Искусственный интеллект в информационной безопасности

Здравствуйте уважаемые слушатели, в данной лекции мы рассмотрим вопросы использования методов машинного обучения для решения традиционных и современных задач обеспечения информационной безопасности.

К традиционным задачам обеспечения информационной безопасности относятся такие вопросы как определение характеристик атак на компьютерные системы, детектирование аномалий в поведении пользователя и определение его портрета.

В рамках лекции рассматриваются вопросы использования особенностей действий пользователя для его идентификации. Например, индивидуальный стиль написания текста или метаданные, относящиеся к действиям пользователя (такие как время его активности, динамика активности и прочие). Помимо оценки влияния пользователя на систему и анализа его действий с точки зрения информационной безопасности, предлагается рассмотреть и проанализировать классические задачи информационной безопасности – это защита от спама, обнаружение аномалий в системе и т. д.

Для решения кейсов и рассмотрения примеров предлагается использовать классические методы машинного обучения, такие как метод k-средних, линейная регрессия, деревья решений и случайный лес.

Современные или «нетрадиционные» задачи информационной безопасности выходят за пределы классической парадигмы, что подразумевает не только использование новых методов, но и иные цели, стоящие перед лицом, ответственным за обеспечение информационной безопасности системы. Одним из ключевых факторов для этого является развитие концепции киберфизических систем, что приводит к уменьшению роли человека в различных областях.

Одной из основных особенностей киберфизических систем является их частичная или полная самоорганизация. Здесь традиционный подход, когда нарушителем является только человек, не достаточен для защиты. Кроме того, одной из ключевых характеристик информационной безопасности становится не только обеспечение основных свойств информации (конфиденциальности, целостности и доступности), но и возможность функционирования системы с минимальными потерями в условиях нарушения этих свойств.

Искусственный интеллект в информационной безопасности

Сегодня информационная безопасность пронизывает все области современной науки и техники, что приводит к росту числа прикладных задач, которые могут быть эффективно решены при помощи методов машинного обучения.

Традиционно, принято выделять пять групп таких задач:

- прогнозирование угроз,
- предотвращение и обнаружение атак,
- реагирование на инциденты и
- мониторинг состояния.

Эти задачи актуальны в совершенно различных областях и на различных уровнях: на сетевом, на уровне рабочих станций, на уровне приложений, на уровне пользователей и так далее. Соответственно и данные, с которыми мы работаем будут различны. Например, это может быть ретроспективный анализ действий пользователя за рабочей станцией или анализ процессов ОС в режиме реального времени.

Машинное обучение может помочь нам обнаружить несанкционированный доступ к информации, определить узел сети, в котором произошла утечка, сформировать некоторые типичные сценарии проведения атак.

В ходе лекции я расскажу вам о некоторых прикладных задачах обеспечения информационной безопасности, которые могут решаться и решаются с помощью методов машинного обучения. Не стоит рассматривать машинное обучение, как инструмент, который способен оказать влияние в области информационной безопасности только на разделы, связанные с человеком. В современном мире всё более актуальным становится вопросы робототехники и взаимодействия сложных самоорганизующихся систем. В рамках решения этой задачи также применяются методы машинного обучения. Информационная безопасность продолжает развиваться, как и ее основополагающие понятия и свойства. Так, получило развитие понятие целостности, которое сегодня включает не только синтаксическую составляющую, но и семантическую или содержательную.

Представьте ситуацию, когда вы получаете недостоверную информацию, которую намеренно кто-то изменил. В таком случае, можно говорить о том, что на вас было оказано деструктивное информационное воздействие, т. е. была нарушена информационная безопасность. Для решения задачи противодействия такого рода воздействию может применяться машинное обучение и искусственный интеллект. Одним из самых важных примеров нарушений содержательной целостности является некорректный анализ

изображений для беспилотных транспортных средств. Например, небольшое изменение положения дорожного знака или нанесение каких-либо дополнительных символов на знак могут привести к тому, что беспилотник не сможет его распознать.

Одной из первых областей применения машинного обучения в информационной безопасности является обеспечение безопасности электронной почты и особенно защита от спама. Методы машинного обучения довольно давно и успешно применяются в разработке спам-фильтров. Одним из первых стали применять наивный байесовский классификатор (который, кстати, и сегодня может работать довольно успешно), сейчас применяют более сложные и современные алгоритмы, например градиентный бустинг и ансамблевые методы.

Стандартный процесс обучения можно описать следующим образом: письмо формализуют по определенным параметрам или выявляют признаки спама. Это могут быть как текстовые признаки (например, определенные фрагменты текста или определенные слова), так и экспертные (например, домены или служебные заголовки). Для обучения необходима размеченная выборка, в которой для каждого конкретного письма задано, является оно спамом или нет. Однако, с формированием таких наборов данных существуют некоторые сложности. Разметку данных производят люди, указывая, является письмо спамом или нет. Но критерии отнесения конкретного письма к спаму у разных людей разные, это приводит к зашумлению данных и возникновению выбросов. Следовательно, до того, как мы перейдем к этапу обучения такие данные должны быть обязательно предварительно очищены и обработаны. Далее тексты необходимо преобразовать в набор признаков, которые мы выбрали ранее. На этапе обучения классификатор “учат” разбивать письма на две категории – «спам» и «не спам». Это обучение с учителем. Чем тщательнее выбираются признаки, по которым производится обучение, тем лучше будет результат и точнее будет работать классификатор.

