Machine learning в информационной безопасности

Искусственный интеллект[1] захватывает мир

IA с каждым годом все больше проникает в повседневную жизнь любого человека и дело не только в моде на такие технологии, главное достоинство такого метода заключается в более эффективной и быстрой работе программ. Например, при обнаружении мошенничества, переводе текстов на разные языки, распознавание изображений и т. д. Чаще всего методы ML используются в голосовых помощниках (Алиса, Google Assistant), программирование беспилотных автомобилей, теперь же технологии умных машин продолжают интегрироваться в многие другие области научных исследований и использоваться в бизнес-решениях.

Неудивительно, что сфера информационной безопасности не осталась в стороне от всеобщего бума на IA, более того сфера ИБ и кибербезопасности являются одними из самых приоритетных для внедрения высокотехнологических решений. С развитием сферы искусственного интеллекта обычные методы решения и предотвращения атак и инцидентов, количество которых растет с каждым годом, больше не являются жизнеспособными. Возможности автоматизации и аналитики IA и ML помогают устранить пробелы в кибербезопасности, с помощью обнаружения скрытых закономерностей при выявлении атак и автоматически нейтрализовать их. Тендем, состоящий из Big Data и высокоскоростного Интернета, предоставляет специалистом возможность разрабатывать с помощью IA передовые инструменты кибербезопасности, помогающие предотвращать кибератаки. В отчете Capgemini Research Institute говорится, что 61% организаций утверждают, что не смогут *выявить* критические угрозы без IA, а 69% считают, что IA будет необходим для *реагирования* на кибератаки. Ожидается, что к 2027 году рынок ИИ в сфере кибербезопасности вырастет до 46,3 млрд долларов[2].

1.5 года назад ВЭФ (всемирный экономический форум) заявил[3], что если не начать принимать меры, то новейшие технологии смогут нанести сокрушительный удар на глобальное сообщество безопасности уже в 2025 году. Именно поэтому IA и ML вторая по важности тенденцией формирования киберпространства. На сегодняшний день еще непонятно, как будут уравновешиваться силы между атакующими, использующими IA, и защищающимися, использующими эту же технологию. Чтобы снизить риск того, что преимущество будет на стороне киберпреступников и других злонамеренных субъектов, крайне важно, чтобы сообщество кибербезопасности быстро подготовиться к борьбе с помощью более быстрых и динамичных средств защиты с поддержкой IA.

Преимущества IA и ML

Организации, использующие IA и ML для защиты своих данных, получают такие преимущества, как:

• **Повышение скорости на реагирование и обнаружение инцидентов информационной безопасности.** IA и ML с легкостью анализируют огромные объемы данных за секунды, поэтому превосходят ручное обнаружение угроз. Более того, параллельно может происходить исправление/устранение угрозы в режиме реального времени. С учетом способности современных кибератак быстро проникать в инфраструктуру организации, быстрое обнаружение и реагирование является одним из залогов успеха в борьбе с ними.

---

[1] Далее в тексте будут использоваться сокращения:

- **AI** (**Artificial intelligence**) – искусственный интеллект
- **ML** (**Machine learning**) – машинное обучение

[2] https://www.meticulousresearch.com/product/artificial-intelligence-in-cybersecurity-market-5101

[3]https://www3.weforum.org/docs/WEF_Future_Series_Cybersecurity_emerging_technology_and_systemic_risk_2020.pdf

• **Снижение затрат на IT-разработку**. IA и ML экономически эффективные технологии. Из отчета Capgemini следует, что среднее снижение затрат составляет 12%, а некоторые организации сократили свои расходы более чем на 15%[4].

• **Повышение эффективности работы аналитиков**. Интеграция методов IA и ML снижает рабочую нагрузку на аналитиков, путем сокращения времени ручного просмотра журналов данных. IA сам классифицирует атаки, находит закономерности, которые мог проглядеть специалист, тем самым сокращая рутинную работу аналитика и оставляет большое количество времени на решение других важных и серьезных задач.

