral load számot és CD4+ számot tartalmaz) a továbbiakban címkézett adatoknak nevezzük. További 692 beteg adatai (genetikai szekvenciák, viral load és CD4+ szám) is adott volt, esetükben azonban a verseny résztvevői nem tudták, hogy melyik beteg melyik csoportba (azok közé, akiknek javul az állapotuk, vagy azok közé, akiknek nem) tartozik. Mivel a feladat annak felismerése volt, hogy ezen 692 beteg közül melyek állapota fog rövidtávon javulni és melyeké nem, így ezen 692 beteg adatait a továbbiakban felismerendő adatoknak nevezzük.
Az általunk fejlesztett modell gyakori mintabányászati technikákon és szupport vektor gépekkel (support vector machines, SVM) végzett felismerésen alapul, az eljárás vázlata az 1. ábrán látható.
Az eljárás első lépésében („Mintabányászat“) gyakori mintázatokat (gyakori részszekvenciákat) keresünk a genetikai adatokban. Ennek során külön-külön tekintjük a kétféle genetikai információt, a Reverse Transcriptase (RT) és a Protease (PR) nukleotid szekvenciákat. Egy-egy RT-szekvencia kb. 300 bázisból (szimbólumból) áll, míg a PR-szekvenciák hossza 1000 körüli. Mind a két szekvencia alapvetően a négy genetikai bázist jelölő karakterekből (C,T,G,A) áll.
A „Mintabányászat“ lépésben az RT-szekvenciák esetében olyan gyakori mintázatokat keresünk, amelyek i) legalább 50-szer fordulnak elő a címkézett adatok között, és ii) a címkézett és felismerendő adatokban összesen legalább 200-szor fordulnak elő. A PR-szekvenciák esetében pedig olyan gyakori mintázatokat keresünk, amelyek i) legalább 50-szer fordulnak elő a címkézett adatok között, és ii) a címkézett és felismerendő adatokban összesen legalább 100-szor fordulnak elő. Például a „CCTGCC“ gyakori RT-mintázat abban az esetben, ha a címkézett adatok közül legalább 50 RT-szekvencia tartalmazza a „CCTGCC“ mintát, és a teljes adatbázisban (címkézett és felismerendő adatok között összesen) legalább 200 szekvencia tartalmazza ezt a részszekvenciát.
Figyelembe véve, hogy a genetikai adatok tartalmazhatnak valamekkora zajt, azaz előfordulhat, hogy míg az egyik szekvenciában „CCTGCCA“-t látunk, addig a másikban „CCTGACA“-t, ezért a mintázatokban megengedünk legfeljebb egy darab „joker“ karaktert, mely az adatbázisbeli szekvenciák egy tetszőleges karakterére illeszkedik. Például a „CCTG*CA“ mintázat (a „joker“ karaktert *-gal jelöltük) esetében úgy tekintjük, hogy ez a mintázat az olyan adatbázisbeli szekvenciában fordul elő, ahol a „CCTG“ részszekvenciát egy tetszőleges karakter követi, ezt a tetszőleges karaktert pedig végül „CA“ követi.
Az ilyen gyakori minták kinyeréséhez az ECLAT-algoritmus [3] taxonomikus szekvenciák esetére adaptált változatát használtuk [4]. (Esetünkben a taxonómia rendkívül egyszerű: a

taxonómia gyökere a „joker“ karakter, minden egyéb szimbólum pedig a gyökér közvetlen leszármazottja).



1. ábra
Az általunk fejlesztett modell vázlata.