Критериев оценки качества классификации может быть несколько: например, скорость классификации, количество отфильтрованных спам-писем (или true positive) и количество хороших сообщений, которые были отмечены как спам (false positive или ошибки первого рода). При решении любой задачи важно соблюдать баланс, повышение качества классификации может вести к значительному увеличению времени, а увеличение скорости реакции может вести к снижению качества. При выборе классификатора и его обучении необходимо это осознавать и основываться на тех метриках качества, которые действительно важны.

Теперь рассмотрим классификацию спама на примере двух писем: первое относится к спаму, второе – нет.

Первое письмо:

LASER PRINTER TONER All right already, enough! Today it comes from UUnet, tomorrow it comes from Level3, the next day it'll come from alltel.net. This one keeps coming back like a bad penny. Don't waste your time calling any phone numbers here as it won't do you any good; You'll just get aggravated. For the record: the MIT postmasters have left a message at 1/16/2000 2129 EST on the listed complaint number (1-888-494-8597) requesting that all MIT.EDU addresses be removed from their distribution. We'll see... Well, we just got spammed again by these idiots. Once again notified them to remove all MIT.EDU addresses at 2/27/2000 at 23:10 EST. Since you're wasting time reading this page, perhaps you'd like to hear the erudite individual whose voice instructs you on how to supposedly get your email address removed from their infernal spamming list. Here's the 700K .au file we transcribed on a recent call.

Второе письмо:

Dear Dennis,

Hope you are well.

I'm writing to you, yet again, in your capacity as "Answer Man."

One of our David English House teachers has just e-mailed me to see if I have any more information on "university listening tests" which are to be administered soon.

I have no information about any such tests. Do you? If so, could you please let me know.

Thank you kindly.

Best regards,

Donna

В каждом письме будут оцениваться следующие признаки:

- средняя длина непрерывных последовательностей заглавных букв в письме (average_length_capital);
- процент слова all в письме (percentage_word_all);
- процент символа ! в письме (percentage_sign_exclamation_mark).

Результаты получаются следующие:

Первое письмо:

- average_length_capital: 2.7
- percentage_word_all: 1.3

- percentage_sign_exclamation_mark: 0.1

  Второе письмо:

- average_length_capital: 1
- percentage_word_all: 0
- percentage_sign_exclamation_mark: 0

Можно заметить различия в значениях признаков, что и составит основу для выявления паттернов и обучения классификатора спама. Одна из основных задач в информационной безопасности, которые сегодня может решать искусственный интеллект, это предсказание будущих атак.

Здесь стоит отметить, что в информационной безопасности один из основных способов машинного обучения – обучение на прецедентах – не всегда будет достаточно эффективен. В частности, новые атаки на систему могут значительно отличаться от предыдущих, и обучение на прецедентах не сможет в полной мере самостоятельно обеспечить защиту. Это происходит в силу специфики области, злоумышленники всегда стараются действовать максимально скрытно, не повторяя прошлые шаги. Повториться и создать закономерность — значит быть обнаруженным и пойманным.

В связи с этим, основной задачей искусственного интеллекта можно считать обнаружение характерных признаков атак для новых событий. Сегодня эта функция возложена на аналитиков безопасности, которые анализируют множество событий, выделяют подозрительные признаки и принимают решение относительно опасности того или иного события. Объем анализируемых данных и число подозрительных событий - огромны, а наличие человеческого фактора приводит к тому, что часть атак будет проведена успешна. Аналитик может просто не успеть дойти в процессе анализа до вредоносного события. Считается, что человек в среднем может проанализировать 15 инцидентов в день. В этом случае все остальные инциденты останутся не разобранными. Кроме того, “ручной” анализ одного инцидента может занимать продолжительное время, что не позволит вовремя нивелировать ущерб от реализации атаки.

Применение же искусственного интеллекта позволит сэкономить значительное время, т. к. задачи анализа и классификации инцидентов, выявления их признаков будут переложено на него. Аналитику же будет достаточно рассмотреть полученные агрегированные характеристики.

Еще одна из областей применения искусственного интеллекта в информационной безопасности — это поведенческий анализ пользователей, создание паттернов типичного

поведения, что позволяет выявлять аномалии и отклонения от штатных действий. Здесь применение искусственного интеллекта основано на прецедентах, но не на прошлых случаях атак, а на типичном поведении пользователя.

Представим, что злоумышленник получил доступ в закрытый контур. После этого он начинает исследовать сеть, что требует дополнительных действий (например, установка дополнительного программного обеспечения). Это существенно отличается от обычного поведения типичного пользователя. Следовательно, в ходе выполнения этих действий, злоумышленник скомпрометирует себя перед системой, основанной на искусственном интеллекте.

Применение такого подхода нашло себя во многих областях, например, SAP (это система планирования ресурсов предприятия). Корпоративная система SAP позволяет анализировать логи действий пользователей. При сборе и анализе множества логов поведения обычных пользователей представляется возможным понять общую тенденцию действий сотрудника. Если система в ходе мониторинга действий пользователей обнаруживает отклонения от этого поведения, она сигнализирует об этом аналитику, который принимает конечное решение о вредоносности действий. Стоит отметить, что такие системы должны со временем переобучаться, т. к. поведение пользователей также меняется со временем.

