Gépi tanulási módszerek a fotometrikus vöröseltolódás-becslésben

Fizika BSc szakdolgozat

Horváth Bendegúz

az ELTE TTK Fizika BSc hallgatója

Témavezető: Dr. Csabai István egyetemi tanár, Komplex Rendszerek Fizikája Tanszék

Budapest, 2018 május

Kivonat

Ide kell megírnom a kivonatot.

Tartalomjegyzék

1. fejezet

Bevezetés

Az Univerum nagy skálás szerkezetének megértéséhez szükséges, hogy térképet tudjunk készíteni a galaxisok elhelyezkedéséről. Két koordinátát, a galaktikusszélességet és galaktikushosszúságot könnyen megkaphatjuk, viszont a távolság meghatározása már nehezebb feladat. A trigonometrikus parallaxis módszer a legjobb technikákkal is csak galaxison belül működik, a jó *seeing* érdekében pedig űrtávcső kell. A standard gyertya módszerek pontos távolságértéket adnak, de csak néhány százmillió fényév távolságon belül alkalmazhatóak, ezért kell egy olyan módszer, amivel távolabb is mérhetünk, pontosan.

A XX. század elején Edwin Hubble és Vesto Slipher a galaxisok színképének tanulmányozása során észrevették, hogy a színképek eltolódnak a nagyobb hullámhosszak, a vörös színtartomány felé a laboratóriumban mért vonalszerkezethez képest, a galaxis vöröseltolódást szenved. Hubble méréseket készített a galaxisok távolságáról és vöröseltolódásáról, és lineáris összefüggést tapasztalt, amit a később róla elnevezett törvény ír le:

$$v = H \cdot d, \tag{1.1}$$

az arányossági tényező a Hubble-állandó, értéke $H = 73.45 \pm 1.66$ *km/Mpc*. Ez az összefüggés lehetőséget ad a pontos távolságmérésre, ha a galaxisok vöröseltolódását meg tudjuk mérni.

Vöröseltolódás mérésre a legpontosabb módszer felvenni az objektum spektroszkópiai képét, és megnézni, hogy a vonalak mennyire csúsztak el. A megfelelő minőségű spektrum

felvételéhez akár egy órányi távcsőidő szükséges, ezért sokáig nem is készült égbolttérkép. Az 1986-ban Margaret Geller és munkatársai által készítet égtérkép csupán egy vékony szeletet fedett le az égből, de már azon is kirajzolódott, hogy a galaxiseloszlás nem egyenletes azon a skálán. Gellerék térképén a legtávolabbi galaxisok körülbelül 200 *Mpc* távolságnyira voltak tőlünk, ezért volt motiváció elkészíteni egy még távolabb látó térképet. A *BEKS* égfelmérés szűkebb, 1×1 négyzetfokos tartományban, ceruzaszerűen[1] nézett 1000-1000 *Mpc* távolságba mindkét irányba, és ezen a skálán is kirajzolódótt struktúra az anyageloszlásban, és a sűrűség ingadozás periodikusnak mutatkozott. Nagy távolságokra csökkent az észlelt galaxisszám, ez a halványabb galaxisok nehezebb észlelhetősége miatt volt. A technikai fejlődésnek köszönhetően elkészülhettek olyan kisérleti berendezések, amelyek lehetővé tették valódi háromdimenzióban a galaxisok pontos helyzetének felmérését. A Sloan Digital Sky Survey első fázisában 1 millió objektum spektrumát vette fel négy év alatt[1]. Ez a sebeség nem kielégítő, ezért felmerült az igény, hogy a vöröseltolódásokat fotometriai úton mérjék. A fotometriával mért vöröseltolódások kicsit pontatlanabbak, de mrérésük gyorsabb mint spekroszkópiával, ezenkívűl halványabb objektumokat is lehet vele mérni, magasabban van a magnitúdó korlát. Ezen tulajdonságok vonzóvá teszi a vöröseltolódás becslését fotometriai módszerekkel.

Dolgozatom célja az általam készített, gépi tanuláson alapuló fotometrikus vöröseltolódás-becslő módszerek bemutatása, amelyek a galaxisok képeit használjaák fel fotometriaia adatként.

