Университет ИТМО

Машинное обучение
Лабораторная работа №2

Студент: Маскайкин А.В.
Группа: P4117

Санкт-Петербург
2017г

1. Постановка задачи

Провести серию экспериментов с построением и тестированием деревьев решений (используя DecisionTreeClassifier и RandomForestClassifier), переразбивая исходное множество данных, заданное в варианте, следующим образом:
Номер эксперимента Размер обучающей выборки Размер тестовой выборки

Номер эксперимента	Размер обучающей выборки	Размер тестовой выборки
1	60 %	40 %
2	70 %	30 %
3	80 %	20 %
4	90 %	10 %

2. Исходные данные

- Датасет: https://archive.ics.uci.edu/ml/datasets/Website+Phishing
- Предметная область: Фишинговые сайты
- Задача: определить, фишинговый, подозрительный или нормальный сайт
- Количество записей: 1353
- Количество атрибутов: 9
- Атрибуты:
  1. SFH {1,-1,0}
  2. Pop-up Window {1,-1,0}
  3. SSL final state {1,-1,0}
  4. Request URL {1,-1,0}
  5. URL of Anchor {1,-1,0}
  6. Web traffic {1,-1,0}
  7. URL Length {1,-1,0}
  8. Age of domain {1,-1}
  9. Having IP Address {1,-1}

Во всех характеристиках значение «-1» означает «фишнговый», «0» - подозрительный, «1» - нормальный.

2.1 Описание параметров

- SFH (Server from handler) — Представление пользовательской информации, которая передается из веб страницы на сервер. Если оно пустое — сайт фишинговый, если передача идет на другой домен — подозрительный.
- Pop-up Window — Наличие всплывающего окна. Если при окне не доступен правый клик, то сайт фишинговый.
- SSL final state — Подлинность SSL сертификата.
- Request URL — Количество запросов к веб странице. Если их много, то, вероятно, сайт подвергся атаке, которая заменяет содержимое (текст/картинки). Если количество запросов велико — сайт фишинговый.

- URL of Anchor — привязка к URL. Если при вводе адреса сайта в браузере происходит редирект на другой домен, то привязки нет. И если процент редиректов большой — сайт фишинговый.
- Web traffic — объем веб трафиика сайта. У нормальных сайтов объем высокий, у фишинговых — низкий.
- URL Length — Длина адреса сайта. Чем больше длина, тем выше вероятность, что в адрес встроен вредоносный код.
- Age of domain — Возраст сайта. Если сайт существует менее полугода, то его можно заподозрить как фишинговый.
- Having IP Address — Наличие IP адреса. Если адреса нет — сайт фишинговый.

3. Реализация алгоритма.

```
# coding=utf-8
from __future__ import division
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
# загрузка датасета
from sklearn.tree import DecisionTreeClassifier

def load_data(filename):
    return pd.read_csv(filename, header=None).values

# разделение датасета на тестовую и обучающую выборку
def split_dataset(test_size):
    dataset = load_data('fs.dataset.csv')
    site_attr = dataset[:, :-1]  # список атрибутов для каждого сайта
    site_class = dataset[:, -1]  # класс (результат) сайта (норм,
подозрительный, фишинговый)
    site_class = site_class.astype(np.int64, copy=False)
    data_train, data_test, class_train, class_test = \
        train_test_split(site_attr, site_class, test_size=test_size,
random_state=55)
    return data_train, class_train, data_test, class_test

def main():
    max_size = 0.4
    min_size = step = 0.1
    for size in np.arange(min_size, max_size, step):
        data_train, class_train, data_test, class_test = split_dataset(size)
        decision_forest = DecisionTreeClassifier()
        decision_forest = decision_forest.fit(data_train, class_train)
        decision_accuracy = decision_forest.score(data_test, class_test)
        random_forest = RandomForestClassifier()
        random_forest = random_forest.fit(data_train, class_train)
        random_accuracy = random_forest.score(data_test, class_test)
        print("Size: ", round(size, 1))
        print('DecisionTree accuracy: ', round(decision_accuracy, 10))
        print('RandomTree accuracy: ', round(random_accuracy, 10))
if __name__ == '__main__':
    main()
```

4. Результаты работы.

('Size: ', 0.1)

('DecisionTree accuracy: ', 0.8823529412)

('RandomTree accuracy: ', 0.8602941176)

('Size: ', 0.2)

('DecisionTree accuracy: ', 0.8450184502)

('RandomTree accuracy: ', 0.8560885609)

('Size: ', 0.3)

('DecisionTree accuracy: ', 0.8596059113)

('RandomTree accuracy: ', 0.8472906404)

('Size: ', 0.4)

('DecisionTree accuracy: ', 0.8708487085)

('RandomTree accuracy: ', 0.8542435424)

По результатм серии экспериментов оба алгоритма (DecisionTree и RandomTree) показали схожий результат на данном датасете при варьированных размерах обучающих и тестовых выборок. Точность обоих алгоритмов в данной серии никак не зависела от размера выборки: в каких-то случаях при увеличении тестовой выборки точность увеличивалась, в других – наоборот, уменьшалась. Но в целом оба алгоритма показали довольно результат с высокой точностью – в среднем 0.85-0.86. Этот показатель для данного датасета выше, чем у алгоритмов Naive Bayes и K Nearest Neighbours (0.78 – 0.8), тестируемых в прошлой работе. Данный факт говорит о том, что использование деревьев решений является хорошим инструментом для определения фишинговых сайтов.