Проблемы применения методов искусственного интеллекта в информационной безопасности

Искусственный интеллект не может быть ответом на все вопросы в области информационной безопасности. Здесь существует несколько основных проблем. Это:

1. Неполная автономность. В системах безопасности, основанных на искусственном интеллекте, так или иначе, задействован человек. Это не является критичным недостатком такого рода систем, но значительно уменьшает их потенциал;
2. Конфиденциальность данных. Анализ больших объемов данных при помощи искусственного интеллекта может привести к ситуации, когда информация не будет предоставляться. Например, уменьшение количества информации, предоставляемой при регистрации, уменьшение активности в сети и т. д.;
3. Отсутствие регулирования. Из-за уникальной и непредсказуемой природы искусственного интеллекта существующие правовые рамки не всегда применимы в данной области. Существующие нормативно-правовые акты не способны в полной

мере охватить весь спектр задач, решаемый искусственным интеллектом в области информационной безопасности;

4. Этические проблемы. Применение искусственного интеллекта становится все шире, при этом недостаточное внимание уделяется вопросам, связанным с морально-этическими аспектами, присущим человеку при принятии решений в области информационной безопасности.

Резюмируя вышесказанное, методы машинного обучения и искусственного интеллекта позволяют решать различные задачи информационной безопасности. Для их решения могут применяться различные методы, начиная от деревьев классификации, заканчивая методами роевого интеллекта.

Дополнительные материалы:

1. https://esputnik.com/blog/primenenie-iskusstvennogo-intellekta-dlya-filtracii-massovyh-email-rassylok
2. http://web.mit.edu/network/spam/examples/
3. http://vu.flare.hiroshima-u.ac.jp/english/writing/intermediate/informal/examples.htm
4. https://securelist.ru/machine-learning-versus-spam/29962/

Выявление аномалий и обучение на прецедентах

В данном фрагменте я кратко расскажу о выявлении аномалий и обучении на прецедентах, о которых я упоминала ранее.

Выявление аномалий является важной задачей интеллектуального анализа данных, рассматриваемой во многих областях исследования и сферах применения. Детектирование аномалий относится к проблеме нахождения паттернов данных, не соответствующих ожидаемому поведению. Такие несоответствующие паттерны в зависимости от прикладной области относят к аномалиям, выбросам, несогласованным наблюдениям, исключениям, аберрациям, сюрпризам, особенностям и загрязнителям. Наиболее часто используемым понятием в контексте обнаружения аномалий является «выброс». В некоторых приложениях аномалии также называют целевыми (или особыми) событиями. Обнаружение аномалий широко используется в таких областях применения как обнаружение вторжений в компьютерной безопасности, обнаружение мошенничества при проведении банковских транзакций, выявление заболеваний раком, поиск отказов в системе безопасности, слежение за вражеской активностью, контроль движения поездов и др.

Ключевым аспектом любого подхода к обнаружению аномалий является природа анализируемых данных, которые в общем смысле являются набором примеров. Каждый пример данных может обладать рядом отличительных признаков. Признаки могут быть различных типов, таких как бинарные, дискретные и непрерывные. Каждый пример может быть описан как одним (одномерные данные), так и множеством (многомерные данные) признаков. В случае многомерности признаки могут быть различных типов.

Природа признаков определяет применимость различных подходов к обнаружению аномалий. В случае непрерывных или дискретных данных, представленных временными рядами, возможно использование статистических методов.

Одним из наиболее распространенных подходов является вывод на основе прецедентов. Это метод принятия решений, в котором используются знания и опыт предыдущих ситуаций (прецеденты).

Прецедент – это описание проблемы или ситуации в совокупности с подробным указанием действий, предпринимаемых в данной ситуации или для решения данной проблемы. Прецедент включает:

1. Описание проблемы.
2. Решение этой проблемы.
3. Обоснование и результат применения решения.

Описание проблемы должно содержать всю информацию, необходимую для достижения цели вывода. Например, если цель врача – диагностировать заболевание, то описательная информация должна содержать симптомы и результаты лабораторных исследований. Если цель врача – выбор лечения, то понадобятся еще динамика состояния больного, аллергия на лекарственные средства и так далее. Все этапы примененного к больному лечения сохраняются в описании решения.

При рассмотрении новой проблемы (текущего случая) в базе данных прецедентов находится похожий в качестве аналога. Для стандартизации и упрощения можно попытаться использовать его решение, возможно, адаптировав к текущему случаю вместо того, чтобы искать решение каждый раз сначала. После того, как текущий случай будет обработан, он вносится в базу прецедентов вместе со своим решением для его возможного последующего использования.

Использование ИИ для предотвращения и расследования инцидентов. Пример на создание портрета пользователя

Одной из основных задач информационной безопасности является предотвращение инцидентов. Либо в случае, если инцидент уже произошел - максимально быстрое реагирование и расследование. Это две независимые задачи, но решены они могут быть при помощи схожих методов, включающих методы искусственного интеллекта.

В сфере информационной безопасности существует такой термин как “профайлинг” — это набор приемов, которые позволяют составить психологический портрет пользователя и понять, как он будет действовать в различных обстоятельствах, как он относится к правилам безопасности, принятым в организации, и является ли он потенциальным “инсайдером”.