[1] pencil-beam survey

2. fejezet

Vöröseltolódás-becslés és gépi tanulás

2.1. Fotometrikus vöröseltolódás-becslés

A vöröseltolódás fotometriával történő becslésének kétféle megközelítése van, empirikus és spektrumokon alapuló. Egy spektrumon alapuló módszert először 1962-ban írt le Baum[2], fotoelektromos fotométert használt kilenc sáváteresztő filterrel, amik 3730 Å és 9875 Å közötti hullámhosszú fényt engedtek át. Ezzel a rendszerrel 6 fényes elliptikus galaxis spektrális energia-eloszlását (spectral energy distribution, SED) mérte meg a Virgo halmazból, majd még háromnak egy másik halmazból. A SED-ek átlagát ábrázolta a hullámhossz logaritmusának függvényében, és képes volt észrevenni az eltolódást a két energia sűrűségeloszlás között, így a második klaszternek a vöröseltolódását is megkapta. Mérése pontos volt, de a módszer arra támaszkodott, hogy a spektrumoknak 4000 Å-nél levágása van, melyet a csillag-atmoszférák fémtartalma okoz, ez az elliptikus galaxisoknál jól látható, de például az aktív csillagkeletkezést mutató irreguláris galaxisokban ez a levágás nem figyelhető meg, ezért ez a módszer csak elliptikus galaxisoknál volt használható[3].

David C. Koo egy másik módszert, a szín-szín diagrammok módszerét vezette be 1985-ös cikkében[4]. Négy sáváteresztő szűrőt használt, melyeken mért magnitúdók különbségével létrehozott színtereken ábrázolva az azonos típusú, különböző vöröseltolódású galaxisokat jól definiált görbét kapott. Szín-szín diagrammokat bármilyen kombinációjából lehet készíte-

ni három vagy több színnek, az optimális választás a várt vöröseltolódás-eloszlástól függhet, a gyakorlati választás a rendelkezésünkre álló szűröktől[4]. Ez a megközelítés a fotometriai vöröseltolódás-meghatározás fontos eleme lett, ugyanis az egyes galaxisok típusának és színszűrőkön mért fényességének ismeretében meghatározható, hogy milyen vöröseltolódást szenved a galaxis.

Egy harmadik, gyakran használt módszer a *template fitting* (sablon illesztés), ez a módszer is a spektrális energiasűrűség-eloszlásra támaszkodik, az ismert vöröseltolódású és SED-ű galaxisok SED-jéből felépül egy könyvtár, és az ismeretlen vöröseltolódású galaxisok SED-jét hozzá lehet párosítani hasonlóság alapján a könyvtárban lévőkhöz. A *template* módszerek nagyon hasznosak lehetnek új égboltfelmérésnél, ha nem áll rendelkezésre megfelelő mennyiségű spektroszkópiai adat. A módszer használatával, különösen ha elméleti sablonokat használnak[5], a vöröseltolódáson kívűl egyéb fizikai tulajdonságát is is ki lehet nyerni a galaxisnak. Megfelelő használatához a sablonkönyvtárnak teljesnek kell lennie, hogy össze lehessen kötni mindegyik mérni kívánt galaxis SED-jét egy sablonnal, különben szisztematikus hiba jelenik meg a mérésben, viszont a túl sok sablon degenerációhoz vezethet. A sablonillesztő és spektrumokon alapuló módszerek egyébb változatait és eredményeik összevetését jól leírják Hildebrandt és társai[6].

Empirikus módszerek alkalmazásához szükség van nagy mennyiségű spektroszkópiával mért vöröseltolódás-adatra és hozzájuk valamilyen, fotometriával mért adatokra. Az egyik legegyszerűbb módszer, hogy a színszűrőkön mért magnitúdóértékekere többváltozós lineáris vagy kvadratikus függvényt illesztenek[7]. Ezeket az adatokat felhasználva a függvényillesztés helyett lehet gépi tanuló eljárásokat is alkalmazni, például *nearest neighbour*[8], *random forest*[9] vagy neurális hálókat[10]. Ezeknek a módszereknek előnye, hogy használatuk viszonylag egyszerű, hatalmas adathalmazokon is működhetnek, illetve nincs szükség a SED-re se, de nagy mennyiségű és jó minőségű tanulóhalmaz kell az előkészületekhez.

2.2. A gépi tanulási módszerek

A gépi tanuló módszerek már a XX. század közepén megjelentek, de a számítástechnika és a rendelkezésre álló adatok mennyisége még nem állt olyan szinten, hogy töretlenül fejlődhessen. Az akkori megközelítés szorosan összefüggött a mesterséges intelligencia kutatásával, de az 1990-es években az irány eltolódott a gyakorlati jellegű problémák megoldása felé. Ma már egyre több használható adat és fejlett grafikus processzorok mellett sokféle gépi tanuló algoritmus lett implementálva, így a nehézségek a megfelelő módszer megtalálása és alkalmazása az adatokra, illetve az adatok használható formába hozása. A gépi tanulás célja, hogy a gép *megértse* az adatok szerkezetét, felismerjen egy szabályt és ez alapján jóslatokat tegyen.

