

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie

Klasifikácia zákrytových premenných hviezd pomocou hlbokého učenia

Bakalárska práca

BINARY STARS

Používateľská príručka

Vedúci bakalárskej práce:

Doc. Ing. Peter Butka, PhD.

Bakalár:

Maximilián Revický

Konzultant bakalárskej práce:

Ing. Viera Maslej Krešnáková, PhD.

Košice 2022

1 Funkcia programu

Úlohou daných skriptov je predpripraviť vstupné dátové množiny do `.pkl` alebo `.csv` súborov a klasifikovať svetelné krivky zákrytových premenných hviezd. Konkrétne vyriešiť nasledujúce klasifikačné úlohy: (1) Binárna klasifikácia dvojhviezd na oddelené a dotykové systémy. (2) Binárna klasifikácia na dvojhviezdy s kritickým sklonom dráhy a škvrnité hviezdy. Úlohy programu sú rozdelené do viacerých skriptov:

V priečinku `data-preparation/` sa nachádzajú:

- Načítanie syntetických dát oddelených kriviek s kritickým sklonom dráhy z databázy a uloženie do formátu `.pkl: binaries_detached_bellow_i_crit_10000.ipynb`
- Načítanie syntetických dát oddelených kriviek z databázy a uloženie do formátu `.pkl: binaries_detached_random.ipynb`
- Načítanie syntetických dát dotykových kriviek s kritickým sklonom dráhy z databázy a uloženie do formátu `.pkl: binaries_overcontact_bellow_i_crit_10000.ipynb`
- Načítanie syntetických dát dotykových kriviek z databázy a uloženie do formátu `.pkl: binaries_overcontact_random.ipynb`
- Načítanie syntetických dát dotykových kriviek z databázy a uloženie do formátu `.pkl: single_spotty.ipynb`
- Načítanie observačných dát dvojhviezd z JSON súborov a uloženie do formátu `.csv: data_observed_binary.ipynb`
- Načítanie observačných dát dvojhviezd s kritickým sklonom dráhy a škvrnitých hviezd z textových súborov a uloženie do formátu `.csv: data_observed_binary_spotty.py`

V priečinku `modeling/` sa nachádzajú:

- Binárna klasifikácia dvojhviezd na oddelené a dotykové systémy, 3 experimenty, 2 modely 1D CNN a BiLSTM RNN, vyhodnotenie modelov na syntetických krivkách: `model_3_experiments_binary_stars_50000.ipynb`
- Binárna klasifikácia dvojhviezd na oddelené a dotykové systémy, vyhodnotenie na observačných dátach: `model_observed_binary.ipynb`
- Binárna klasifikácia na dvojhviezdy s kritickým sklonom dráhy a škvrnité hviezdy, 1 experiment, 1 model 1D CNN a BiLSTM RNN, vyhodnotenie modelu na syntetických krivkách: `model_experiment_single_spotty_critical_binary.ipynb`
- Binárna klasifikácia na dvojhviezdy s kritickým sklonom dráhy a škvrnité hviezdy, vyhodnotenie modelu na observačných krivkách: `model_observed_spotty_binary.ipynb`
- Funkcia šumu `stochastic_noise_generator`, ktorá sa aplikuje na syntetické dáta, aby sa priblížila viac observačným dátam zo skriptu `noise_generator.py`
- Funkcia na náhodné zarovnanie kriviek oddelených a dotykových systémov s kritickým sklonom dráhy `curve_alignement_randomizer` zo skriptu `curve_alignement.py`

2 Súpis obsahu dodávky

Priložené CD obsahuje stav github repozitáru `github.com/MaxRevicky/Classification-of-eclipsing-binary-stars` v čase odovzdania:

- zdrojové kódy použité pri predspracovaní datasetov: `data-preparation`,
- datasety s observačnými dátami: `datasets`,
- zdrojové kódy použité pri vytvorení a vyhodnotení modelov: `modeling`,
- uložené natrénované modely: `models-checkpoints`,
- základné inštrukcie: `README.md`,
- systémovú príručku v elektronickej forme: `SP_Revicky_Maximilian.pdf`,
- používateľskú príručku v elektronickej forme: `PP_Revicky_Maximilian.pdf`,
- bakalárska práca v elektronickej forme: `BP_Revicky_Maximilian.pdf`.

3 Inštalácia programu

3.1 Požiadavky na programové prostriedky

- Python 3.8.6
- Anaconda Navigator
- Jupyter Notebook
- Knižnice: `numpy`, `pandas`, `json`, `pickle`, `sklearn`, `keras`, `tensorflow`, `imblearn`, `collections`, `scikitplot`

4 Použitie programu

V tejto kapitole si predstavíme použitie skriptov uvedených v prvej kapitole používateľskej príručky.

