Обнаружение ботов в социальных сетях на примере ВКонтакте

Орлова А. С.

Научный руководитель: Александр Русланович Швец

Санкт-Петербургская школа физико-математических и компьютерных наук

ВШЭ, Санкт-Петербург

11 Июня 2020

Мотивация

- Боты — пользователи, распространяющие спам, управляемые злоумышленником.
- Взломанные — пользователи, утратившие доступ к аккаунту, управление которым получил злоумышленник.
- Поиском ботов занимаются модераторы. Модератор - сотрудник, чья задача заключается в поиске ботов и взломанных пользователей на основе жалоб от пользователей или анализе подозрительного поведения пользователей.
- Возникают следующие проблемы:
  - При увеличение количества пользователей необходимо нанимать большее число модераторов.
  - Модераторы могут совершать ошибки, их сложно обнаружить.

Обзор работ

1. Метод построения и анализа структуры социального графа

Метод заключается в построении соцального графа и его дальшейшем анализе. Для анализа графа используются алгоритмы поиска компонент сильной связности, случайные блуждания[1].
Недостатки подхода:

- Долгое время работы
- Низкая точность метода

Из-за имеющихся недостатков данный подход не используется для поиска ботов в современной разработке. Однако, метод используется для генерации признаков для алгоритмов машинного обучения. [2]

[1] SybilInfer: Detecting Sybil Nodes using Social Networks [G. Danezis and P. Mittal, 2009]

[2] TweetScore: Scoring Tweets via Social Attribute Relationships for Twitter Spammer Detection [Yihe Zhang, Hao Zhang, Xu Yuan, and Nian-Feng Tzeng. 2019]

Обзор работ

2. Использование ручной разметки от доверенных пользователей

Спустя некоторое время боты обнаружили уязвимости предыдущего метода и стали образовывать больше социальных связей с доверенными пользователями. Тогда предыдущий метод был усовершенствован дополнительным уровнем фильтрации - разметкой от доверенных пользователей. [3]
Недостатки подхода:

- Долгое ожидание разметки
- Проблема приватности данных пользователей

3. Современный подход - методы машинного обучения

[3] Social turing tests: Crowdsourcing sybil detection [G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. Metzger, H. Zheng, 2012]

Методы машинного обучения

3.1 Анализ содержимого публикаций пользователя

Для анализа используется текст сообщений, фотографии, видео. Опубликованный контент обрабатывается методами машинного обучения, иногда объединяется с признаками пользователя. Такое решение эффективно работает для социальных сетей с одним типом контента. [4]

3.2 Анализ данных аккаунта пользователя

Используются только метаданные аккаунтов, такие как имя, время создания, количество опубликованного контента. На основе полученных данных проводится анализ методами классификации и кластеризации. Сложность представляет выбор и получение признаков. [5]

---

[4] Detecting spammers on social networks [Xianghan Zhengab, Zhipeng Zengab, Zheyi Chenc, Yuanlong Yuab, Chunming Rong, 2015]

[5] Detection of spam-posting accounts on Twitter [Isa Inuwa-Dutse, Mark Liptrott, Ioannis Korkontzelos, 2018]

Проблемы ML решений

- Работы рассматривают уже собранные датасеты и не работают в режиме онлайн. Из-за этого не оптимизируется количество признаков и сложность их получения
- Решения адаптированы под конкретные социальные сети и используют их особенности, а значит, не все признаки можно переиспользовать
- Используются небольшие или плохо размеченные наборы данных для обучения и тестирования. Например, разметка производится на основе публикуемого контента
- Сложно сравнивать результаты работ между собой - разные социальные сети и наборы признаков не позволяют переиспользовать датасеты для тестирования моделей

Цель и задачи

Цель

Создать модель для классификации пользователей на добросовестных, ботов и взломанных в инфраструктуре социальной сети ВКонтакте

Задачи:

- Создать инструмент для сбора данных пользовательских аккаунтов из социальной сети ВКонтакте
- Собрать датасет для обучения и тестирования моделей
- Создать модель для классификации пользователей на три класса: добросовестные, фальсифицированные и взломанные
- Протестировать модель на собранных датасетах и на поступающем онлайн потоке пользовательских данных

Инструмент сбора данных

Требования к инструменту:
Инструмент должен собирать данные из социальной сети ВКонтакте, обрабатывая 600 млн пользователей меньше чем за 24 часа.