Составление психологического портрета сотрудника, как и многие другие задачи информационной безопасности, раньше решались в “ручном режиме”. Но это достаточно долгий процесс, требующий работы узко квалифицированного специалиста, очных интервью с сотрудниками.

Проблема “ручного” профайлинга заключается в отсутствии четкого алгоритма, по которому составляется психологический портрет, и ясного представления, как получить максимальную пользу от типизации и классификации. Кроме того, нет заранее определенного стандарта, с которым можно было бы сверять полученные результаты.

На помощь специалистам по информационной безопасности в последнее время пришли методы искусственного интеллекта, позволяющие выявлять скрытые

закономерности в поведении. Чем больше характеристик пользователя будет проанализировано, тем точнее и достовернее будет его портрет. Для описания портрета пользователя применяются технологии компьютерного зрения, распознавания речи, методы анализа действий пользователя и т. д.

Системы профайлинга могут быть независимыми, либо являться одним из компонент DLP-систем (это системы предотвращения утечек). В этом случае профайлинг учитывает особенности, которые проявляются в письменной речи пользователей. Тексты собираются, а затем анализируются по различным критериям.

Результаты такого анализа и профилирования позволяют, например, выявлять потенциальных нарушителей, определять является ли сотрудник надежным и лояльным к компании, а также узнавать может ли ему быть предоставлен доступ к некоторой конфиденциальной информации.

Подобные методы могут применяться и при расследовании инцидентов информационной безопасности или компьютерных преступлений, например, в случае если некто опубликовал сообщения с конфиденциальной информацией на форуме или в социальной сети, прислал сообщение с угрозами, либо распространяет информацию, порочащую репутацию. Здесь необходимо определить, кто именно совершил данное преступление, т. е. идентифицировать пользователя-отправителя.

На сегодняшний день существуют две основные группы методов идентификации пользователя: по техническим характеристикам его устройства и по характеристикам сообщений, размещаемых пользователем.

Первая группа применяется довольно давно и является хорошо проработанной. Идентификация производится по характеристикам аппаратного и программного окружения устройства, с которого пользователь осуществляет доступ к ресурсу. Однако, здесь есть одно существенное ограничение - производится идентификация самого устройства, с которой осуществляется доступ, а не конкретного пользователя, который находился за компьютером и отправлял сообщения.

Рассмотрим пример онлайн-магазина, в котором имеется возможность оставлять свои комментарии под товаром или услугой. Как вы знаете, сегодня задача формирования положительного или отрицательного мнения часто сводится к накрутке комментариев, в большинстве случаев это делается автоматизированными средствами, так называемыми

“ботами”. Задача обеспечения информационной безопасности может быть в этом случае сведена к их автоматическому обнаружению.

Автор каждого отзыва или даже простой пользователь-посетитель обладает определенным набором характеристик. Это, например, версия операционной системы; часовой пояс; версия и язык интернет-браузера; разрешение экрана; IP и MAC адрес. А также те данные, которые пользователя вводит на сайте: адрес его электронной почты, дата регистрации, статистика поведения на сайте, комментарии под определенными фирмами или авторами (т. е. перекрестное комментирование). Все признаки можно разделить на пять основных групп - физическое нахождение автора, поведение авторов, инструментарий пользователя, активность автора и характеристика отзыва.

Возьмем одного из пользователей: на сайте он зарегистрирован недавно, его IP-адрес позволяет судить, что автор находится вне России, почта зарегистрирована на сервисе, где легко пройти регистрацию. Также он часто пишет об одной и той же фирме и отмечает отзывы “дружественных авторов”. В результате использования логистической регрессии удалось выявить информативные признаки, которые оказывали влияние на итоговое принятие решений относительно этого автора. Общее количество признаков, необходимых для выявления бота сократилось до 10, хотя изначально признаковое пространство включало более 20 признаков. Итоговый перечень значимых признаков выглядит следующим образом - местоположение автора (в России или нет), является ли автор представителем компании, были ли еще отзывы из подсети той же компании, были ли отзывы об этой же компании за последние 27 часов, использовалось ли мобильное приложение, отзыв об одной компании с одного почтового сервиса, количество отзывов об одной компании с одного аккаунта, значение рейтинга автора, количество жалоб, поданных пользователем и количество уникальных проектов, о которых писал автор.

Результаты анализа позволяют сделать выводы, что нахождение автора вне России повышает вероятность того, что это бот, а, если отзыв оставлен через мобильное приложение или у автора высокий рейтинг, вероятность этого снижается. Оставшиеся признаки повышают вероятность того, что отзыв - ложный (написан ботом), если их значения положительны. Таким образом, может быть построена система автоматической идентификации ботов при анализе их активности на сайте, точность прогноза системы составляет порядка 90%. Стоит отметить, что подобная система в первую очередь направлена на фильтрацию ложных отзывов, что приводит к несколько грубому отбору правдивых отзывов, т. е. отзыв может быть настоящим, но будет отнесен к ложным.

Дополнительная настройка и использование дополнительных значимых признаков могут увеличить точность оценки. Стоит отметить, что с точки зрения функционирования подобного рода систем в реальных условиях такое решение позволило значительно автоматизировать процесс модерации.