Három nagyobb kategóriába sorolhatjuk a módszereket a rendelkezésre álló adatok és problémák alapján. Az egyik csoport a felügyelt tanulás, ilyenkor rendelkezésünkre álló adatok jelöltek, van egy *ground truth*, amit a modell jóslatainak meg kell közelítenie. A felügyelt tanulási módszereket osztályozás és regressziós problémák megoldásához használják, napjainkban az egyre nagyobb felcímkézett adathalmzoknak köszönhetően egyre több problémára tudják alkalmazni. A felügyelet nélküli tanulási módszerek felcímkézetlen adatokkal dolgoznak, feladatuk, hogy felfedjék a az adtokban rejtett struktúrákat. Az *ground truth* hiánya miatt nem lehet jellemezni a tanulás minőségét számértékkel. A fotometrikus vöröseltolódásbecslésben a várt végeredmény jól meghatározott, ezért felügyelt tanulási módszereket alkalmaznak [10], [12]. Ezeknek a módszerek megértéséhez fontosnak tartom bemutatni a munkám során használt eljárások általános a működési elvét, és alkalmazhatóságának határait.

Felügyelt tanulásnál rendelkezésünkre áll egy $(x_1, y_1), (x_2, y_2), ...(x_N, y_N)$ adathalmaz, ahol x_i egy mintát jelöl és a tulajdonságok(features) számával megegyező dimenziójú vektor, y_i jelöli az osztálycímkét vagy a minta értékét regressziós problámákban, dimenziója az osztályok számával megegyező. A feladat, hogy a gép megtalálja azt a leképzést, ami a lehető legkisebb hibával képez x_i-ből y_i-be még nem látott minták esetén is. A hiba mérését az adatokhoz és a feladathoz illő metrikával kell végezni. A kutatáshoz kétféle módszert használtam, *random forest regressort* és neurális hálókat.

2.2.1. *Random Forest*

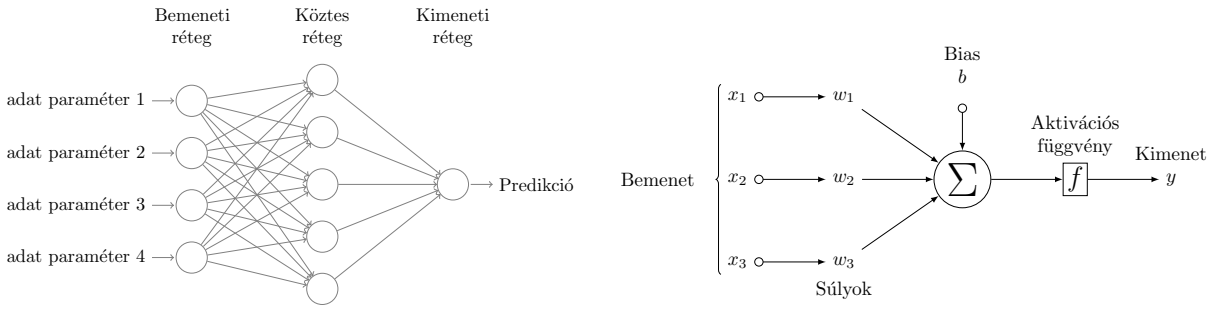
A *random forest* algoritmus egyik jó tulajdonsága, hogy alkalmazható klasszifikációs és regressziós problémákra is, alapját a döntési fák sokasága képezi, amiket összefűzve pontosabb becslést tud adni. A döntési fák felépülése a gyökér csomópontnál kezdődik, ami tartalmazza a tulajdonságokat és a célértéket. A csomópontnál meglévő jellemzőket felosztják az így létrejövő utódpontok között, igyekezve a hibát minimalizálni. A folyamatot tovább iterálva létrejön egy rétege a csomópontoknak, amelyek így egy fát alkotnak. A fa kialakulását szabályozni lehet, paraméterek lehetnek, hogy hány hány rétegű, milyen mély legyen a fa, minimum hány elem kerüljön egy levélbe[1], minimum hány felé legyen osztva a minta egy csomópontnál, vagy milyen módon mérje a szétválasztás minőségét. Véletlenség az erdőbe úgy kerül, ha az egyes fáknál a szétválasztás a csomópontoknál nem a lehető legjobb *split* szerint történik, hanem véletlenszerűen kiválasztott részét kapják meg a tulajdonságoknak. Erre azért van szükség, mert a fontosabb tulajdonságok dominálnának mindegyik fa döntésében, így az egyes fák predikciói korreláltak lennének [13]. Az erdőbe több fát adva az algoritmus nem tanul túl, becslései jobban konvergálnak a kívánt végeredményhez, de határértéke van a hibának [14]. Előnye még, hogy felfedi a prediktálás szempontjából fontos tulajdonságokat, sok attribútummal rendelkező adathalmazoknál ez segíthet az összefüggések megértésében. A fő korlátai, hogy a túl sok fa lassú prediktálást eredményez, illetve regressziós problámáknál a modell nem tud extrapolálni, csak a tanulóhalmaz értékkészletein belüli eredményeket ad. Komplexebb vagy zajosabb adatoknál érdemes máshogy próbálkozni, ezen esetekben például mesterséges neurális hálókkal jobb ereményt lehet elérni.

