2021 年度卒業

# 修士論文

空間解像度差のあるデータセットを用いた深層学習による銀河形状分類精度

日本語タイトル暫定版なため、ここ英語タイトルも未完！


| | |
|---|---|
| 所属 | 新潟大学自然科学研究科 電気情報工学専攻・飯田佑輔研究室 |
| 在籍番号 | F20C026D |
| 氏名 | 本間 裕也 |

## 概要

日本語のアブストラクト

## Abstract

English Abstract Here

# 目 次

# 1　はじめに

# 2　深層学習

## 2.1　パーセプトロン

深層学習の説明を行う前に,

## 2.2　ニューラルネットワーク

ニューラルネットワークとは,

### 2.3　損失関数と重み更新

深層学習の学習で用いられる指標を，損失関数と呼ぶ．損失関数には様々な種類が存在し，解く問題の種類によって使い分ける．一般的な損失関数として，式 (1) の 2 乗平均誤差 (主に回帰問題に使用) や，式 (2) のクロスエントロピー誤差 (主に分類問題に使用) が挙げられる．

$$E = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{1}$$

$$E = sans \tag{2}$$

# 3　使用するデータセット

## 3.1　Sloan Digital Sky Survey(SDSS)

### Sloan Digital Sky Survey (以下，SDSS) について

SDSS**??**とは，天文学史上最も重要なサーベイ観測プロジェクトの一つとも称される大規模プロジェクトである．このプロジェクトは全天の約 1/4 の天域において，天体の画像および分光データを取得し，天体カタログを作成することを目的としたものである．SDSS での撮像・分光データ取得は CCD を搭載した地上望遠鏡によって行われる．SDSS の最初のプロジェクトである SDSS-I は 2000 年から 2008 年まで行われ，また対象を銀河系や超新星に絞った SDSS-II が 2005 年から 2008 年に SDSS-I と並行して実施された．その後，太陽系外惑星調査や天の川銀河の構造及び進化などに焦点を当てた SDSS-III が 2008 年から 2014 年に，南天・北天からの銀河系探索などを目的とした SDSS-IV が 2014 年から 2020 年に行われた．

SDSS で撮影された天体のうち低赤方偏移銀河は後述する Galaxy Zoo によって形態分類ラベル付けが行われている．SDSS と Galaxy Zoo から天体の画像データと分類ラベルを取得できることから，これらのセットを今回の銀河形態分類モデル作成に用いる．

### 実験に用いる銀河画像データの作成方法

今回分類モデル学習に用いる銀河画像データとして，SDSS-II におけるデータリリースの中から，Data Release 7 (以下，DR7)**??**より画像データの取得を行った．DR7 を選んだ理由としては，Galaxy Zoo における銀河形態ラベル付けに DR7 の銀河画像が用いられたからである．

データベースから取得できるのはある程度大きな天域の全体画像のため，用いたい銀河の画像を取得したい場合は，全体画像から切り出しを行う必要がある．今回は Galaxy Zoo にて形態ラベル付けが為されている銀河の中から 15,000 天体を対象に，銀河毎に 64 ピクセル四方のサイズで切り出しを行った．銀河切り出し画像の生成概略図を図**??**に示す．

DR7 におけるデータの撮影が行われた SDSS-II において，銀河撮像に用いられた測光フィルタは u, g, r, i, z の 5 つが存在し，これらのフィルタを使用し 5 つの帯域画像が撮影された (図 1 参照)．これら 5 つの帯域画像のうち，今回の実験では r フィルタから得られた帯域画像 (r バンド画像) を使用している．r バンド画像を使用した理由としては，5 つの帯域画像のうち可視光帯画像である g, r, i バンド画像の中で，最も平均値に近い波長を捉えている r バンド画像がより多くの銀河形態的特徴を有していると考えられること，また r フィルタが 5 つの測光フィルタのうち最も感度がよいことが挙げられる**??**．

| *u* | *g* | *r* | *i* | *z* |
|---|---|---|---|---|
| 3551Å | 4686Å | 6165Å | 7481Å | 8931Å |
| 22.0 | 22.2 | 22.2 | 21.3 | 20.5 |