Существует возможность идентифицировать преступника или пользователя-злоумышленника и по стилю его письменной речи, т. е. по лингвистическим или стилистическим характеристикам электронных сообщений. Это чем-то похоже на биометрическую идентификацию. Каждый человек имеет свой стиль письма, который составляет своеобразный уникальный «отпечаток» - набор характеристик, позволяющих его идентифицировать. В качестве признаков могут быть использованы различные характеристики текстов: символьные, лексические, синтаксические, семантические, контентно-специфические.

Это могут быть простейшие частоты n-грамм слов или символов, частоты слов определенной длины или любимая длина предложений. Или более сложные - преобладающая лексика, любимые или уникальные слова.

Здесь задача обеспечения информационной безопасности родственна задаче обработки естественного языка. Т. е. могут применяться те же признаки, методы предварительной обработки и алгоритмы машинного обучения.

Рассмотрим задачу определения нарушителя - некто разместил на странице в социальной сети информацию, составляющую коммерческую тайну вашей компании. Эта страница была специально создана только для этой цели и по ее анализу злоумышленника определить не удается: электронная почта не указана, имя и фамилия пользователя - подделка, друзей у него нет, пользователь использовал анонимайзеры для доступа к сайту.

В компании работает порядка 100 человек, а предварительный анализ службой безопасности позволил сузить число подозреваемых до 20. Также в нашем распоряжении имеется собственный сервер электронной почты, на котором содержатся все сообщения электронной почты всех сотрудников. Можно произвести анализ этих сообщений, сравнить их с сообщением размещенном на сайте. Т. е. у нас имеется обучающая выборка и сообщение, которое необходимо отнести к одному из пользователей (т. е. классифицировать).

При решении этой задачи на качество идентификации могут оказывать влияние два основных фактора:

- используемые характеристики электронного сообщения,

- метод классификации.

В качестве признаков могут быть использованы, например частоты n-грамм символов, которые достаточно легко вычислить. Однако, в ходе реального решения данной задачи использовалось множество признаков различных типов: бинарные, количественные и пр.

Изначально признаковое пространство было достаточно высокой размерности. Было принято решение производить отбор информативных признаков. Применение таких методов позволяет решить несколько основных задач: во-первых, снизить время обучения и, во-вторых, повысить точность классификации.

Существуют различные методы отбора признаков. При решении нашей задачи использовался алгоритм Relief-F, в котором отбор производится на основе вычисления расстояния по значению признака до k-ближайших соседей. Каждому из признаков назначается определенный коэффициент или вес, рассчитанный путем оценки расстояния до ближайших сообщений того же пользователя и до сообщений других пользователей.

Вес тем выше, чем лучше он отделяет сообщения одного пользователя от сообщений других пользователей, и чем хуже разделяет сообщения пользователя между собой (Формула 1 и Рис. 1).

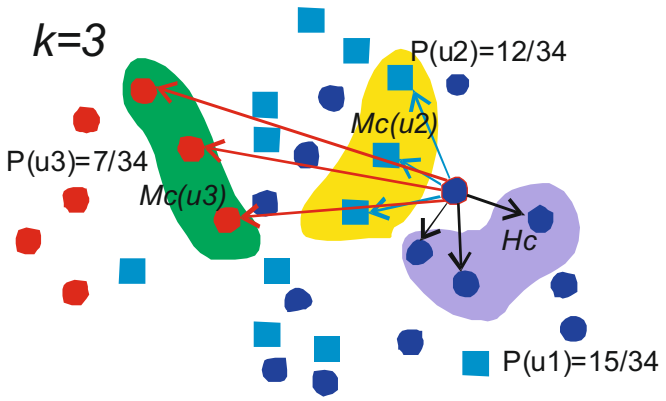


Рисунок 1. Отбор признаков на основе вычисления расстояния по значению признака до k-ближайших соседей

Вес признака уменьшается на значение по расстоянию до сообщений того же пользователя, и увеличивается на значение до сообщений других пользователей.

$$W_{q+1}(f_i) = W_q(f_i) - \sum_{c=1}^{k} \frac{d(f_i,t_j,t_{xc})}{l \times k} + \sum_{z=1}^{y} \left[\frac{P(u_z)}{1-P(u_x)} \times \sum_{c=1}^{k} \frac{d(f_i,t_j,\tilde{t}_{zc})}{l \times k} \right], \text{ где } P(u_x) < 1 \quad (1)$$

$$d(f_i,t_j,t_o) = \frac{|f_{ij} - f_{io}|}{\max(f_i) - \min(f_i)}, \text{ если } \max(f_i) \neq \min(f_i)$$

Вес признака тем больше, чем меньше расстояние по значению данного признака до сообщений того же пользователя и чем больше расстояние по значению этого признака до сообщений других пользователей.

При расчете весов используется нормализованное расстояние. Т. к. разные признаки имеют разные абсолютные значения. Нормализация позволяет привести все используемые числовые значения к одинаковой области их изменения, благодаря чему появляется возможность свести их вместе.

Также в расчете влияния признака на качество разделения учитывается априорная вероятность появления сообщений данного пользователя. Т. к. если сообщений пользователя мало и вероятности не учитываются, то получится что вес признака будет существенно увеличиваться на значение расстояния для пользователей, у которых мало сообщений.