2.2.2. Mesterséges neurális hálók

A mesterséges neurális háló egy leegyszerűsített modellje a biológiai neurális hálónak, megtartva annak jó tulajdonságait és a tanulás mechanizmusát. Alap építőkövei a neuronok és a súlyok[2], a neuronok a súlyokon keresztül vannak összekötve egymással és ezek adják

[1] levél a döntési fának az a része, amiből nem nő ki több ág

[2] biológiai analógiája az axonok



2.1. ábra. Egy nerális háló vázlatos modellje és egy neuron aktivációjának a folyamata.

meg, hogy az egyik neurontól a másik milyen súllyal kapja meg az értékét. A neuronok rétegekbe vannak rendezve: bemeneti réteg, köztes réteg és kimeneti réteg[3], a köztes réteg több rétegből is állhat. Az egyazon rétegben lévő neuronok nincsenek összekötve egymással, értéküket az alattuk lévő neuronoktól kapják a súlyokkal számolva. Matematikai formában az értékátadás:

$$z_i = w_{i1}x_1 + w_{i2}x_2 + ... + w_{in}x_n + b_i, \tag{2.1}$$

az egyenletben z_i a rétegben az i-edik neuron, w_{ij} az x_j előző rétegbeli neuron és z_i neuront összekötő súly értéke, b_i pedig a *bias* vektor i-edik eleme.

Ez átírható egy egész rétegre:

$$\underline{z} = \mathbf{W}\underline{x} + \underline{b} \tag{2.2}$$

A $\mathbf{W}$ a w_{ij} súlyokból képzett súlymátrixot jelöli. Ahhoz, hogy a háló bonyolultabb problémákat is meg tudjon oldani, szükséges nemlinearitást vinni a rendszerbe. Ehhez aktivációs függvényt alkalmazunk, ami eldönti, hogy a neuron aktivizált legyen vagy sem. Gyakran alkalmazott aktivációs függvény a sigmoid:

$$S(z) = \frac{1}{1 + e^{-z}} \tag{2.3}$$

Illetve a *rectified linear unit*(ReLU):

$$R(z) = \max(0, z) \tag{2.4}$$

[3] input layer, hidden layer és output layer

A ReLU előnye, hogy a súlyok optimalizálásánál történő deriválásoknál az eltünő gradiens problémája nem áll fenn, így gyorsabb tanulást eredményez és a sigmoidhoz képest nem olyan szűk intervallumon ad vissza értékeket. A bemenő adatok ilyen transzformációkon mennek keresztül, míg kimenő adatként össze lehet hasonlítani a várt kimenettel. Az összehasonlítás a hibafügvénnyel történik, ami kiszámolja a két érték közötti távolságot valamilyen metrikával. Legtöbbször a hibafüggvényt több különböző adat becslése után értékeltetjük ki, a háló *batch*-okban kapja meg az adatot. Regressziós problémákban kedvelt hibafüggvény a *mean sqared error*:

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - p(x_i))^2 \tag{2.5}$$

a képletben a $p(x_i)$ az x_i adatokból prediktált érték, n pedig a *batch* méret. A hiba minimalizálása egy optimalizációs probléma, a p függvény a súlyoktól függ, amiket úgy kell megváltoztatni, hogy a hiba minimális legyen. A súlyok optimalizálása *backpropagation*-nel történik, a hibafüggvényt deriválva a súlyok szerint a láncszabállyal, az utolsó rétegtől kezdve az elejéig, megkapjuk a gradiensmátrixokat rétegenként, amit szorozva egy kis számmal(*learning rate*), levonva a súlymátrixokból megkapjuk az új súlyokat. A súlyok frissítése *batch*-onként történik, így a *batch*mérettel és a *learning rate*-tel is lehet a tanulást gyorsítani, majd pontosabbá tenni [16].

A hálók architektúrájának kialakítására nincs általános szabály, intuíciónkra és hasonló problémámkon jól szereplő hálók mintájára kell hagyatkozni, tanulásából következtetéseket levonni és alakítgatni. Általánosságban a komplexebb problémákhoz *mélyebb*, több neuronból álló hálók kellenek, de ekkor a beállítandó paraméterek száma is nagyobb lesz, a tanítás lassabb lesz a megnövekedett számítási igény miatt. Képi adatoknál a paraméterek száma nagyon magas lenne, valamint a képnek nem az egész része érdekes, csupán részletek, és ezek a részletek bárhol lehetnek a képen, ezért nem célszerű teljesen összekötött neurális hálókat használni. Ezeknek a problémáknak a megoldására találták ki a konvolúciós neurális hálókat.