図 1: SDSS Data Release 7 における，銀河撮像に用いられた測光フィルタ一覧
(フィルタ名，各フィルタによって撮影された画像の波長平均値)

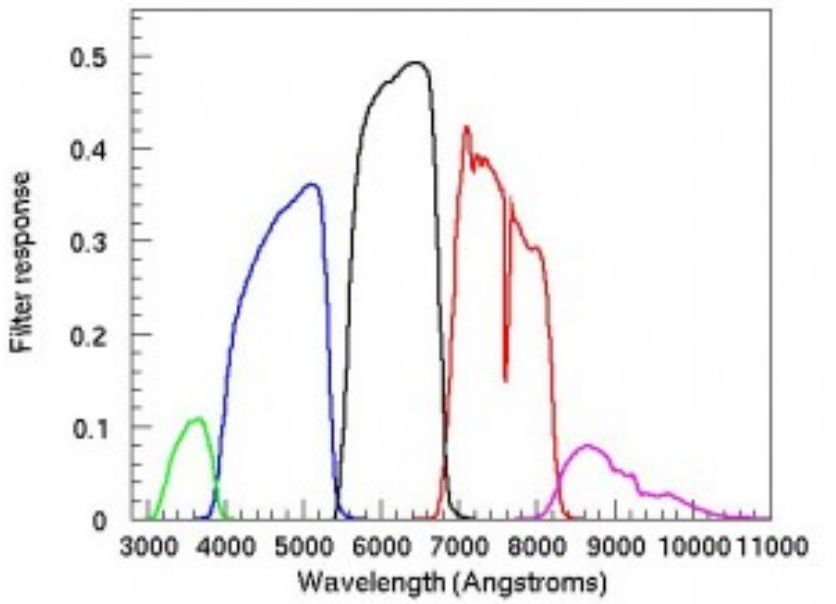


図 2: SDSS における測光フィルタのスループット曲線
(https://www.sdss.org/instruments/camera/ より引用)

### 3.2 Galaxy Zoo (GZ)

**Galaxy Zoo (以下，GZ) について**

Galaxy Zoo**??**とは，人間の目による分類を行い銀河形態カタログを作成したプロジェクトである．従来の銀河形態分類は専門知識を持った天文学者のチームによって行われてきたが，SDSS のような何十万もの銀河が格納されたデータセットが登場するに従い，天文学者だけでは銀河データの増加に追いつけなくなってきた．このような状況を打破する方法として，インターネットを通じて専門知識を持たない有志の一般人に投票形式で形態分類を行わせる方法が提案された．こ

れが GZ である．

GZ の最初の調査である Galaxy Zoo 1??では渦巻銀河か楕円銀河の区別や，渦巻銀河であった場合にどちら周りに渦巻いているかなどを投票形式で集計し，6 つのカテゴリへと分類が行われた．GZ1 の後継プロジェクトである Galaxy Zoo 2??では，GZ1 より詳細な形態分類を行うため 11 つの質問が用意されている．

### GZ1 における形態ラベル付け方法

GZ1 では，ウェブサイト (www.galaxyzoo.org) を通じて有志の一般人の形態分類を投票形式で集計した．GZ1 における分類の投票ページを図 3 に示す．サイトを訪れた有志の一般人は，楕円銀河・時計回り渦巻銀河・反時計回り渦巻銀河・エッジオン (渦が地球から観測できない向きについている銀河)・星もしくは区別できなかった天体・マージャー (2 つの銀河が衝突し合体している，合体銀河) の 6 つのうちいずれかに投票を行う．形態分類カタログは，対象となった銀河に対し，各形態分類の投票率を付与され作成される．


図 3: GZ1 における形態分類の投票ウェブページ

(??より引用)

### 実験に用いる正解ラベルの作成方法

今回分類モデル学習に用いる正解ラベルとして，GZ1 から取得できる Table2 の分類フラグを用いた．Table2 の記載内容を示した図を図 4 に示す．Table2 は SDSS から取得した天体のうち，分