Свойства инструмента:

- Для каждого пользователя собирается 110 признаков, признаки читаются из различных источников и баз данных
- Запуск сбора данных каждый день, обновляем данные для пользователей, которые совершали действия за последнюю неделю

Основные ограничения:
Количество запросов на запись и время получения ответов от баз данных.

Реализация инструмента сбора данных

- Используем хранилище ключ-значение вместо реляционной базы данных для увеличения скорости работы
- Сохранение только ненулевых данных в хранилище, однако записываем ошибку получения данных
- Объединение запросов пользователей с учетом имеющегося шардирования, распределения данных пользователей по серверам
- Использование и реализация сетевого ThreadPool для распределения нагрузки по времени

Результат:
Было достигнуто время работы 8 часов, в день собираются данные для 90 млн активных пользователей

Предобработка данных

В собранном датасете присутствуют бинарные, вещественные и категориальные данные. Бинарные и вещественные не требуют дополнительных преобразований, однако возникает необходимость выбора представления для категориальных данных. Для этого были протестированы следующие методы:

- Сопоставление значению признака числа пользователей с этим значением, выбирая таким образом наиболее распостраненные значения
- Сопоставление значению процентное отношение заблокированных пользователей, чтобы выявить наиболее подозрительные значения признаков
- Ручная группировка значений признака, основанная на анализе активности пользователей

В итоговой реализации был выбран третий способ кодирования признаков как самый эффективный относительно важности кодируемого признака

Увеличение количества размеченных данных

Проблема:

Имеется множество пользователей у которых не выбран класс, но имеется допустимое множество классов.

Для решения этой проблемы используется метод Partial Label Learning, чья задача выбрать один верный класс из множества допустимых; использован алгоритм PALOC [6]. Для использования этого метода был добавлен четвертый класс "сервисный пользователь" - пользователи, чье поведение похоже на ботов, однако они принадлежат людям.

Проведена дополнительная проверка для доверенных пользователей: из датасета были удалены те, чья активность была подозрительной. Для дальнейшей работы были использованы три класса: 120 тыс. ботов, 170 тыс. взломанных, 10 млн. доверенных пользователей, из которых были случайно выбраны 200 тыс. для дальнейшего обучения.

[6] Towards Enabling Binary Decomposition for Partial Label Learning [Xuan Wu, Min-Ling Zhang, 2018]

Разработка модели классификации

	Один против всех	Мультиклассификация
Число классификаторов	Необходимо создать классификатор для каждого класса	Один классификатор
Число определяемых классов	Пользователь может классифицироваться множеством классов, либо не классифицироваться ни одним из них	Пользователь всегда попадает в один класс
Удобство обучения	При обучении классификаторы не зависят друг от друга. Возможность обновлять модель для каждого класса независимо	Вычисление ошибки при обучении происходит по всем классам

Был выбран подход "Один против всех". Подбор параметров для моделей осуществлен с помощью поиска по сетке. Для выбора полезных для модели признаков использовался итерационный алгорим удаления признаков.

Тестирование

Сравнение с другими работами:

	Precision	Recall	F1
SybilSCAR	0.661	0.436	0.436
TweetScore	**0.989**	0.914	0.946
Tweet Random Forest	0.93	0.92	0.93
Эта работа	0.971	**0.946**	**0.958**

Среди найденных модераторами взломанных пользователей для разных дней:

День	Всего решений, сделанных модераторами	Число ошибочных решений модели	Процент ошибки
#1	8200	764	9.3%
#2	2478	197	7.9%
#3	2424	165	6.8%

Результаты

- Реализован инструмент, генерирующий датасеты из данных пользователей ВКонтакте. Инструмент собирает данные среди 600 млн пользоветалей за 8 часов.
- Собран датасет из активных пользователей, чья последняя активность была не более чем неделю назад. Собрано 110 признаков для 90 млн. пользователей
- Созданы и протестированы модели на поступающих от модераторов решениях, проведено сравнение с моделями из других работ
- Полученное решение интегрировано в инфраструктуру ВКонтакте
- По итогам работы планируется выступление на конференции Antispam & Antifraud meetup в Mail.ru Group