A konvolúciós hálókban a rétegek között nincs teljes összeköttetés, a neuron csak az alatta lévő neuronnal, és annak szomszédaihoz kötődik. Az objektumfelismerő-osztályozó problémáknál az objektum bárhol előfordulhat a képen, ezért a neuronokat összekötő súlyoknak is

eltolás invariánsank kell lennie, ezért egy rétegben minden neuron ugyanazokkal súlyokkal összegzi az allata lévő neuron szomszédainak kimenetét. A figyelembe vett szomszédok száma meghatároz egy *filter* méretet, a *filter* technikailag egy tenzor, aminek elemei a súlyok. A bemeneti képet végig páztázza a filter, és végrehajta az összegzéseket, az egyes kimenetek egy *aktivációs térképet alkotnak*, amire alkalmazhatjuk a következő réteg konvolúciós réteget. Ez a műveleletet kétdimenziós diszkrét konvolúció a matematikában. A kimeneti *aktivációs térkép* dimenziói egy n széles és c csatornával rendelkező képen alkalmazott $f \times f$ méretű, s ugrással mintavételező *filter*-rel, a képre p *padding*[4]-et alkalmazva a konvolúció során a következő képpen alakulnak:

$$a = \frac{n + 2p - f}{s} + 1, \tag{2.6}$$

a képleteben a létrejövő *aktivációs térkép* szélessége, k *filter*t alkalmazva $a \times a \times k$ dimenziós

Eddigi eredmények és módszerek

[4] kép szélén nullákkal létrehozott keret

3. fejezet

Adatok

Az eredményes gépi tanuláshoz a nagy mennyiségű és jó minőségű adat majdnem annyira fontos, mint a jó modell megválasztása. Általában nem áll rendelkezésünkre egyből az algoritmusnak adható adat, a nyers adatokat előbb preprocesszállni kell, ki kell nyerni a fontos tulajdonságokat, melyek előztetes ismereteink alapján fontosak lehetnek, és össze kell állítani az adathalmazt, amelyet majd szétválasztunk tanuló- és teszthalmazra.

A kutatómunkámhoz sok, jó minőségű galaxis képére volt szükségem, és mindegyik galaxisnak a spektroszkópiával mért vöröseltolódására is, ezekhez az adatokhoz a Sloan Digital Sky Survey adatbázisában fértem hozzá.

3.1. Sloan Digital Sky Survey

A Sloan Digital Sky Survey az eddigi legrészletesebb égboltfelmérés, mély, több színsávos képekkel az ég egyharmadáról, 500 milló asztrofizikai objektumról, 3 millió felvett spektrumadattal. A felmérést 2000-ben kezdte, több ciklusban, 2014-ben kezdődött a negyedik fázis (SDSS-IV), és már elkezdődtek a megbeszélések az SDSS-V elindításáról[17]. A megfigyelési adatokat egy külön erre a célra megépített 2.5 m széles optikai teleszkóp szolgáltatja az Apache Point Obszervatóriumból, Új Mexikóban. Öt színsávaban mér, a látható fény és az infravörös tartománya között, u, g, r, i és z színszűrőkkel (ultraviolet, green, red,

near infrared és infrared), amik 3000 Å és 10000 Å közötti tartományt fedik le. A különböző szűrőkön keresztül a CCD chipek egymás után rögzítenek, 71.2 másodperc késéssel, r, i, u, z és g sorrendben, így a különböző színű képeken előfordulhat, hogy egy objektum kicsivel arrébb. Ennek volt előnye is, mozgó objektumokat könnyebben lehetett azonosítani, például létre tudtak hozni aszteroida katalógusokat. Az SDSS képek alapvető egysége a *field*, ami 10-szer 13 szögperces szeletet tartalmaz az égből, és 1489-szer 2048 pixelt tartalmaz és az egyes *fieldeket* három jelzőszám teszi egyedivé, a *field number*, a *run*, ami a szkennelés száma, a *camera cloumn*, ami a megmutatja melyik oszlop CCD kamera készítette a képet egy 1 és 6 közötti szám, mindegyik oszlop egy *fieldet* készít, ami 128 pixel szélességben fed át a szomszédossal. Az elkészült felvételek az SDSS képfeldolgozó pipelinejába kerülnek, ahol kalibrált FITS formátumú képet csinálnak, és a katalógushoz hozzá adják a képi paramétereket.

