ID: 455

A Database Dedicated to the Development of Machine Learning Based Disruption Predictors

Qiqi Wu, Wei Zheng, Ming Zhang, Yuxing Wang.

wuqiqi@hust.edu.cn

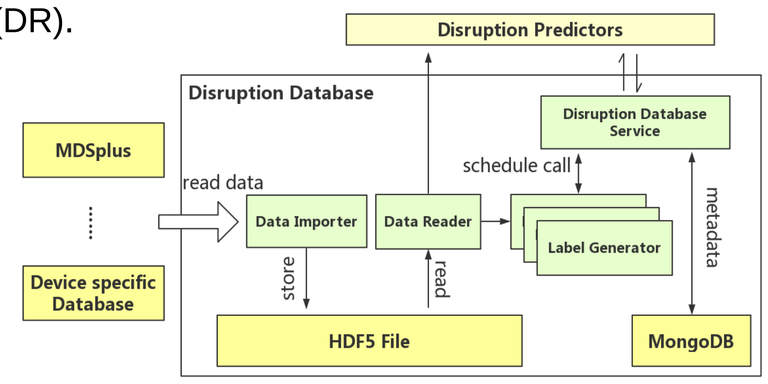
International Joint Research Laboratory of Magnetic Confinement Fusion and Plasma Physics, Huazhong University of Science and Technology

Introduction

- Data plays an important role in machine learning, so data is very important for machine learning based disruption prediction algorithms.
- There are a lot databases build with disruption related information in it in many Tokamak devices, but they are not designed for machine learning based disruption predictor's development.
- In order to develop the disruption prediction algorithm conveniently, we created the disruption database(DDB).
- It allows developers to design disruption prediction algorithms without the consideration of complex processing of data, it provides a data searching, data filtering and predictive performance evaluation function.

The Structure of the DDB

- The disruption database is delivered as a python package, it works with a MongoDB and a file system.
- The data is in 2 categories: the diagnostic data and the labels of each shot. All the labels is stored in MongoDB while the diagnostic data is stored in file system in format of HDF5.
- The package has 4 main components: Data Importer(DI), Label Generator(LG), Disruption Database Service(DDBS) and Data Reader(DR).



The structure of the DDB

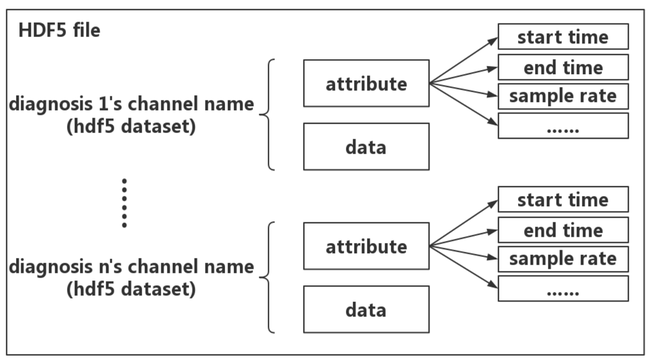
- The work flow:

1. First use Data Importer to import device specific database diagnostic data into unified HDF5 files.
2. Then the Disruption Database Service will call the Label Generator to load the different label generator plugins and generate labels of each shot, the labels are stored in MongoDB.
3. After step 1 and 2, the Disruption Database Service is ready, you can query the shot you need by query API.
4. Then use the Data Reader to read out the diagnostic data to train or do inference.
5. Finally, you can use Disruption Database Service to get a performance evaluation.

The Main Components of DDB

Data Importer

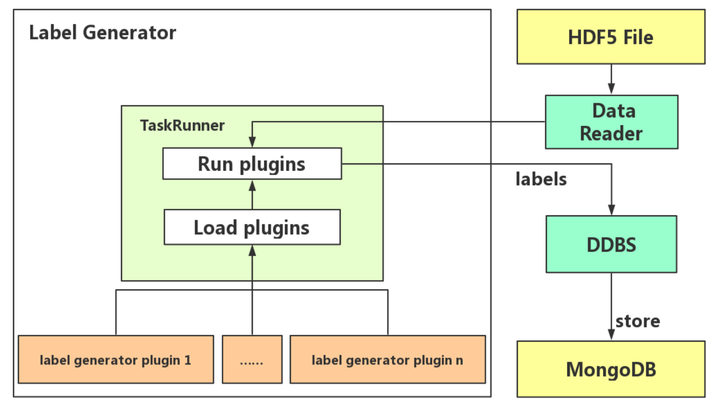
- It's device specific meaning it is defined uniformly and implemented by the people on different tokamak because the diagnostic data may be stored in different format or database on different tokamak.
- We provide API to write the data into unified format HDF5 file(including unified folder structure, each shot stored in one file).
- You just read data from tokamak's database and then use our API for storing data.