光スペクトルデータが観測された天体に関して収録されている．Table2 には以下の情報が記載されている．

- SDSS における天体オブジェクト ID
- 天体の天球座標 (RA, Dec)
- 投票数
- 各カテゴリの得票率
- バイアスが除去された投票率
- 分類フラグ (渦巻銀河・楕円銀河・不確かな天体)

分類フラグは，渦巻銀河・楕円銀河・不確かな天体の 3 つが存在する．それぞれの銀河種に対し，得票率が 8 割を超えた場合にフラグが立つ．今回はこの分類フラグを深層学習モデルの学習に用いる．

| OBJID | RA | DEC | NVOTE | P_EL | P_CW | P_ACW | P_EDGE | P_DK | P_MG | P_CS | P_EL_DEBIASED | P_CS_DEBIASED | SPIRAL | ELLIPTICAL | UNCERTAIN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | hms | dms | | | | | | | | | | | | | |
| 587727178986356823 | 00:00:00.41 | -10:22:25.7 | 59 | 0.61 | 0.034 | 0.0 | 0.153 | 0.153 | 0.051 | 0.186 | 0.61 | 0.186 | 0 | 0 | 1 |

図 4: Table2

# 4　SDSS & Galaxy Zoo を用いた分類モデル学習

この章では当論文の実験で用いられる SDSS と GZ にて，学習データとテストデータの解像度が揃っているという条件のもと，高精度分類が行える分類モデルを学習させることを目的とし，その結果 2 つの実験を行った．第 4 章を行う動機は，当論文で掲げている将来展望の前提条件を達成することである．

当論文の将来展望は，高空間分解能観測装置データを用いてモデル学習を行うことで，既存の低空間分解能データセットに対し更なる高精度形態分類を提供するというものである．この将来展望にまつわる実験の最も初段階の前提条件として，まずは学習データとテストデータの解像度が揃っている条件にて高精度分類が行えるかを検証する必要がある．そこで 2 つの実験を行った．

## 4.1　実験概要

第 4 章では SDSS から取得した銀河画像と GZ から取得した分類フラグを学習データとし，渦巻銀河・楕円銀河・不確かな天体のいずれかを予測する分類モデルを学習する．

**学習データ**　学習に用いる画像データは，SDSS から取得した 64 ピクセル四辺の銀河切り出し画像を使用し，正解ラベルは GZ から取得した分類フラグを用いる．実験には 15,000 天体を用いた．15,000 天体の赤方偏移別の個数を示したグラフを図 5 に示す．一般的に赤方偏移の値が小さいほど，地球から近い天体といえる．15,000 天体のうち，GZ の分類フラグにて渦巻銀河と分類されている天体は 4,058 天体，楕円銀河は 1,561 天体，不確かな天体は 9,310 天体であった．

モデルの学習および評価の際，モデルの学習データとテストデータの比率は 7:3 とした．

**モデル構造**　今回用いた深層学習モデルは，cheng et al.(2019)**??**にて用いられていた銀河形態分類モデルを参考にした．今回用いたモデルの構造を図 6 に示す．このモデルは畳み込み層を合計 3 つ有しており，それぞれのカーネルサイズは 3x3, 3x3, 2x2 である．それぞれの畳み込み層の後には，2x2 の max-pooling 層が存在する．全畳み込み層の後に全結合層が 2 層配置されており，それぞれ 1024 個のノードを有している．

**モデルの評価方法**　分類モデルの評価指標として，主に accuracy (正解率)，そして一部実験において True Skill Statistics (以下，TSS) を用いた．accuracy は全テストデータの中で正しく分類できたデータがどれだけあるかというものであり，モデルの正確性を表す指標である．TSS も同じく正確性を表す指標であるが，テストデータ内のデータインバランス性に対しロバストな性質がある．