A gyakori mintázatok közül kiválasztjuk azokat, amelyek leginkább jellemzőek egy-egy csoportra. Ehhez KHI-négyzet próbát használunk. A „jellegzetes“ gyakori mintázatok kiválasztása után az adatainkat egy sokdimenziós vektortérbe transzformáljuk a mintázatok segítségével. Minden egyes beteg adatainak pontosan egy vektort feleltetünk meg. A vektor egy-egy komponense egy-egy mintázatnak felel meg: egy-egy vektorkomponens számértéke azt adja meg, ahányszor az adott mintázat előfordul a beteg genetikai szekvenciái között. A beteg adatait leíró vektor a mintázatoknak megfelelő komponenseken túl további két komponenst tartalmaz: a CD4+ számot és a viral load számot. A vektortérbe (mintatérbe) történő leképezés azért előnyös, mert ez után már bármilyen klasszikus osztályozó algoritmust (döntési fák, legközelebbi szomszéd modellek, neurális hálók, stb.) használhatunk. Mivel az utóbbi években a szupport vektor gépek (support vector machines, SVM) rengeteg feladatban bizonyultak sikeresnek, választásunk erre a modellosztályra esett. Konkrétan a WEKA szoftvercsomagbeli [5] SMOreg [6,7] implementációt használtuk.
A szupport vektor gépek működése (más osztályozó modellekhez hasonlóan) két fázisból áll: először a címkézett adatokat elemezve mintegy „megtanulják“ azokat a szabályszerűségeket, hogy mikor tartozik egy beteg az egyik vagy a másik csoportba, majd ez után használhatók arra, hogy a felismerendő adatokat besorolják valamelyik csoportba. Amennyiben a tanítás során használt adatokban a csoportok eloszlása kiegyensúlyozatlan (imbalanced), azaz valamelyik csoportbeli adatok „túl gyakran“ fordulnak elő, míg más csoportbeliek „túl ritkán“, akkor előfordulhat, hogy a felismerési fázisban a szupport vektor gépek hajlamosak lesznek „túl sokszor“ a többségi csoportba sorolni a felismerendő adatokat. Esetünkben a címkézett adatokat képező 1000 beteg közül mindössze 206 állapota javult, míg a felismerendő adatok között a betegek felének javult az állapota. Ezért tehát a kiegyensúlyozatlansági probléma esetünkben fennáll. A kiegyensúlyozatlanság negatív hatását mintavételezéssel igyekszünk csökkenteni: a mintavételezett adathalmazba 0,8 valószínűséggel kerül be egy-egy olyan beteg, akinek az állapota javult, míg csupán 0,2 valószínűséggel kerül be egy-egy olyan beteg, akinek az állapota romlott. Az így mintavételezett adathalmazban a csoportok eloszlása (várható értékben) kiegyensúlyozott. Ezt a mintavételezést összesen 100-szor végezzük el, melynek eredményeként 100 darab kiegyensúlyozott adathalmazunk lesz. (Ezen adatbázisok természetesen nem függetlenek egymástól: ugyanazon beteg több adathalmazban is előfordulhat.)
Minden egyes kiegyensúlyozott adathalmazon egy-egy szupport vektor gépet „tanítunk“, így összesen 100 szupport vektor gépünk lesz, amely képes a felismerendő adatokat valamely csoportba besorolni. A felismerendő adatok esetében minden egyes beteget mind a 100 szupport vektor géppel külön-külön besorolunk valamely csoportba. Az egyes csoportokat 0-val és 1-gyel kódoljuk, végső eredményként a 100 szupport vektor gép kimenetének átlagát kerekítjük 0-ra illetve 1-re.
Az itt leírt modell a felismerendő adatokat 62 %-os pontossággal ismerte fel, ezzel 44-dik helyezést értük el a versenyben, melyen a világ különböző tájairól 109 résztvevő összesen több mint 800 modellel, illetve modell-változattal vett részt. (A győztes 77 %-os felismerési pontosságot ért el.)

3. Kitekintés

Bízunk abban, hogy a bemutatott és a hozzájuk hasonló modellek a jövőben alkalmazásokra találnak majd a betegek gyógyításában: ha az AIDS (vagy más betegség) kezelésére különböző terápiák léteznek, és sikerül olyan modelleket készíteni, melyek képesek előre jelezni egy-egy terápia sikerét (javulni fog-e a beteg állapota az adott terápia hatására), és ez