При решении задачи поиска автора сообщения в качестве классификатора может быть использован, например, один из следующих алгоритмов:

1. Нейронная сеть.
2. Деревья решений.
3. Случайный лес.
4. Метод опорных векторов.
5. Логистическая регрессия.
6. Наивный байесовский классификатор.

Рассмотрим решение этой задачи методом Случайный лес. Этот алгоритм был выбран т. к. он:

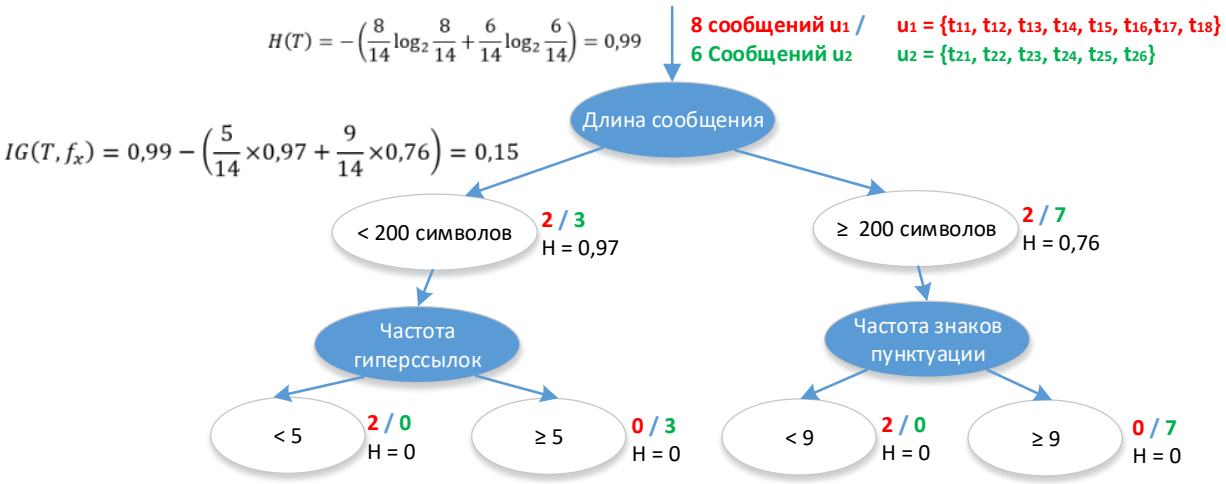
- способен обрабатывать модели высокой размерности;
- имеет возможность работы со всеми типами признаков (дискретные, непрерывные, бинарные, символьные и пр.);
- обладает высокой точностью на несбалансированных данных и в случаях малого количества обучающих примеров;
- он достаточно быстро обучается и показатели качества его работы сопоставимы с методом опорных векторов и нейронных сетей, достаточно часто применяемых в задачах идентификации.

Данный метод достаточно эффективно применяется для решения различных задач классификации и прогнозирования. Основной идеей алгоритма RF является построение ансамбля (или леса) случайных деревьев принятия решений. В основе метода лежит идея бэггинга – комбинация независимых моделей позволяет повысить точность классификации. Т. к. предполагается, что большинство сгенерированных деревьев сами по себе верно

предсказывают пользователя, и что, деревья, которые ошибаются, выдают различные результирующие классы.

Структура дерева представляет собой древовидный граф, имеющий в своем составе ребра (ветви) и узлы двух типов: листья и внутренние узлы. В листьях дерева содержатся значения целевой функции, в нашем случае класс – пользователь, на ребрах записаны значения признаков, от которых зависит значение целевой функции, в узлах – сами признаки. Чтобы классифицировать сообщение, необходимо спуститься от корня дерева к листу, и получить значение класса, в нем содержащееся.

Каждое из деревьев ансамбля строятся по случайной подвыборке сообщений (с повторением) и случайному набору признаков. При построении дерева производится ветвление по одному из признаков, дающему лучшее разбиение примеров (в нашем примере - по принципу минимизации энтропии, здесь используется критерий прироста



$H(T)$ – энтропия до разделения по признаку,

$IG(T, f_x)$ – (Information Gain), прирост информации при разделении T по f_x,

f_x – значение признака, по которому производится ветвление.

информации). Помимо того, что необходимо отобрать лучший признак, еще необходимо выбрать значение этого признака, которое разделит примеры наиболее хорошо (отбирается одно значение также максимизирующее прирост информации).

В классическом алгоритме в качестве критерия разделения выборки применялся индекс Джини. И индекс Джини, и прирост информации показывают, насколько неоднородна выборка. Ветвление заканчивается, когда в узле оказываются сообщений одного пользователя.

Классификация производится по принципу голосования: каждое дерево леса классифицирует объект к одному из классов, т. е. голосует за определенный класс. Далее объект относится к тому классу, за который проголосовало наибольшее число деревьев.

Обучив нашу модель и проведя классификацию сообщения неизвестного автора, удалось выявить нарушителя, который впоследствии на допросе признал свою вину.

Поведение пользователя может анализироваться не только в онлайн, но и в оффлайн сервисах. В таких случаях в качестве свойств для анализа могут браться визуальные характеристики пользователя (например, черты лица) и акустические (его голос). Одним из интересных примеров является использование особенностей взаимодействия пользователя со смартфоном для идентификации. В этом случае процесс идентификации может быть построен на таких характеристиках, как особенность взаимодействия с экраном смартфона (т. е. продолжительность прикосновений, особенности позиционирования и т. д.). Здесь первоначально собирается статистика взаимодействия с устройством, т. е. особенности движений при работе со смартфоном. После сбора и обработки информации получается портрет пользователя, основанный на его действиях. В следующий раз при работе со смартфоном последний будет проверять информацию, получаемую в ходе текущего контакта пользователя, с той информацией, которая была получена ранее.

