Предсказание пульсаров на основе датасета HTRU2

...

Артём Давыдов
Группа 205

Постановка задачи

Одним из объектов исследования современной астрофизики являются пульсары - космические источники электромагнитного излучения, быстро вращающиеся вокруг своей оси. Сигналы от пульсаров приходят на Землю в виде периодических импульсов, которые отслеживаются с помощью телескопов.

Однако такой метод поиска затруднен ложными срабатываниями телескопов на другие сигналы, которые не относятся к пульсарам.

Актуальность данной задачи обусловлена тем, что если удастся разработать хорошую технологию для отслеживания ложных срабатываний, то её можно будет предложить для применения учёным-астрономам.

Зачем нужно машинное обучение?

Поскольку в работе телескопов часто бывают ложные срабатывания, то необходимо их отследить, но как именно - непонятно. Поэтому здесь полезно будет применить машинное обучение для классификации таких сигналов. Это - классическая задача машинного обучения: на основе характеристик полученных сигналов, отследить пульсары.

Набор данных (DataSet)

Набор данных называется HTRU2, он представляет собой 17898 различных сигналов, которые были зарегистрированы телескопами, из них:

1. 16259 ложных срабатываний
2. 1639 истинных срабатываний

Каждому пульсару соответствует строка из 8 чисел, каждое из которых - это интегральные характеристики двух кривых - интегрированного профиля и DM-SNR кривой.

Методы решения поставленной задачи

Для решения задачи будет использоваться обычная полносвязная нейронная сеть (Dense-слои), количество нейронов в которых подбирается (примерно) так, чтобы не было переобучения(не слишком много). На выходе должен быть 1 нейрон, т.к. рассматриваемая задача - задача бинарной классификации.

Возможные проблемы

При решении задачи с помощью нейронных сетей могут возникнуть следующие проблемы:

1. Данные сильно несбалансированны: “не пульсаров” гораздо больше, чем пульсаров. Поэтому нейронная сеть может научиться говорить, что всегда “не пульсар”
2. Слишком мало самих данных: всего 17898 пульсаров, а на каждый пульсар приходится только 8 характеристик.

Методы решения озвученных проблем

Чтобы решить указанные выше проблемы были использованы следующие методы:

1. Для несбалансированных данных, согласно статье [1], используется метод “пересчета веса класса”, при этом вес класса “пульсар” становится больше веса класса “не пульсар”. И сети становится невыгодно выдавать на выходе все ответы “не пульсар”
2. Чтобы увеличить количество данных, согласно статье [2], используется метод “Feature engineering”, который на основе имеющихся данных, с помощью стандартных математических операций, составляет новые данные. Это также помогает обнаружить скрытые закономерности в данных.

Обучение и проверка работы сети

Исходные данные были разбиты на 3 группы:

1. 80% - на обучение сети
2. 10% - на проверку
3. 10% - на тест

Работа была проведена с 3-мя типами данных:

1. Обычные данные из HTRU2 без применения методов “Feature engineering”
2. Центированные данные из HTRU2 без применения методов “Feature engineering”
3. Данные с применением метода “Feature engineering”

Достигнутые результаты

Обучение и проверка качества работы сети показали:

1. Обучение на центрированных данных не дает существенных улучшений: сеть по-прежнему из 27 пульсаров угадывает 20-23, а количество “не пульсаров”, которые она классифицирует как пульсары остается равным 30-40.
2. Применение метода “Feature engineering” ухудшает работу сети: пульсары вообще не предсказываются: ответ на выходе: “всегда не пульсар”.

Дополнительных ограничений на результаты нет: максимум можно угадать 27 пульсаров, а остальные 1763 ответа должны быть “не пульсар”

Список использованной литературы

1. [1] - www.tensorflow.org/tutorials/structured_data/imbalanced_data
2. [2] - medium.com/dataexplorations/tool-review-can-featuretools-simplify-the-process-of-feature-engineering-5d165100b0c3