4.1 Načítanie syntetických dát a uloženie do formátu .pkl

Načítaním dát z databázy a uložením do formátu .pkl sa zaoberajú nasledujúce skripty v priečinku `data-preparation/`:

- `binaries_detached_bellow_i_crit_10000.ipynb`,
- `binaries_detached_random.ipynb`,
- `binaries_overcontact_bellow_i_crit_10000.ipynb`,
- `binaries_overcontact_random.ipynb`,
- `single_spotty.ipynb`.

Pre načítanie dát je potrebné zadať cestu k `.db` suborom. Načítame tabuľky z databázy. Pre účely tejto práce sú potrebné iba dve, `Parameters` a `Curves`. K jednotlivým krivkám z tabuľky `Curves` priradíme parameter `overcontact` tak, aby vznikla iba jedna tabuľka. Hlavným cieľom skriptu je vytvorenie súboru s danými syntetickými krivkami vo formáte `.pkl`.

4.2 Načítanie observačných dát dvojhviezd z JSON súborov a uloženie do formátu .csv

Načítaním observačných dát dvojhviezd z JSON súborov a uložením do formátu `.csv` sa zaoberá Jupyter Notebook `data-preparation/data_observed_binary.ipynb`. Cestu k JSON súborom je potrebné nastaviť do premennej `path_to_json`. Hlavným cieľom skriptu je vytvorenie súboru s danými observačnými krivkami vo formáte `.csv`.

4.3 Načítanie observačných dát dvojhviezd s kritickým sklonom dráhy a škvrnitých hviezd z textových súborov a uloženie do formátu .csv

Načítaním observačných dát dvojhviezd s kritickým sklonom dráhy a škvrnitých hviezd z textových súborov a uložením do formátu .csv sa zaoberá Jupyter Notebook `data-preparation/data_observed_binary_spotty.py`. Cestu k textovým súborom je potrebné nastaviť do premennej `source_folder`. Hlavným cieľom skriptu je vytvorenie súboru s danými observačnými krivkami vo formáte .csv.

4.4 Binárna klasifikácia dvojhviezd na oddelené a dotykové systémy

Úlohu binárnej klasifikácie dvojhviezd na oddelené a dotykové systémy sme rozdelili na dva skripty v priečinku `modeling/`:

- `model_3_experiments_binary_stars_50000.ipynb`
- `model_observed_binary.ipynb`

Skript `model_3_experiments_binary_stars_50000.ipynb` pracuje so syntetickými dátami svetelných kriviek dotykových a oddelených binárnych systémov, ktoré sú vo formáte .pk1. Cestu k datasetu oddelených kriviek je potrebné nastaviť do premennej `data_overcontact_0` a cestu k datasetu dotykových kriviek do premennej `data_overcontact_1`. Hlavným cieľom tejto časti je natrénovať 3 klasifikačné modely a uložiť ich do formátu .hdf.

Skript `model_observed_binary.ipynb` pracuje s observačnými dátami svetelných kriviek dotykových a oddelených binárnych systémov, ktoré sú vo formáte .csv. Cestu k datasetu oddelených a dotykových kriviek je potrebné nastaviť do premennej `data`. Hlavným cieľom tejto časti je vyhodnotenie modelu. Vyhodnotenie sa skladá z kontingenčnej tabuľky klasifikácie a metrík: presnosť, návratnosť a F1 skóre.

4.5 Binárna klasifikácia na dvojhviezdy s kritickým sklonom dráhy a škvrnité hviezdy

Úlohu binárnej klasifikácie na dvojhviezdy s kritickým sklonom dráhy a škvrnité hviezdy sme rozdelili na dva skripty v priečinku `modeling/`:

- `model_experiment_single_spotty_critical_binary.ipynb`
- `model_observed_spotty_binary.ipynb`

Skript `model_experiment_single_spotty_critical_binary.ipynb` pracuje so syntetickými dátami svetlených kriviek škvrnitých hviezd a dvojhviezd s kritickým sklonom dráhy, ktoré sú vo formáte `.pkl`. Cestu k datasetu oddelených kriviek s kritickým sklonom dráhy je potrebné nastaviť do premennej `data_detached_1`, cestu k datasetu dotykových kriviek s kritickým sklonom dráhy do premennej `data_overcontact_1` a cestu k datasetu škvrnitých hviezd do premennej `data_single_spotty_0`. Cieľom tejto časti je natrénovať klasifikačný model a uložiť ho do formátu `.hdf`.

Skript `model_observed_ell_rot.ipynb` pracuje s observačnými dátami svetlených kriviek škvrnitých hviezd a dvojhviezd s kritickým sklonom dráhy, ktoré sú vo formáte `.csv`. cestu k datasetu škvrnitých hviezd a dvojhviezd s kritickým sklonom dráhy je potrebné nastaviť do premennej `data`. Cieľom je vyhodnotenie modelu pomocou kontingenčnej tabuľky klasifikácie a metrík: presnosť, návratnosť a F1 skóre.