Дополнительные материалы:

1. Robnik-Šikonja M., Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF //Machine learning. – 2003. – Т. 53. – №. 1-2. – С. 23-69. (https://link.springer.com/article/10.1023/A:1025667309714)

Использование ИИ для обнаружения атак на систему. Пример на сетевой безопасности.

Для того, чтобы говорить об атаках и тем более о методах и системах их обнаружении, необходимо для начала дать определения базовым понятиям информационной безопасности.

Уязвимость — дефект, возникший на этапе проектирования, реализации или эксплуатации и потенциально способный привести к нарушению информационной безопасности.

Угроза безопасности — совокупность условий и факторов, создающих потенциальную или реально существующую опасность нарушения информационной безопасности.

Атака — это поиск и использование злоумышленником уязвимостей. Другими словами, атака — это реализация угрозы.

Нарушителем информационной безопасности является физическое лицо или логический объект, случайно или преднамеренно совершивший действие, следствием которого является нарушение информационной безопасности.

Атака — любая область, где неуполномоченный пользователь может работать и внедрять свой данные или код. Атаки можно разделить на три области: сети, программное обеспечение, атаки на человеческий фактор.

В этом разделе мы рассмотрим сетевые атаки. Под сетевой атакой принято понимать действия с применением программных и технических средств и с использованием сетевого протокола, направленные на реализацию угроз несанкционированного доступа к информации, воздействия на нее или на ресурсы автоматизированной системы.

В своей работе "Компьютерные атаки: что это и как им противостоять" Питер Мэлл выделил следующие типы атак:

- удаленное проникновение — это тип атак, которые позволяют реализовать удаленное управление компьютером через сеть;
- локальное проникновение — это тип атак, которые приводят к получению несанкционированного доступа к узлу, на который они направлены;
- удаленный отказ в обслуживании — тип атак, которые позволяют нарушить функционирование системы в рамках глобальной сети;
- локальный отказ в обслуживании — тип атак, позволяющих нарушить функционирование системы в рамках локальной сети;

- атаки с использованием сетевых сканеров — программ, которые анализируют топологию сети и обнаруживают сервисы, доступные для атаки;
- атаки с использованием сканеров уязвимостей — т. е. программ, осуществляющих поиск уязвимостей на узлах сети, которые в дальнейшем могут быть применены для реализации сетевых атак;
- атаки с использованием взломщиков паролей — программ, подбирающих пароли пользователей;
- атаки с использованием анализаторов протоколов — они основаны на использовании программ, "прослушивающих сетевой трафик.

Какие же существуют подходы к обнаружению атак и идентификации источника атаки.

В первую очередь, когда мы говорим о сетевой безопасности и машинном обучении, стоит сказать о системах обнаружения вторжений или Intrusion detection systems (IDS).

История IDS начинается еще в 1980-ых годах. К этому периоду относятся экспертные системы, основанные на правилах, и действовавшие на основании статистических методов. Анализируя сетевой трафик и данные приложений пользователей, они выявляли подозрительную или вредоносную активность. В 1993 году появились первые системы, в которых применялись нейронные сети.

Искусственный интеллект дает возможность увеличить скорость обработки больших объемов получаемых данных или событий, анализа инцидентов, позволяет находить зависимости, в соответствии с которыми коррелируют различные события, происходившие в разное время в разных местах и таким образом находить атакующие источники.

Есть две основные группы методов построения систем обнаружения вторжений — это:

- обнаружение аномалий (anomaly-based);
- обнаружение злоупотреблений (misuse detection или signature-based).

Методы обнаружения аномалий

Эти методы основаны на анализе поведения и позволяет детектировать новые еще неизвестные типы атак.

Системе известны некоторые признаки, характеризующие правильное или допустимое поведение объекта наблюдения. Под нормальным или правильным поведением понимаются действия, выполняемые объектом и не противоречащие политике безопасности.

Примерами аномального поведения могут быть большое число соединений за короткий промежуток времени, высокая загрузка процессора и т. п. Если можно однозначно описать профиль нормального поведения, то любое отклонение от него можно идентифицировать как аномальное.

Для выявления сетевых аномалий необходимы входные данные — т. е. сетевой трафик, собранный за фиксированный промежуток времени и имеющий определенные признаки. Эти признаки сравниваются с имеющимся шаблоном нормального поведения и на основании величины расхождения значений делается вывод о наличии сетевой аномалии. В случае, если аномалия не была зафиксирована, шаблон нормального поведения корректируется с учетом полученных значений признаков сетевого трафика. Наиболее распространенным способом обнаружения аномалий является сравнение полученных данных сетевого трафика с заданными предельными значениями — при этом в случае значительного превышения предельного значения определяется сетевая аномалия. Одним из примеров может выступать детектирование подозрительной активности пользователей. Например, обнаружение сетевой активности, нехарактерной для ПК может служить индикатором несанкционированного доступа к машине. В качестве нехарактерной сетевой активности можно привести пример сканирования портов, проводимый из отдела кадров и т. д.