• **Улучшение общего уровня безопасности организации.** За счет внедрения IA и ML безопасность компании будет со временем улучшаться из-за постоянного анализа больших данных умными машинами, постоянное обучение позволяет выявлять и предотвращать больше возможных угроз. Кроме того, происходит защита инфраструктуры организации на макро- и микроуровнях, что приводит к созданию более эффективного барьера против вторжений злоумышленников, в сравнении с ручными методами.

Разберемся с понятиями

Настало время чуть подробнее разобраться в том, что же такое ML и IA, являются ли эти понятия синонимичными и какие методы ML существуют и используются на текущий момент.

IA— это способность компьютерной системы имитировать когнитивные функции человека, такие как обучение и решение проблем. Один из способов обучить компьютер имитировать мышление человека— использовать нейронную сеть.

ML - это способ развития интеллекта компьютерной системой. По словам Артура Самуэля «машинное обучения – это возможность обучения компьютера без явного программирования.» Если выражаться более формально, то процесс ML можно описать так: компьютерная программа учится на опыте **E (experience)** на классе задач **T (task)** с производительностью P **(performance measure),** если ее P улучшается во вовремя выполнения T на E.

Нейронная сеть (Neural Networks) - это вычислительные системы с взаимосвязанными узлами, которые работают подобно нейронам в человеческом мозге.

- **Глубокое обучение** является частью более широкого семейства методов машинного обучения, основанных на представлениях обучающих данных, а не на алгоритмах для конкретных задач.
- **Machine learning** использует математические алгоритмы. Их использование предоставляет компьютерным системам возможность "учиться" используя данные, а не быть явно запрограммированными. Примеры таких алгоритмов: логистическая и линейная регрессия, дерево решений, алгоритм К-ближайших соседей и т.д. Методы ML хорошо работают при решении конкретной задачи.

Виды обучения для решения задач с помощью ML алгоритмов:

- **Обучение с учителем (Supervised learning)** – в этом способе есть размеченный набор данных (признаки модели) и целевые переменные (верные ответы). Использую известные данные, алгоритм при таком виде обучения пытается предсказать наиболее вероятные ответы для неизвестных объектов. Именно наличие ответов создает некую иллюзию существования учителя.
- **Обучение без учителя (Unsupervised learning)** - с помощью такого метода, мы можем подходить к решению задач, практически не имея понятия о том, как должны выглядеть наши результаты. При обучение без учителя, обратная связь по результатам прогнозов отсутствует.

[4] https://www.capgemini.com/wp-content/uploads/2019/07/AI-in-Cybersecurity_Report_20190711_V06.pdf

- **Обучение с подкреплением (Reinforcement learning)** - частный случай обучения с учителем, при котором «учителем» является среда функционирования, дающая обратную связь информационной системе в зависимости от принятых ею решений.

Какие задачи в сфере ИБ могут решаться

IDS/IPS

Глубокое обучение с подкреплением (DLS) оказалось хорошим инструментом[5] для разработки **систем обнаружения и предотвращения вторжений (IDS/IPS)**, но это не единственная хороший метод решение этой задачи[6-7]