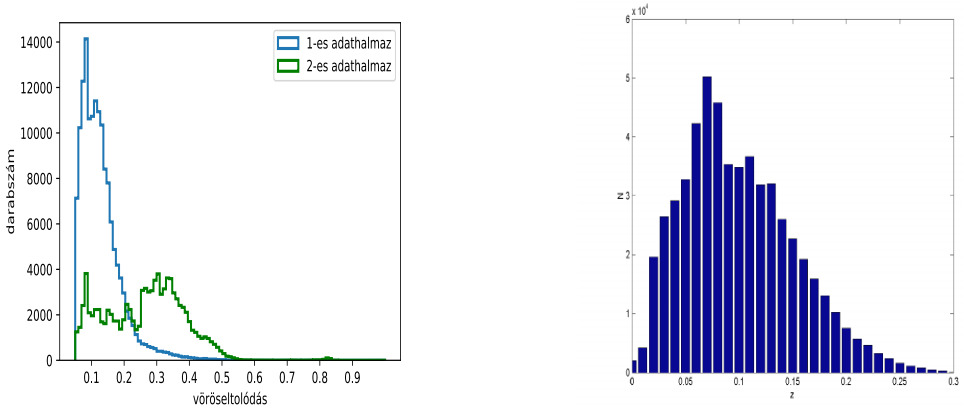
A megfigyeléseket folyamatosan végzik, az adatbázisokat viszont csak évente frissítik, ezeket a *data release*-nek (DR) hívják és tartalmazzák az előző megfigyelésekből származó adatokat is. Az SDSS virtuális oszervatórium keretrendszerben működik, adatai publikusak. Az objekumok és paramétereik relációs adatbázisba vannak rendezve, a SQL nyelvet használva lehet lekérdezéseket indítani.

3.2. A tanulítóhalmaz elkészítése

3.3. Adatexploráció

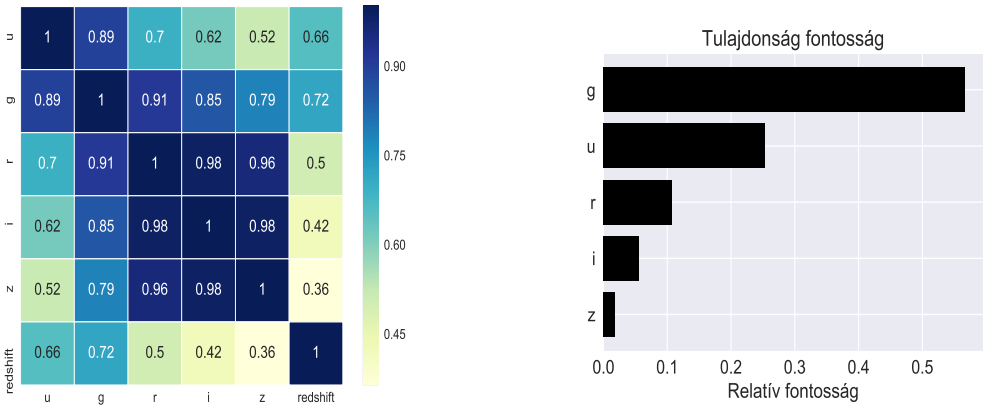
Mielőtt elkezdenénk a tanítást, érdemes megvizsgálni a tanulóhalmazunk statisztikai jellemzőit, megnézni mennyire jól reprezentálja a valóságot, valamint feltárni az összefüggéseket. Ehhez az SDSS *DR7*-es adatbázisának vöröseltolódás-eloszlásával hasonlítottam össze a tanulóhalmaz eloszlását, alkalmaztam egy *Random Forest regreesort* az *ugriz* magnitúdóértékekre, és készítettem egy *korrelációs térképet* a magnitúdókkal és vöröseltolódásokkal.

Az SDSS *DR7*-ben a galaxisok vöröseltolódásának mediánja $z \sim 0.07$ [18], a teszthal-



3.1. ábra. Az általam előállított tanulóhalmaz és az SDSS DR7 vöröseltolódás-eloszlása. A jobboldali kép forrása: [18]

mazomé $z \sim 0.11$, ami a 0.05-ös vöröseltolódásbeli vágásomnak tudható be. Összehasonlítva a 3.2. ábrán a baloldalon az 1-es adathalmaz hisztogramját a jobboldalival, a két eloszlás jó hasonlóságot mutat, a vágásban van különbség, a legtöbb galaxisnak mind a kettő esetben 0.3-nál kisebb a vöröseltolódása.



3.2. ábra. A korrelációs térkép, a Random forest tanulásában a különböző tulajdonságok relatív fontossága.

