DLI Accelerated Data Science Teaching Kit

Module 24 Team Project:
Fake News Detection (cuML)

OBJECTIVE

Source: https://venturebeat.com/2020/06/01/ai-tools-could-improve-fake-news-detection-by-analyzing-users-interactions-and-comments/

Social media (e.g., Twitter and Facebook) has become a new ecosystem for spreading news [1]. Nowadays, people are relying more on social media services rather than traditional media because of its advantages such as social awareness, global connectivity, and real-time sharing of digital information. Unfortunately, social media is full of fake news. Fake news consists of information that is intentionally and verifiably false to mislead readers, which is motivated by chasing personal or organizational profits [2]. For example, fake news has been propagated on Twitter like infectious virus during the 2016 election cycle in the United States [3], [4]. Understanding what can be done to discourage fake news is of great importance.

Fake news detection [5], [6], [7] is to determine the truthfulness of the news by analyzing the news contents and related information such as propagation patterns. It attracts a lot of attention to resolve this problem from different aspects, where supervised learning based fake news detection dominates this domain. The goal of this team project is to build classifiers via support vector machine (SVM), Random Forest, and Logistic Regression, and speed up the process with RAPIDS. It will require 2 ~ 3 student to complete this team project.

PREREQUISITES

Install the necessary Python packages below:

- **cuML (https://github.com/rapidsai/cuml)** is a suite of libraries that implement machine learning algorithms and mathematical primitives functions that share compatible APIs with other RAPIDS projects.

- **sklearn** contains a number of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

INSTRUCTIONS

- Figure out support vector machine (SVM) with following references:
    1. Suthaharan, S., 2016. Support vector machine. In *Machine learning models and algorithms for big data classification* (pp. 207-235). Springer, Boston, MA.
    2. Pisner, D.A. and Schnyer, D.M., 2020. Support vector machine. In *Machine Learning* (pp. 101-121). Academic Press
    3. Noble, W.S., 2006. What is a support vector machine?. Nature biotechnology, 24(12), pp.1565-1567.
- Download the training data from here (https://www.kaggle.com/c/fake-news/data?select=train.csv) as the dataset (fake_news.csv)
- Config Google Colab environment to install all required packages
- Data Preprocessing
    1. Remove rows with missing values
    2. Stemming (https://en.wikipedia.org/wiki/Stemming)
    3. Remove stop words (https://en.wikipedia.org/wiki/Stop_word)
    4. Extract features with **Term Frequency — Inverse Document Frequency (TFIDF)** (https://en.wikipedia.org/wiki/Tf–idf) to build samples
    5. Split the samples into training and testing datasets with the ratio 0.2
- Building fake news detection models
    1. Load training data and testing data
    2. Train a classifier 1 with traditional random forest from sklearn on the training data and record the training time
    3. Train a classifier 2 with traditional Logistic Regression from sklearn on the training data and record the training time
    4. Train a classifier 3 with traditional SVM from sklearn on the training data and record the training time
    5. Train a classifier 4 with random forest from cuML on the training data and record the training time
    6. Train a classifier 5 with Logistic Regression from cuML on the training data and record the training time
    7. Train a classifier 6 with SVM from cuML on the training data and record the training time
    8. Complete the fake news detection on testing data with these six models and record the classification performance such as accuracy
- Compare the training time and the classification performance

References:

[1] G. Pennycook and D. G. Rand, “Fighting misinformation on social media using crowdsourced judgments of news source quality,” Proceedings of the National Academy of Sciences, p. 201806781, 2019.
[2] K.Shu,A.Sliva,S.Wang,J.Tang,andH.Liu,“Fakenewsdetectionon social media: A data mining perspective,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.
[3] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” Journal of economic perspectives, vol. 31, no. 2, pp. 211–36, 2017.
[4] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on twitter during the 2016 us presidential election,” Science, vol. 363, no. 6425, pp. 374–378, 2019.
[5] D. Hovy, “The enemy in your own camp: How well can we detect statistically-generated fake reviews–an adversarial study,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, 2016, pp. 351–356.
[6] W. Y. Wang, “” liar, liar pants on fire”: A new benchmark dataset for fake news detection,” arXiv preprint arXiv:1705.00648, 2017.
[7] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, “A stylometric inquiry into hyperpartisan and fake news,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 231–240.