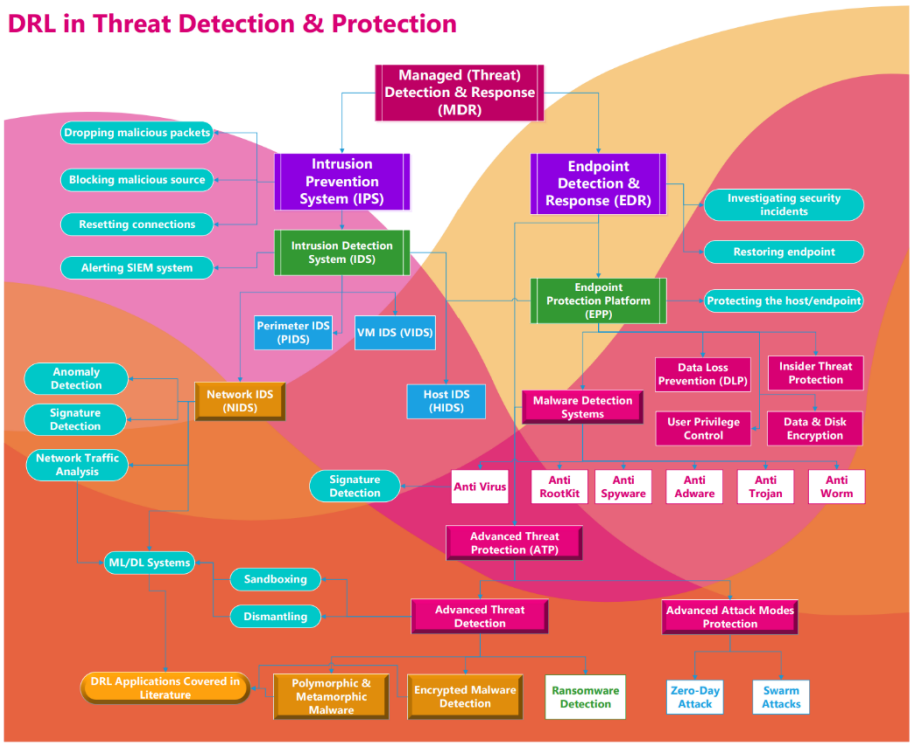

Рис. 1. Схема интегрирования механизмов DLR в MDR.

В NIDS (сетевая система обнаружения вторжений) может **выполняться задача идентификации ботнетов** (распространенные атаки: DDoS, финансовые нарушения, спам по электронной почте и т.д). Согласно статье[8] ботнеты представляют собой основной источник нежелательного интернет-трафика, кроме того, в исследовании было обнаружено, что признаки ботнет-инфекции были обнаружены в 11% из 800 000 исследованных DNS-доменов, что указывает на большое разнообразие жертв ботнетов.

Дополнительно, может выполняться **задача бинарной классификации: является ли трафик аномальным или нет**[9], если да, то можно построить другой классификатор, который будет определять вид аномалии (рис. 2).

---

[5] https://arxiv.org/pdf/2206.02733.pdf
[6] https://www.mdpi.com/2076-3417/12/3/1205
[7] https://www.mdpi.com/2076-3417/12/2/852
[8] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.541.4324&rep=rep1&type=pdf
[9] Anomaly_Detection.ipynb

Чтобы **защитить конечные точки от вредоносных программ**[10,11,12,13,14] также можно использовать DLR или другие методы ML(рис. 2). Программа с использование ML технологий сможет построить классификатор определяющие самые разнообразные угрозы (черви, криптоджекинг, трояны и т.д.). Обработка больших данных в режиме реального времени позволяет таким методам **прогнозировать угрозы и нейтрализовывать их** до того, как они успеют нанести вред. Более того они позволяют **прогнозировать утечки данных.**