5 Popis vstupných, výstupných a pracovných súborov

Vstupné dáta potrebné pre načítanie syntetických dát z databázy a uloženie do formátu `.pkl`:

- dáta syntetických kriviek dotykových systémov vo formáte `.db`
- dáta syntetických kriviek oddelených systémov vo formáte `.db`
- dáta syntetických kriviek dotykových systémov s kritickým sklonom dráhy vo formáte `.db`
- dáta syntetických kriviek oddelených systémov s kritickým sklonom dráhy vo formáte `.db`
- dáta syntetických kriviek škvrnitých hviezd vo formáte `.db`

Výstupné dáta z načítania syntetických dát z databázy a uloženie do formátu `.pkl`:

- dataset syntetických kriviek dotykových systémov vo formáte `.pkl`
- dataset syntetických kriviek oddelených systémov vo formáte `.pkl`
- dataset syntetických kriviek dotykových systémov s kritickým sklonom dráhy vo formáte `.pkl`
- dataset syntetických kriviek oddelených systémov s kritickým sklonom dráhy vo formáte `.pkl`
- dataset syntetických kriviek škvrnitých hviezd vo formáte `.pkl`

Jupyter notebook `data-preparation/data_observed_binary.ipynb` pracuje s dátami:

- dáta observačných kriviek dvojhviezd vo formáte `.json`

Výstupom Jupyter notebooku `data-preparation/data_observed_binary.ipynb` je súbor:

- dataset observačných kriviek dvojhviezd vo formáte `.csv`

Skript `data-preparation/data_observed_binary_spotty.py` pracuje s dátami:

- dáta observačných kriviek dvojhviezd a škvrnitých hviezd vo formáte `.txt`

Hlavným výstupom skriptu `data-preparation/data_observed_binary_spotty.py` je súbor:

- dataset observačných kriviek dvojhviezd a škvrnitých hviezd vo formáte `.csv`

Jupyter notebook `modeling/model_3_experiments_binary_stars_50000.ipynb` pracuje s dátami:

- dataset syntetických kriviek dotykových systémov vo formáte `.pkl`
- dataset syntetických kriviek oddelených systémov vo formáte `.pkl`

Hlavným výstupom Jupyter notebooku `modeling/model_3_experiments_binary_stars_50000.ipynb` sú nasledovné súbory:

- natrénovaný model klasifikácie `models-checkpoints/model_experiment_1_50000.hdf5` z prvého experimentu
- natrénovaný model klasifikácie `models-checkpoints/model_experiment_2_50000.hdf5` z druhého experimentu
- natrénovaný model klasifikácie `models-checkpoints/model_experiment_3_50000.hdf5` z tretieho experimentu

Jupyter notebook `modeling/model_observed_binary.ipynb` pracuje s dátami:

- dataset observačných kriviek dvojhviezd vo formáte `.csv`

Hlavným výstupom Jupyter notebooku `modeling/model_observed_binary.ipynb` sú súbory:

- obrázok ROC krivky `roc_auc.png`
- obrázok vzorky zle predikovaných oddelených systémov `image.png`
- obrázok vzorky zle predikovaných dotykových systémov `image2.png`

Jupyter notebook `modeling/model_experiment_single_spotty_critical_binary.ipynb` pracuje s dátami:

- dataset syntetických kriviek dotykových systémov s kritickým sklonom dráhy vo formáte `.pkl`
- dataset syntetických kriviek oddelených systémov s kritickým sklonom dráhy vo formáte `.pkl`
- dataset syntetických kriviek škvrnitých hviezd vo formáte `.pkl`

Hlavným výstupom Jupyter notebooku `modeling/model_experiment_single_spotty_critical_binary.ipynb` je súbor:

- natrénovaný model klasifikácie `models-checkpoints/model_experiment_4.hdf5` zo štvrtého experimentu

Jupyter notebook `modeling/model_observed_spotty_binary.ipynb` pracuje s dátami:

- dataset observačných kriviek dvojhviezd s kritickým sklonom dráhy a škvrnitých hviezd vo formáte `.csv`

Hlavným výstupom Jupyter notebooku `modeling/model_observed_spotty_binary.ipynb` je súbor:

- obrázok ROC krivky `roc_auc.png`

- obrázok vzorky zle predikovaných škvrnitých hviezd `image.png`
- obrázok vzorky zle predikovaných dvojhviezd s kritickým sklonom dráhy `image2.png`

6 Príklad použitia

V tejto kapitole opíšeme jednotlivé kroky použitia skriptu `modeling/model_experiment_single_spotty_critical_binary.ipynb`.

1. Otvoríme Jupyter notebook s názvom `model_experiment_single_spotty_critical_binary.ipynb`.
2. Nastavíme správnu cestu k potrebným súborom, ktoré sú potrebné na spustenie projektu.
3. Spustíme všetky bunky kliknutím na horné menu -> Run -> Run All cells.
4. Výstupom Jupyter notebooku je natrénovaný model, ktorý dokáže danú krivku klasifikovať na dvojhviezdu s kritickým sklonom dráhy a škvrnitú hviezdu.