The data structure of hdf5 file

Label Generator

- A program that load different label generator plugins and generate a set of labels with a value.
- TaskRunner can distribute the generator plugins and the diagnostic data into different worker threads, it may even run on different machines at the same time.



The structure of Label Generator

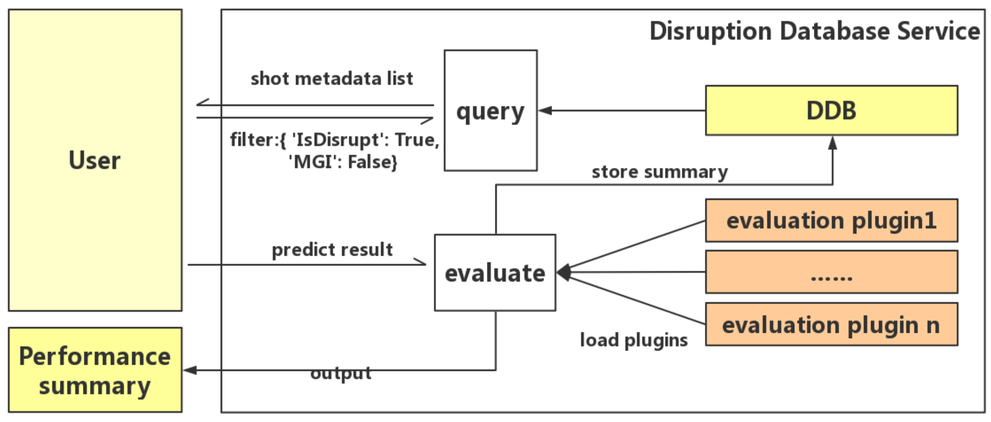
- The base class of plugins has been implemented and can be inherited, so you just need to focus on the label generating algorithms.
- We have already developed a few label generator plugins and worked out dozens of labels of the data from J-TEXT.

Part of the labels for J-TEXT data(others are not listed)

Labels	Data type	Physical meaning
IsDisrupt	Bool	Whether a shot is disruptive or not
RampDownTime	Float	Plasma current's drop time of non-disruptive shot
CqTime	Float	Current quenching time
CqDuration	Float	Duration of the current quenching
IsLockedMode	Bool	Whether a shot is locked mode or not
LockedModeTime	Float	Time of the locked mode happen

Disruption Database Service

- DDBS provides query function. You set a query criteria, and the function returns a list of shot satisfying these criteria and metadata of those shots.
- DDBS will provide performance evaluation function. You submit the prediction result of your predictor to DDBS, it will give you a folder of evaluation results containing a summary in format of json, and a set of figure ready for publication.



The structure of Disruption Database Service

Data Reader

- It provide API to read from the unified HDF5 files.
- The key parameters of read data function is: shot number, a list of diagnoses' name, the path of HDF5 files' root directory.
- Return value is a python dictionary whose keys are diagnoses' name and value are diagnostic data.

Future work and obstacles

- In the future, we will open sources of the disruption database on GitHub.
- We will host a MongoDB server for various tokamak so the researchers can use this tool to contribute to the disruption database and use the data. This will also promote the cross machine disruption predictors' development.
- A web UI for human analyzing and labeling the shots will be added into the disruption database.
- One of the obstacles is the diagnostic data is too big, a massive storage is necessary.
- Current tokamak experiment is poorly logged at least on J-TEXT. Many useful data is missing and diagnostic changes or malfunctions are not logged, making labeling the shot extremely difficult.

Contact Us: wuqiqi@hust.edu.cn, zhengwei@hust.edu.cn

Address: School of Electrical and Electronic Engineering, Luoyu Rd. #1032, Huazhong University of Science and Technology, Wuhan 430074, China