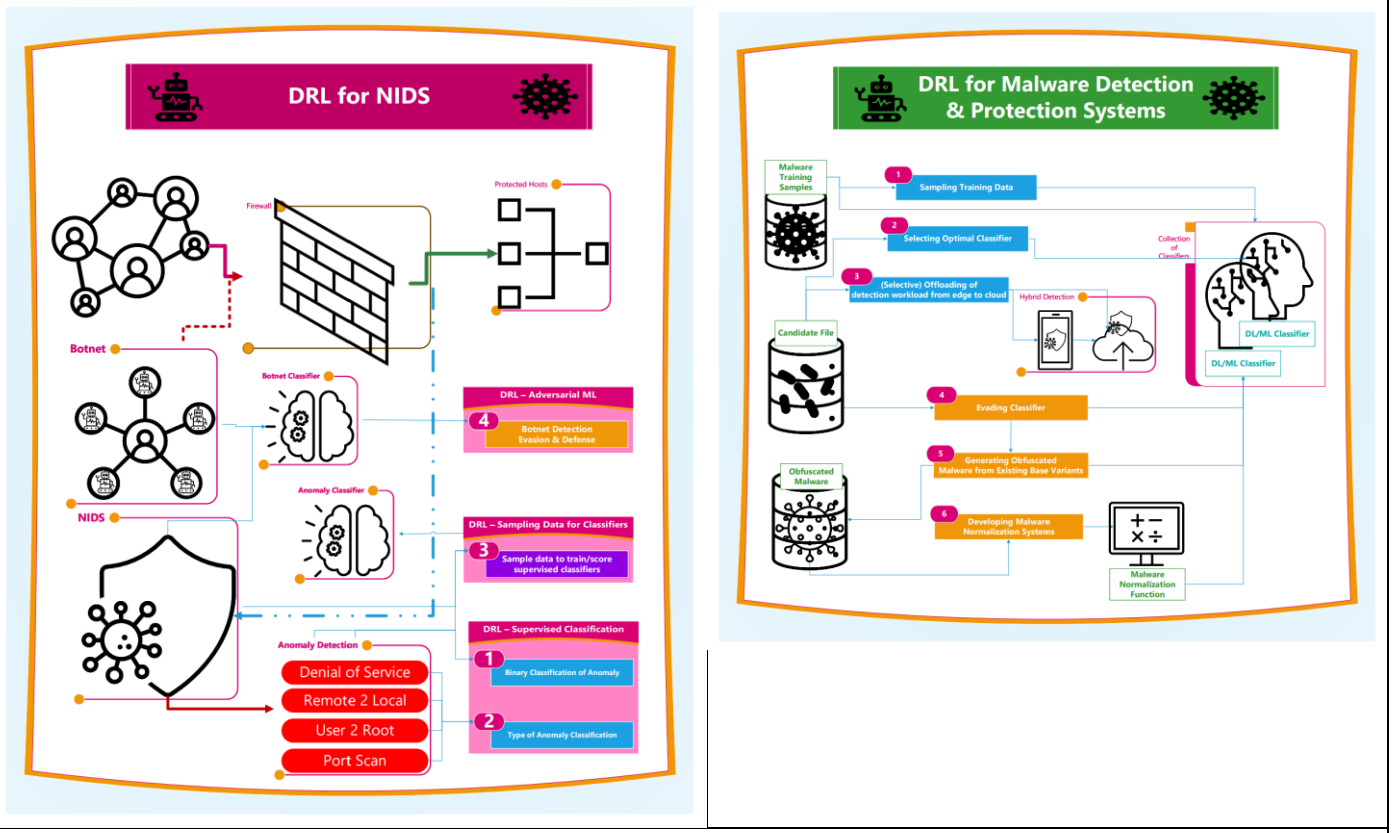

Рис. 2. Слева показаны задачи, которые может решать DLR в сетевой системе обнаружения вторжений (NIDS). Справа показаны задачи, которые может решать DLR в целях обнаружения вредоносных программ и защите конечных точек.

Тестирование на проникновение

Методы ML также позволяют устраивать **автономные тестирования на проникновения**[15]. Традиционные методы в значительной степени зависят от знаний экспертов, что требует непомерных трудовых и временных затрат. Автономное тестирование на проникновение является более эффективным и интеллектуальным способом решения этой проблемы.

Защита от внутренних угроз

Человеческие ошибки, такие как своевременно необнаруженные неточности в коде, ошибки конфигурирования и другие проблемы, часто являются причиной нарушения безопасности данных.

[10] https://www.mdpi.com/2076-3417/11/14/6446
[11] https://arxiv.org/pdf/2206.02733.pdf
[12] https://media.kaspersky.com/en/enterprise-security/Kaspersky-Lab-Whitepaper-Machine-Learning.pdf
[13] https://www.sciencedirect.com/science/article/pii/S2405959522000637#b11
[14] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00318-5
[15] https://www.mdpi.com/2076-3417/11/19/8823

Ручное выявления требуется много людей и времени, в течение которого система, сервер или другие ресурсы с данными будут уязвимыми. ML в этом случае можно использовать для **обнаружения уязвимых мест и их точной локализации.**

Еще одной важной практикой безопасности является **отслеживание и анализ действий и поведения пользователей**[16][17], что гораздо сложнее, чем распознать традиционные вредоносные действия против сетей. Зная типичное поведение пользователя, система может отправить предупреждение аналитикам ИБ в случае существенного изменения модели работы сотрудника (посещение подозрительных сайтов, длительное отсутствие за рабочим ПК, изменение круга общения при переписке в корпоративном мессенджере и т.д. Более того, алгоритмы ML могут соотносить и анализировать различные признаки файлов, тем самым **обнаруживать утечки из-за неправомерных действий привилегированных пользователей**.