Системы обнаружения злоупотреблений или злоумышленного поведения основаны на том, что известны методы вторжений и признаки атаки или признаки, характеризующие поведение злоумышленника.

Такие системы производят детектирование атак, основываясь на сигнатурном анализе процессов компьютерной системы или анализе сетевого трафика. Атаки описываются в виде сигнатуры (signature) и далее производится поиск этой сигнатуры в контролируемом пространстве. Основным недостатком является то, что они неэффективны против новых атак, необходимо постоянное обновление и обучение по новым данным.

При обнаружении злоупотреблений получаемые входные данные (т. е. сетевой трафик - так же, как и в методе выявления аномалий) сравниваются с заранее определенным шаблоном атак. Шаблоном атаки в данном случае является набор признаков, описывающий определенную атаку. Основным отличием метода обнаружения злоупотреблений от метода обнаружения аномалий является сравнение атрибутов сетевого трафика не с пороговыми значениями, а с шаблонами атак — например, совокупностью команд, позволяющих получить информацию об атакуемой системе. При работе метода обнаружения

злоупотреблений исследуемый трафик относится к вредоносному в случае совпадения с каким-либо из имеющихся шаблонов атак

Одно из новых направлений применения машинного обучение для сетевой защиты — это анализ сетевого трафика, который позволяет проводить углубленный анализ всего трафика на каждом уровне и выявлять атаки и аномалии. Машинное обучение применяется здесь для решения трех основных задач:

- регрессия для прогнозирования параметров сетевых пакетов и сравнения их с обычными параметрами;
- классификация для идентификации различных типов атак, таких как спуфинг;
- кластеризация для расследования инцидентов.

Рассмотрим решение задачи обнаружения и идентификации различных атак в сетевом трафике. В качестве исходных данных будем использовать классический дата сет — UNSW-NB15 (https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/), содержащий данные, сгенерированные на основе современных паттернов «нормального» сетевого трафика и различных атак, что позволяет применять алгоритмы, обученные на этом датасете к реальным задачам.

Рассмотрим задачу определения наличия сетевой атаки в трафике. Необходимо на основании значений признаков определить значение поля label. Именно оно идентифицирует, является ли данный запрос валидным, или же несет угрозу системе. В данном наборе данных представлено 9 видов атак: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms.

Возьмем набор данных, разделим его на обучающую и тестовую выборку. Большая часть - 90% записей будет использована для обучения модели, а 10% для тестирования. Обучим нашу модель. Будем снова использовать Случайный Лес. В результате обучения получим классификатор, позволяющий отнести трафик к одной из двух категорий - аномальный и нормальный. Проверим показатели качества классификации: точность или достоверность (accuracy) классификации — это показатель, который говорит о том, какая доля записей была классифицирована верно. В нашем случае она является достаточно высокой.

Обнаружение подозрительной активности пользователей и сетевого трафика – самое очевидное применение машинного обучения. Нынешние системы все успешнее

справляются с выявлением необычных событий в больших потоках данных, решением стандартных задач анализа и рассылкой уведомлений.

Следующий шаг – использование ИИ для борьбы с более сложными проблемами. Например, уровень киберриска для компании в каждый конкретный момент зависит от множества факторов, в том числе от наличия систем без “заплат” (Hot Fix), незащищенных портов, поступления сообщений направленного фишинга, уровня надежности паролей, объема незашифрованных конфиденциальных данных, а также от того, является ли организация объектом атаки со стороны спецслужб другого государства.

Доступность точной картины рисков позволила бы рациональнее использовать ресурсы и разработать более детальный набор показателей эффективности обеспечения безопасности. Сегодня соответствующие данные либо не собираются, либо не преобразуются в осмысленные сведения, и это создает широкие перспективы для работы с машинным обучением.

В компаниях уже начали пользоваться искусственным интеллектом и машинным обучением для распознавания угроз безопасности и реагирования на них. Например, в банке Barclays Africa применяют искусственный интеллект для обнаружения признаков компрометации систем в локальной корпоративной сети и в облаке. В перспективе искусственный интеллект может помогать компаниям определяться, в какие новые технологии безопасности следует вкладываться.

Дополнительные материалы:

1. В качестве инструмента для анализа данных предлагается использовать RapidMiner - программная платформа для обработки данных - https://rapidminer.com/educational-program/
2. Общее описание данной платформы представлено в статье - Введение в RapidMiner — https://habr.com/ru/post/269427/
3. В примере, рассмотренном в данной части лекции используется часть данных из набора UNSW-NB15. Файл с данными в формате .csv доступен для скачивания по ссылке — https://drive.google.com/open?id=1weZZ_syA-JXmmx33_FInh2hYjjHiyMXY
4. Файл процесса, созданного в среде RapidMiner, демонстрирующего решение задачи классификации трафика – https://drive.google.com/open?id=14hWiVSDf9QTkBGI-Q313jIsCTr_Vjm60

5. Описание UNSW-NB15 dataset
   https://docs.google.com/document/d/1Pw2sZdv3VTiNeCwxTlqphPv0GC31lWr98tAOpmUMges
6. Виды атак —
   https://docs.google.com/document/d/1_eawdXI29HlrdqyAfEUreF7smsMJsraplcpvcJ_02eE/