4. fejezet

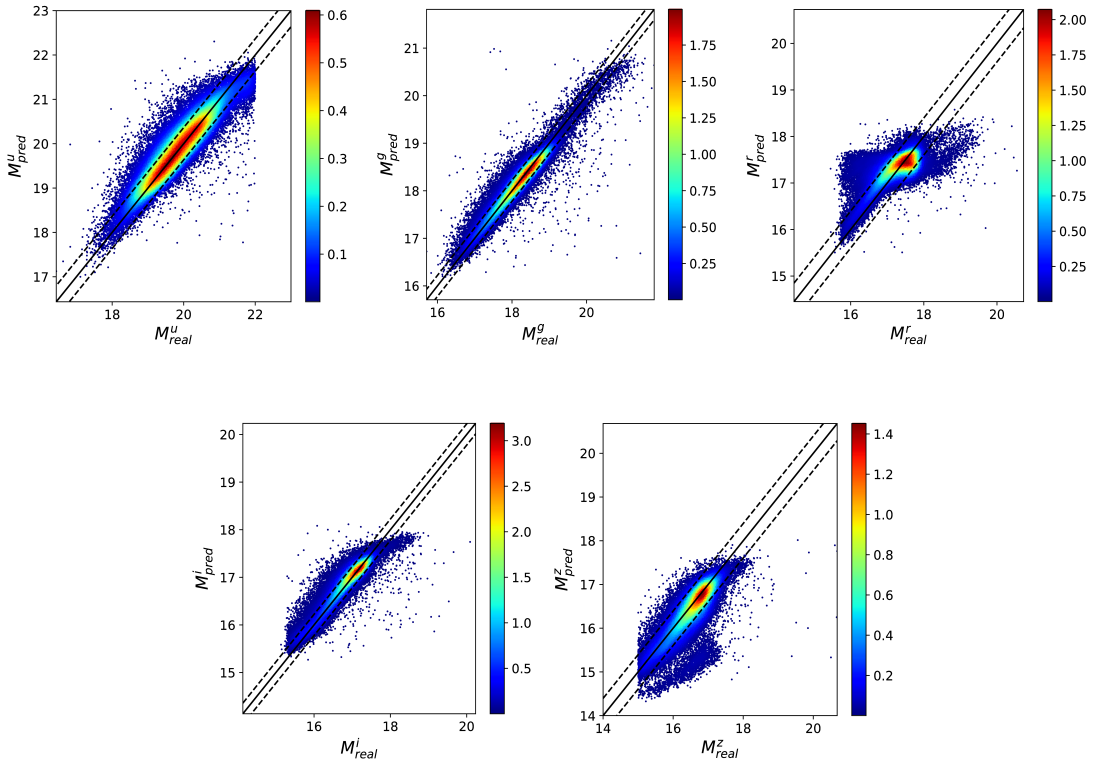
A kutatás

4.1. Neurális háló és Random Forest kombinálva

A *random forest* algoritmus jól teljesít a magnitúdókból prediktálás során, így célszerű megvizsgálni, hogy mennyire működik jól, ha a magnitúdóértékek is becslésből származnak. A képekből való fényességbecsléshez mind az öt színszűrőhöz elkészítettem egy-egy konvolúciós neurális hálót, ezeket a hálókat az első 50.000 képen tanítottam be. A következő 50.000 képpel prediktáltam magnitúdóértékeket, ezeket a becsült magnitúdókat használtam, a *random forest regressor* tanításához. A harmadik 50.000 képnek is megbecsültem a magnitúdóit, és ezekből predikált a véletlen erdő vöröseltolódás-értékeket, így nem volt átfedés a tanuló-és teszthalmazok között. A neurális hálók tanításánál 80 *epoch*-ot használtam mindegyik hálóra, de nem egyben ment végig a tanulás, a *learning rate*-t csökkentettem, ha a hibafüggvény értéke nem csökkent, illetve a *batch* méretet is növeltem közben. Az egymás melletti[1] magnitúdók korrellálása, illetve a célváltozók és a bemeneti adatok hasonlóság miatt az volt a feltételezésem, hogy ugyanolyan architektúrájú neurális hálókat lehet használni mindegyik fényességérték számításához.

Az egyes hálók bemenetként megkapták az $50 \times 50 \times 1$ méretű, az adott színszűrőn keresztül készített képet inputként. Az első réteg egy konvolúciós réteg volt, 32 darab 2×2-es

[1] hulámhossz szerint



4.1. ábra. Az egyes neurális hálók magniúdópredikciója ábrázolva a valódi magnitúdóértékek függényében.

filterből állt, alkalmaztam rá a *ReLu* aktivációt és utána egyből egy 2×2 *MaxPooling*-ot. Utána a következő konvolúciós réteg 16 db 2×2-es filterből állt, *sigmoid* aktivációval, majd egy 2×2-es *MaxPooling* következett. A kijövő értékek kilapítása után következtek a teljesen összekötött rétegek, az első réteg 1024 neuronból állt, *sigmoid* aktivációval. Utána 15%-os *Dropout* regularizáció következett, ami a túlillesztés elkerülésére szolgált. Ezt követően 512 neuron, *Relu* aktiváció és 10%-os *Dropout*, 256 neuron *ReLu*-val és végül az egy darab neuron. A hibafüggvény *mean squared error* volt. A *random forest regressor* 250 fából állt, maximális mélysége 15, egy levélbe minimum 120-elemnek kellett kerülnie, egy szétválasztáshoz legalább 28 minta volt szükséges és a szétválasztás jóságát *mean squared error*-ral mérte.