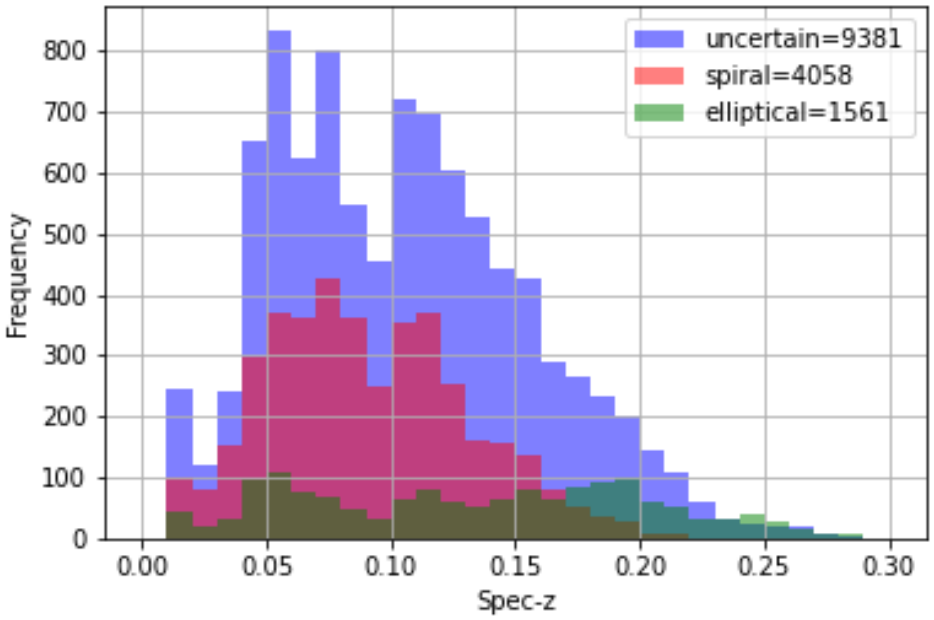


図 5: 実験に用いる 15,000 天体の赤方偏移別の個数グラフ

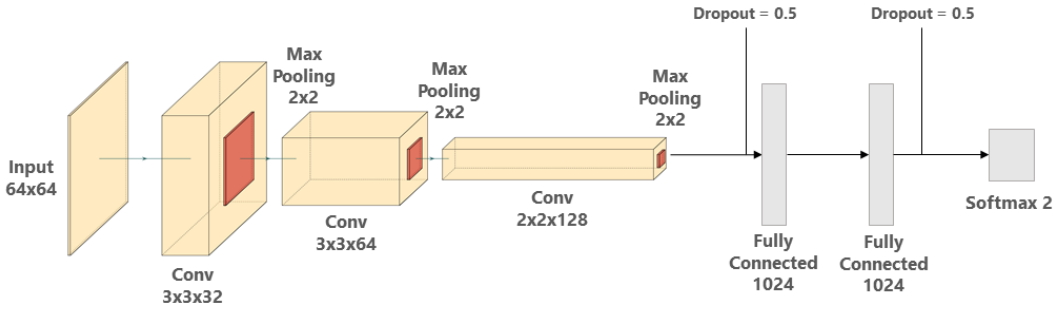


図 6: 用いた分類モデルの構造図

以下に 2 値分類問題を例とした場合の，accuracy, TSS の導出方法を示す．ここで混同行列とは，分類問題においてモデルが予測した値および真の値を行列形式で表したものであり，分類モデルの性能を評価または可視化するのによく用いられる指標である．モデルの予測値と真の値が交差する対角成分における数が多いほど，モデルが正確な予測を行っているといえる．

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{TN + FP} \tag{4}$$

また，深層学習モデルには実行毎に学習のブレが存在するため，モデルの評価の際には学習・テ

ストを 30 回実行し，評価指標の平均値，標準偏差および標準誤差を導出する．

## 4.2　不確かラベルが与える分類精度への影響

GZ が与える形態分類フラグのうち「不確かな天体」というフラグが立っている天体は，渦巻銀河・楕円銀河のどちらの得票率も 8 割を超えなかった天体である．これらの天体は人間の目による形態分類が比較的難しい天体，つまり特徴があやふやな天体であると考えられる．

この節では特徴があいまいだと考えられる不確かな天体群が，分類精度に与える影響を調べる．具体的には以下の 2 条件で実験を行い，テストデータに対する予測の accuracy や混同行列の比較を行う．

