Лаборатория Касперского

13 апреля 2017

Хакатон по анализу данных. Очный этап.

README

Привет!

Поздравляем Вас с выходом в очный этап хакатона от “Лаборатории Касперского”!

В этом файле Вы найдёте всю необходимую информацию о специфике задачи, формате входных и выходных данных. Также, в архиве есть файл sample_submission.csv, который поможет Вам понять, в каком формате мы хотим получить ответ.

Описание данных

В архивах train.zip и test.zip лежат по 500 файлов с именами формата:

“<номер>_<train/test>.csv”

В архивах zips_train.zip и zips_test.zip лежат по 500 запакованных файлов с именами формата:

“<номер>_<train/test>.csv.zip”

Один ряд в сжатом виде занимает 10 Мб, а в распакованном 40 Мб.
В распакованном виде train.zip и test.zip суммарно занимают примерно 40 Гб.
В распакованном виде zips_train.zip и zips_test.zip суммарно занимают примерно 10 Гб.

Для экономии места на жестком диске рекомендуем работать с zips_train.zip и zips_test.zip и воспользоваться скриптом из архива (“load_data.py”) для загрузки с одновременной распаковкой в python.

Каждый из файлов в папках test и train содержит многомерный временной ряд, моделирующий 96 часов автономной работы некоторого химического завода. Скорость записи показаний - 1000 показаний в час. Таким образом, всего в каждом временном ряде 96000 показаний.

Структура временного ряда

Каждый многомерный временной ряд состоит из 56 переменных (см. описание завода в конце документа):

0 – time
1 – 41 – meas – measurements
42 – 53 – mv – manipulated values
54 – product rate
55 – hourly cost

Постановка задачи

Бинарная классификация рядов.

Во время некоторых симуляций, записанных в папки train и test, на завод были совершены кибер-атаки. При этом для рядов из папки train факт наличия или отсутствия атак известен и записан в файл train_labels.csv, который имеет структуру: “имя файла” - была атака или нет. Например:

SeriesId,Attack
0_train.csv,0.0
1_train.csv,1.0
2_train.csv,0.0

Ваша задача - для симуляций из папки test определить в каких из них завод подвергался атаке за время наблюдений, а в каких не подвергался и работал в штатном режиме. Отметим, что штатный режим работы завода может содержать корректные переходные процессы, немного напоминающие кибер-атаки.

Файл ответа должен иметь следующую структуру (см. пример “sample_submission.csv”):

SeriesId,Attack
0_test.csv,1.0
1_test.csv,0.0
2_test.csv,1.0

Метрика: roc_auc_score

Подробнее:
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

Подсчет результатов

Результаты разбиты на private и public части в соотношении 70/30. В течение соревнования Вы в реальном времени будете видеть свой результат на public части. Финальный результат рассчитывается только на private части.

Правила проведения хакатона

1. Файл решения должен уметь получать на вход csv-файл для 96 часов работы завода и выдавать бинарный ответ - была ли атака или нет.
2. Работы лидеров будут проверены экспертной комиссией, которая имеет право дисквалифицировать команду в случае нарушения правил.
3. Победителем становится команда, чье решение прошло экспертную проверку и показало лучший результат на private части.

Не будут засчитаны решения:

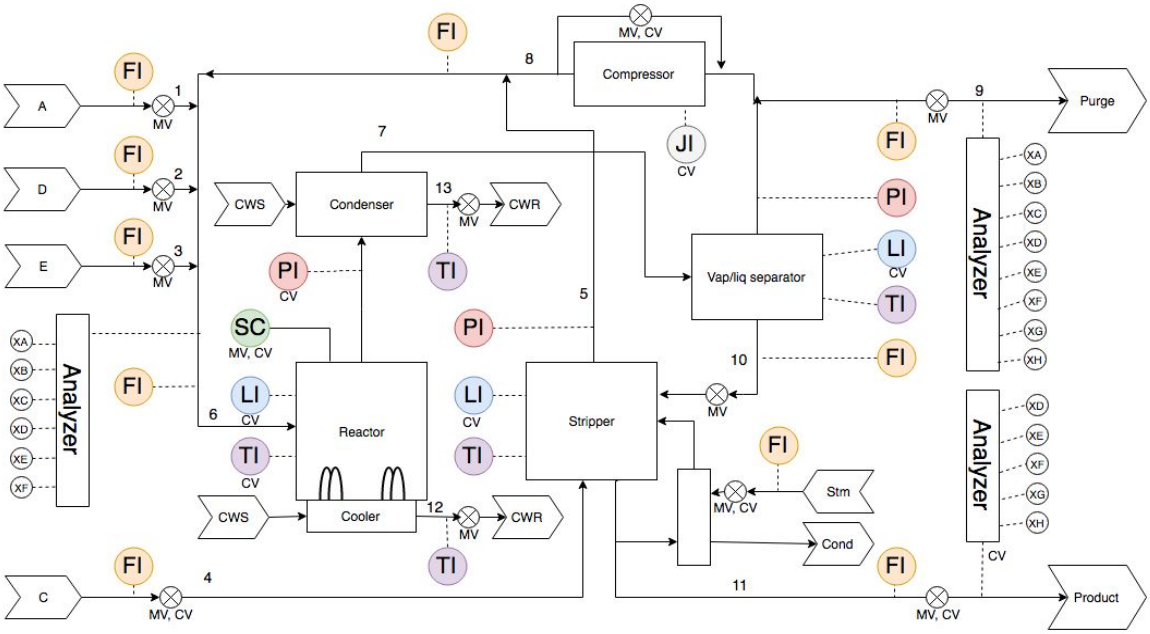
1. **Разметка руками.**
2. **Подбор параметров под скрипт оценки.**
3. **Нахождение решения методом посылки ответов каждые 5 минут.**

Дополнительная информация о заводе

В папках train и test лежат данные, полученные с модели химического завода «Tennessee Eastman Process». Вследствие реакции неизвестных химических компонент A, C, D, E получаются новые соединения (“g”-gas, “liq”-liquid) G и F:

$$A(g) + C(g) + D(g) \rightarrow G(liq), \text{Product 1}$$
$$A(g) + C(g) + E(g) \rightarrow H(liq), \text{Product 2}$$
$$A(g) + G(g) \rightarrow F(liq), \text{Byproduct}$$
$$3D(g) \rightarrow 2F(liq), \text{Byproduct}$$

Схема промышленного процесса имеет вид:



Indicators:
FI — Flow, TI — Temperature, PI — Pressure, LI — Level, JI — Compressor Power,
SC — Speed Controller, CWS — Cold Water Source, CWR — Cold Water Return
MV — Manipulated variables, CV — Controlled Variables

MEAS – Measurements:
1. A Feed, kscmh
2. D Feed, kg/h
3. E Feed, kg/h
4. A + C Feed, kscmh
5. Recycle flow, kscmh
6. Reactor feed, kscmh
7. Reactor pressure, kPa
8. Reactor level, %
9. Reactor temperature, 'C
10. Purge rate. kscmh
11. Separator temperature, 'C
12. Separator level, %
13. Separator pressure, kPa
14. Separator underflow, m'/h
15. Stripper level, %
16. Stripper pressure, kPa
17. Stripper underfow, m]/h
18. Stripper temperature, 'C
19. Steam flow, kg/h
20. Compressor work, kW
21. Reactor coolant temperature, 'C
22. Condenser coolant temperature, 'C
23. Feed %A, mol%
24. %B, mol%
25. %C, mol%
26. %D, mol%
27. %E, mol%
28. %F, mol%
29. Purge %A, mol%
30. %B, mol%
31. %C, mol%
32. %D, mol%
33. %E, mol%
34. %F, mol%
35. %G, mol%
36. %H, mol%
37. Product %D, mol%

38. %E, mol%
39. %F, mollYo
40. %G, mol%
41. %H, mol%

MV - Manipulated variables:
42. D feed flow, %
43. E feed flow, %
44. A feed flow, %
45. C feed flow, %
46. Compressor recycle valve, %
47. Purge flow, %
48. Separator liquid flow, %
49. Stripper liquid product flow, %
50. Stripper steam flow, %
51. Reactor cooling water flow, %
52. Condenser cooling water flow, %
53. Agitator speed, %