Irodalomjegyzék

[1] Z. Frei and A. Patkós, Inflációs Kozmológia: (Typotex, Budapest, 2005).

[2] Baum, W. A.: 1962, Problems of Extra-Galactic Research, Proceedings from IAU Symposium no. 15. Edited by George Cunliffe McVittie. International Astronomical Union Symposium no. 15, Macmillan Press, New York, p.390

[3] "Photometric Redshifts." NASA/IPAC Extragalactic Database - NED, ned.ipac.caltech.edu/level5/Glossary/Essay_photredshifts.html.

[4] Koo, D. C. "Optical Multicolors - A Poor Person's Z Machine for Galaxies." The Astronomical Journal, vol. 90, 1985, p. 418., doi:10.1086/113748.

[5] Bruzual, G., and S. Charlot. "Stellar Population Synthesis at the Resolution of 2003." Monthly Notices of the Royal Astronomical Society, vol. 344, no. 4, 2003, pp. 1000–1028., doi:10.1046/j.1365-8711.2003.06897.x.

[6] Hildebrandt, H., et al. "PHAT: PHoto-ZAccuracy Testing." Astronomy & Astrophysics, vol. 523, 2010, doi:10.1051/0004-6361/201014885.

[7] Connolly, A. J., et al. "Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry." The Astronomical Journal, vol. 110, 1995, p. 2655., doi:10.1086/117720.

[8] Csabai, I., et al. "The Application of Photometric Redshifts to the SDSS Early Data Release." The Astronomical Journal, vol. 125, no. 2, 2003, pp. 580–592., doi:10.1086/345883.

[9] Carliles, S., et al. “Random Forests For Photometric Redshifts.” The Astrophysical Journal, vol. 712, no. 1, 2010, pp. 511–515., doi:10.1088/0004-637x/712/1/511.

[10] Collister, AdrianÂ A., and Ofer Lahav. “ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks.” Publications of the Astronomical Society of the Pacific, vol. 116, no. 818, 2004, pp. 345–351., doi:10.1086/383254.

[11] Csabai, I., et al. “Multidimensional Indexing Tools for the Virtual Observatory.” Astronomische Nachrichten, vol. 328, no. 8, 2007, pp. 852–857., doi:10.1002/asna.200710817.

[12] Collister, AdrianÂ A., and Ofer Lahav. “ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks.” Publications of the Astronomical Society of the Pacific, vol. 116, no. 818, 2004, pp. 345–351., doi:10.1086/383254.

[13] Ball, Nicholas M., and Robert J. Brunner. “Data Mining And Machine Learning In Astronomy.” International Journal of Modern Physics D, vol. 19, no. 07, 2010, pp. 1049–1106., doi:10.1142/s0218271810017160.

[14] “Random Forests Leo Breiman and Adele Cutler.” Statistics at UC Berkeley, www.stat.berkeley.edu/ breiman/RandomForests/.

[15] Sadeh, I., et al. “ANNz2: Photometric Redshift and Probability Distribution Function Estimation Using Machine Learning.” Publications of the Astronomical Society of the Pacific, vol. 128, no. 968, 2016, p. 104502., doi:10.1088/1538-3873/128/968/104502.

[16] L., Samuel, et al. “Don’t Decay the Learning Rate, Increase the Batch Size.” SAO/NASA ADS: ADS Home Page, 1 Nov. 2017, adsabs.harvard.edu/cgi-bin/bib_query?arXiv%3A1711.00489.

[17] Zasowski, Gail. Science Blog from the SDSS, blog.sdss.org/2018/02/21/sdss-v-is-underway/.

[18] Verevkin, A. O., et al. “The Non-Uniform Distribution of Galaxies from Data of the SDSS DR7 Survey.” Astronomy Reports, vol. 55, no. 4, 2011, pp. 324–340., doi:10.1134/s1063772911020089.

Nyilatkozat

Név: Horváth Bendegúz

ELTE Természettudományi Kar, szak: Fizika BSc

Neptun azonosító: ZNL3LK

Szakdolgozat címe: Fotometrikus vöröseltolódás becslés

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló munkám eredménye, saját szellemi termékem, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest 2018. majus ?.

_____