- 不確かを含めた、渦巻・楕円・不確かの 3 値分類
- 不確かを除いた、渦巻・楕円の 2 値分類

### 4.2.1　実験条件

モデルの学習およびテストを行う際，1000 天体を使用した．使用する天体は，GZ による形態分類が行われた 15,000 天体 (図 5 参照) からランダムに取得を行う．使用天体を 1000 天体取得したあと，学習データとテストデータの比率が 7:3 となるように切り分けを行った．

モデルの評価を行う際，モデルの学習およびテストを 30 回行い，accuracy の平均値，標準偏差および標準誤差を導出した．なお，30 回の学習およびテストの際，用いられる天体は毎回ランダムに取得される．

### 4.2.2　実験結果

**学習結果**　不確かを含めた渦巻・楕円・不確かの 3 値分類モデル，および不確かを除いた渦巻・楕円の 2 値分類モデルの学習結果を図**??**にに示す．

- 一番上の図は、横軸 epoch 数・縦軸 loss（損失関数）となっている。
- 中段の図は、横軸 epoch 数・縦軸 accuracy（正解率）となっている。
  - これら 2 つの図は、青色のラインが学習データに対するスコア、黄色のラインがテストデータに対するスコアとなっている。

- loss のグラフについて、学習データに対する損失関数は学習が進むにつれ順当に下がっていくものの、ある一定の epoch からテストデータに対する損失関数が上昇していく現象が見受けられる。この現象は**過学習**と呼ばれている。この現象が起こると、学習データに対し分類モデルが過剰に適合した結果、学習データに対する予測精度に比べテストデータに対する予測精度が低下していく。

- 一番下の図はテストデータに対する予測結果から生成した混同行列であり、横軸がモデルによる予測ラベル、縦軸がテストデータの真のラベルである。

**結果**　不確かを含めた渦巻・楕円・不確かの 3 値分類モデル，および不確かを除いた渦巻・楕円の 2 値分類モデルの学習結果を表**??**にに示す.

表 1: fugafuga

| | mean ± std (2) | mean ± ste (2 σ) |
|---|---|---|
| 3 値分類 | 0.688 ± 0.031(148epoch) | 0.688 ± |
| 2 値分類 | pine | banana |

# 5　空間解像度差のあるデータセットを用いた分類モデル学習

# 6　議論

## 6.1　今回の実験から得た結論

## 6.2　将来課題

## 6.3　将来展望

## 7　おわりに

## 謝辞

本研究を進めるにあたり，ご指導を頂いた飯田佑輔准教授および東京理科大学の大井渚様，そしてデータセット作成や宇宙関連知識取得に際し助力いただいた同研究室の津田様に，厚く感謝申し上げます.
また，日常の議論を通じて多くの知識や示唆を頂いた飯田佑輔研究室の皆様に感謝いたします.

SDSS および SDSS-II の資金はアルフレッド・P・スローン財団から提供され，また参加機関は米国科学財団，米国エネルギー省，米国航空宇宙局，日本の文部科学省，マックスプランク協会，英国高等教育基金協会です．SDSS の Web サイトは、http://www.sdss.org/ です。

SDSS は参加機関のための天体物理学研究コンソーシアムによって運営されています。参加機関は、アメリカ自然史博物館、ポツダム天体物理学研究所、バーゼル大学、ケンブリッジ大学、ケース・ウェスタン・リザーブ大学、シカゴ大学、ドレクセル大学、フェルミラボ社、高等研究所、日本参加グループ、ジョンズ・ホプキンス大学、原子核宇宙物理学合同研究所、カブリ粒子宇宙物理学研究所、韓国科学者グループ、中国科学者グループ、中国科学者グループ、韓国科学者グループ韓国科学者グループ、中国科学院（LAMOST）、ロスアラモス国立研究所、マックスプランク天文学研究所、マックスプランク天体物理学研究所、ニューメキシコ州立大学、オハイオ州立大学、ピッツバーグ大学、ポーツマス大学、プリンストン大学、米国海軍天文台、ワシントン大学です.