Аутентификация и распознавание фальшивых документов[18,19], биометрических данных[20,21] и прочих идентификаторов

Системы защиты, оснащенные компьютерным зрением и обработкой речи, могут **оперативно выявлять «подмену личностей» или несанкционированное присвоение прав доступа** и оповещать центр безопасности.

Обнаружение атак, для которых используется метод перебора. Также бывают случаи, когда злоумышленники запускают вброс учетных данных или атаки методом перебора для доступа к учетным записям, что может привести к нарушениям в сети компании, алгоритмы ML могут легко предотвратить атаки такого типа. Обученная система может замечать и останавливать эти атаки на этапе проб и ошибок.

Безопасность в финансовой сфере

Использование ML, как средство для **обнаружение инсайдерской торговли или других мошеннических разговоров в электронных письмах и телефонных звонках. [22].** Также применение методов ML помогает быстро и эффективно **обнаруживать мошеннические операции**[23][24][25](antifraud), например, когда сценарий использования банковской карты отличается от привычного. Внедрение таких технологий так же, как и в многих других случая не только улучшает распознавание мошеннических операций, но и уменьшает время ручного просмотра и кол-во сотрудников в штате. Вместо того, чтобы нанимать сотни сотрудников достаточно иметь только одну система ML для обработки всех данных. Это идеально подходит для предприятий с сезонными приливами и отливами трафика, касс или регистраций. Система машинного обучения — отличный союзник для масштабирования компании без резкого увеличения затрат на управление рисками. Более того, такие

[16] https://www.imperva.com/learn/application-security/insider-threats/
[17]https://www.researchgate.net/publication/276175716_An_Internal_Intrusion_Detection_and_Protection_System_by_Using_Data_Mining_and_Forensic_Techniques
[18] https://www.ijert.org/fake-education-document-detection-using-image-processing-and-deep-learning
[19] https://arxiv.org/pdf/2102.00653.pdf
[20] https://iopscience.iop.org/article/10.1088/1742-6596/1098/1/012017/pdf
[21]https://www.researchgate.net/publication/344663459_Person_Authentication_Based_on_Biometric_Traits_Using_Machine_Learning_Techniques
[22] https://www.stevens.edu/news/insider-trading-financial-fraud-misconduct-theres-stevensaccenture-communications-surveillance-app
[23] https://seon.io/resources/fraud-detection-with-machine-learning/
[24] https://sdk.finance/all-you-need-to-know-about-machine-learning-based-fraud-detection-systems/
[25] https://cyberleninka.ru/article/n/primenenie-algoritmov-mashinnogo-obucheniya-k-zadache-vyyavleniya-moshennichestva-pri-ispolzovanii-plastikovyh-kart/viewer

системы могут работать круглосуточно и им не нужны перерывы в отличии от людей, а мошеннические атаки могут происходить в любой момент времени.

Adversarial Machine Learning (AML)

Как бы прекрасно ни звучало все вышеизложенное, **методы ML и распознавания образов создают специфические уязвимости**, которые опытные злоумышленники могут использовать для компрометации всей системы, т. Е. само ML может быть самым слабым звеном в цепи безопасности.[30]
AML – это ветвь ML, ставшая теоретической основой для разработки инструментов, способных создавать помехи в работе систем на основе ML. По аналогии с malware этот термин можно назвать как «вредоносное машинное обучение».
Самая первая, основополагающая работа в области AML датируется 2004 годом. В то время исследовали проблему в контексте фильтрации спама, показав, что линейные классификаторы можно легко обмануть с помощью небольшим количеством тщательно продуманных изменений в содержании спама, без существенного влияния на читаемость спам-сообщения. Это действительно были первые примеры противника против линейных классификаторов для фильтрации спама.
Поэтому безопасность алгоритмов глубокого обучения, в контексте компьютерного зрения и задач кибербезопасности заслуживает особого внимания.
Возможные виды атак[31]:

- **атака с отравлением набора данных**. Для этой атаки злоумышленник использует различные методы для вторжения в обучающие и тестовые данные, чтобы повлиять на нормальное функционирование задачи ML. Злоумышленник может использовать враждебные примеры для атаки на сервер данных, откуда должны быть извлечены исходные данные. Компрометация источников данных помогает вставить ошибочные данные, которые, возможно, изменяют функционирование модели ML. Это еще больше изменяет результаты работы системы, основанной на ML.
- **атака с отравлением модели**. Здесь злоумышленник изменяет параметры, через которые он может заставлять алгоритм ML генерировать ошибочный результат, вмешиваясь в работу классификатора. Параметры, с помощью которых классификатор готовит ML-модель, изменяются. Злоумышленник может изменить пределы чувствительности, скорость присоединения и вызвать заниженную или завышенную подгонку, что в дальнейшем влияет на нормальное выполнение задачи ML
- **атака с нарушением конфиденциальности**. Незащищенные файлы и отсутствие механизма шифрования на этапах обучения и развертывания ML-задачи могут привести к утечке данных. Это позволяет неавторизованному пользователю вмешиваться в работу модели и увеличивает риски конфиденциальности, связанные с данными, поскольку конфиденциальность чувствительных данных может быть нарушена.
- **атака с нарушением времени выполнения**. Злоумышленник использует эту задачу для задержки или завершения текущей задачи ML, обычно нацеливаясь на сервер во время фазы развертывания. Именно на этой фазе он пытается удаленно нарушить текущий процесс ML. В результате чего нормальное функционирование задачи ML нарушается, что приводит к потере времени и ресурсов. Злоумышленник определяет слабые места (уязвимости) и проникает на сервер времени выполнения с помощью различных атак, таких как фишинг, атака отказа в обслуживании (DoS) и атака SQL-инъекции. Эта атака может быть ослаблена путем децентрализации рабочего пространства ML.
- **Бэкдоры и троянские атаки**[32,33]. Такие атаки злонамеренно манипулируют предварительно обученными сетевыми моделями, чтобы создания определенных уязвимостей в виде бэкдоров. Испорченные модели затем публикуются в открытом доступе, чтобы способствовать их

[30] https://arxiv.org/pdf/1712.03141.pdf
[31] https://www.sciencedirect.com/science/article/pii/S2405959522000637#b11
[32] https://www.arxiv-vanity.com/papers/1708.06733/
[33] https://arxiv.org/pdf/1712.05526.pdf

внедрению в собственные системы (например, с помощью тонкой настройки). Когда это происходит, злоумышленник может активировать бэкдоры, используя определенные входные образцы, которые неправильно классифицируются как нужные.

Все вышеупомянутые атаки могут быть успешно проведены при различных уровнях знаний атакующего.

Вывод

Несмотря на огромный потенциал и важность интегрирования ML алгоритмов в информационные системы безопасности, не стоит забывать, что эти инструменты тоже могут подвергаться серьезным атакам причем более трудно обнаружимым. Архитектура нулевого доверия должна применяться и к ML. Нет систем без уязвимостей. При наличии достаточного количества времени каждую систему можно взломать или использовать, поэтому мы должны относиться к системам ML с осторожностью и принимать необходимые меры предосторожности. Также стоит отметить, что ML может быть и должно быть дополнительным звеном для противодействия нарушителям, но не может полностью заменить обычных методов защиты безопасности и обычных людей. Полному отказу от прежних методов в пользу ML препятствуют такие причины, как:

- отсутствие непосредственной обратной связи в когнитивном плане. Зачастую для нас остается загадкой то, почему именно ML модель обучалась именно так на выходных данных и никак иначе и получила, основываясь на них именно такой результат. Именно такое отсутствие обратной связи не дает возможности полноценно отказаться от человеческого контроля.
- отсутствие достаточного для корректного обучения ML-моделей кол-во данных по всем направлениям киберугроз, от компьютерных вирусов до приемов социальной инженерии
- возможность специфических атак на ML-алгоритмы и используемые датасеты, что может привести к неверным решениям, пропущенным атакам или ложным срабатываниям;