az előrejelzés megfelelő pontosságú, akkor érdemes lehet azon terápiák közül valamelyiket próbálni először, amelyik a statisztikai előrejelző modell szerint valószínűleg sikeres lesz.
Ahhoz, hogy ilyen távlati alkalmazások valóra válhassanak, egyrészt a modellek pontosságának növelése szükséges, másrészt pedig annak megértésére, hogy mikor (mitől, miért) működik jól vagy rosszul egy modell. Egy lépésként ebbe az irányba eszmecserét szeretnénk motiválni azáltal, hogy néhány további ötletet is közre adunk (köztük olyanokat is, melyek a mi kísérleteinkben sikertelennek bizonyultak: érdemes lenne tudni, hogy ezek vajon a feladat jellege miatt nem vezettek sikerre, vagy egyszerűen nem megfelelően „építettük be“ ezeket az ötleteket a kipróbált modelljeinkbe).
Hasonló felismerési feladatoknál, idősorok esetében (itt a feladat annak eldöntése, hogy melyik csoportba tartozik egy idősor), igen sikeresek a legközelebbi szomszéd alapú, DTW távolságfüggvényt alkalmazó eljárások [8]. A genetikai szekvenciák, amelyekkel ebben a cikkben dolgoztunk, az idősorokkal ellentétben nem valós számok sorozatából, hanem diszkrét jelek sorozatából állnak. Az idősorokkal kapcsolatos eredmények alapján azt várnánk, hogy a DTW-hez hasonló, Lehevenstein-távolságfüggvényt alkalmazó legközelebbi szomszéd alapú modellek jól fognak működni genetikai szekvenciák felismerésekor. Sajnos kísérleteinkben ennek ellenkezőjét tapasztaltuk.
Érdekesnek tűnik az angol szakirodalomban semi-supervised néven hivatkozott felismerési protokoll is, melynek lényege: miután a felismerő modellt (pl. szupport vektor gépet) a címkézett adatokon „tanítottuk“, a felismerő modellel először azokat az adatokat ismerjük fel, melyeket a modell legnagyobb biztonsággal képes felismerni. Ezután ezeket az adatokat (melyeknek a címkéjét a modell immár felismerte) hozzávesszük a tanító adatokhoz, és a modellt újra tanítjuk. Az eljárást addig ismételjük, amíg minden adatot fel nem címkéztünk. Viszonylag kisméretű adathalmaz esetében, különösen, amikor a címkézett adatok és a felismerendő adatok karakterisztikája különbözik, előnyösnek tűnik ez az eljárás. Sajnos kísérleteink során ez az ötlet sem bizonyult sikeresnek.
Sikeresnek bizonyult azonban a címkézett adatok „kiegyensúlyozottabbá“ tétele, olyannyira, hogy blogbejegyzésének [9] tanúsága szerint ezt az ötletet a verseny győztese is alkalmazta. Szintén hasznosnak bizonyult a győztes azon ötlete, hogy a genetikai szekvenciákban olyan pozíciókat keresett, amelyek korrelálnak azzal, hogy melyik csoportba tartozott a beteg.

4. Konklúzió

Ebben a cikkben bemutattunk egy általunk fejlesztett, gyakori mintabányászati eljárásokon és szupport vektor gépeken alapuló modellt, melynek célja annak felismerése, hogy mely AIDS-betegek állapota fog rövidtávon javulni. Ezzel a modellel egy közelmúltbeli versenyen (melyen 109 kutató és fejlesztő, összesen több mint 800 modellel és modellváltozattal vett reszt) a 44-dik helyezést értük el. Bízunk benne, hogy a modell részleteinek publikálásával és nyilvános eszmecsere motiválásával sikerült hozzájárulnunk a betegség elleni küzdelemhez.

Irodalomjegyzék

[1] A verseny weblapja: http://www.kaggle.com

[2] http://index.hu/tudomany/2010/07/09/megelozheto_az_aids/

[3] Han, J.; Pei, J.; Yin Y.; Mao R.: *Mining frequent patterns without candidate generation.* Data Mining and Knowledge Discovery, 8:53 – 87, 2004;

[4] Blohm, S.; Buza, K.; Cimiano, P.; Schmidt-Thieme, L.: *Relation Extraction for the Semantic Web with Taxonomic Sequential Patterns*, in „Applied Semantic Web Technologies“. Boca Raton, FL: Taylor and Francis Group, ISBN: 978-14398-01567, to appear;

[5] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten I. H.: The *WEKA Data Mining Software: An Update*. In: SIGKDD Explorations, Volume 11, Issue 1., 2009;

[6] Smola, A. J.; Scholkopf, B.: *A Tutorial on Support Vector Regression*. NeuroCOLT2 Technical Report Series, NC2-TR-1998-030, 1998;

[7] Shevade, S. K.; Keerthi, S. S.; Bhattacharyya, C.; Murthy, K. R. K.: *Improvements to SMO Algorithm for SVM Regression*. Technical Report CD-99-16, Control Division Dept. of Mechanical and Product Engineering, National University of Singapore.

[8] Xi, X.; Keogh, E.; Shelton, C.; Wei, L.; Ratanamahatana, C. A.: *Fast Time Series Classification Using Numerosity Reduction*. In: Proceedings of the 23th International Conference on Machine Learning, 2006;

[9] http://kaggle.com/blog/2010/08/09/how-i-won-the-hiv-progression-prediction-data-mining-competition/

[10] http://data.unaids.org/pub/Report/2009/JC1700_Epi_Update_2009_en.pdf