Вопросы Применения Методов Машинного Обучения для Решения Проблем Информационной Безопасности

Ядигар Имамвердиев[1], Людмила Сухостат[2]
[1,2]Институт Информационных Технологий НАНА, Баку, Азербайджан
[1]*yadigar@lan.ab.az*, [2]*lsuhostat@hotmail.com*

***Аннотация* – в статье представлен обзор основных методов обнаружения аномалий на основе машинного обучения для решения проблем информационной безопасности. Рассмотрены методы классификации и кластеризации. Приведены проблемы машинного обучения во враждебной среде.**

Ключевые слова – машинное обучение, информационная безопасность, машинное обучение во враждебной среде, обнаружение аномалий.

I. ВВЕДЕНИЕ

Машинное обучение изучает алгоритмы, которые улучшают характеристики работы путем накопления опыта. Методы машинного обучения доказали свою большую практическую эффективность во многих областях применения. Они особенно полезны: (а) в плохо понимаемых предметных областях, где существует мало знаний, чтобы разработать эффективные алгоритмы; (б) в областях, где имеются большие базы данных, содержащие ценные неявные закономерности, которые могут быть обнаружены, или (в) в областях, где программы должны адаптироваться к меняющимся условиям. Неудивительно, что поле информационной безопасности оказывается благодатной почвой, где многие задачи могут быть сформулированы как проблемы обучения и изучены с точки зрения алгоритмов обучения.

По сути, машинное обучение изучает задачи применения математических методов к большому количеству данных для прогноза вероятностей, например, что электронное письмо является спамом. Эти системы работают эффективно благодаря поступлению большого количества данных, на основе которых они могут строить свои прогнозы. Более того, системы спроектированы таким образом, чтобы со временем улучшаться за счет отслеживания самых полезных сигналов и моделей по мере поступления новых данных.

Технологии машинного обучения позволяют максимально эффективно использовать накопленные в организации данные о безопасности и поэтому получают все большее распространение в сфере информационной безопасности. Но информационная безопасность имеет дополнительные трудности для успешного применения методов машинного обучения.

Данная работа посвящена анализу вопросов применения методов машинного обучения для решения проблем информационной безопасности.

II. ОБЗОР МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ

Методы обнаружения вторжений могут быть разделены на два класса:

1) методы, основанные на обнаружении злоупотреблений (misuse detection);
2) методы, основанные на выявлении аномалий (anomaly detection, AD).

Сигнатурный подход занимает центральное место в большинстве современных антивирусов и обеспечивает на уже известных вторжениях 100-процентную точность обнаружения. Но сигнатурный анализ не работает на вирусах, не известных антивирусу, т.е. если не известны их сигнатуры. Так называемые вторжения нулевого дня (zero-day). Их обнаружить позволяет аномальный подход. Такие методы строят классификаторы или решающие правила. В последнее время наряду с этими методами [1, 2] большой интерес представляют методы, использующие машинное обучение.

Три основные задачи машинного обучения – классификация, кластеризация и регрессия – успешно применялись для обнаружения атак, идентификации и анализа угроз и вредоносных программ [3].

Методы классификации. Классификация (обучение с учителем) является классической задачей анализа данных, с корнями, уходящими в машинное обучение [4].

Процесс классификации включает в себя следующие этапы:

1) создание обучающего набора данных;
2) определение атрибута и классов;
3) определение полезных атрибутов для классификации (релевантный анализ);
4) обучение модели использованию примеров из обучающей выборки;
5) использование модели классификации неизвестных образцов данных.

Можно выделить следующие методы классификации.

а) Деревья принятия решений

Дерево решений – это древовидная структура, где каждый нетерминальный узел представляет испытание или решение о рассматриваемом элементе данных. Выбор определенной ветки зависит от исхода испытания. Решение принимается тогда, когда найден терминальный узел. Дерево решений – особая форма набора правил, характеризуемая их иерархической организацией.

б) Нейронные сети

Нейронные сети (НС) являются системами, смоделированными на основе работы мозга человека, состоящего из миллионов нейронов, соединенных между собой с помощью синапсов. Нейронная сеть – набор соединенных модулей входа/выхода, в котором каждое соединение имеет вес, связанный с ним. Обучение сети направлено на определение набора весов с целью минимизации ошибки классификации. Искусственная НС используется для классификации и предсказания латентных переменных, которые сложно измерить, и решения задачи нелинейной классификации. Результаты классификации НС трудно интерпретировать в отличие от других методов классификации, где имеется функциональная зависимость между данными (например ассоциативные правила).

в) Наивные байесовские классификаторы

Наивные байесовские классификаторы используют теорему Байеса для классификации новых экземпляров данных. Каждый экземпляр представляет собой набор значений атрибутов, описанных вектором $X=(x_1,x_2,\ldots,x_n)$. Учитывая M - классов, образец X присваивается классу C_i, тогда и только тогда, когда $P(X\mid C_i)P(C_i)>P(X\mid C_j)P(C_j)$ для всех $j=\overline{1,M}$ таких, что $i\neq j$. Образец принадлежит к классу с максимальной апостериорной вероятностью для образца. $P(X_k\mid C_i)$ рассчитывается как отношение значения частоты X_k для атрибута A_k к общему количеству образцов в обучающем наборе. В наивном байесовском подходе атрибуты, как предполагается, условно независимы. Несмотря на это предположение, наивные байесовские классификаторы дают удовлетворительные результаты, потому что акцент делается на идентификацию классов, а не точных вероятностей. Подход применяется для классификации спама и текста. Теоретически байесовские классификаторы менее склонны к ошибкам. С увеличением числа классов или атрибутов пространство и вычислительная сложность байесовских классификаторов экспоненциально возрастают.

г) Нечеткие множества

Нечеткие множества образуют ключевую методику представления и обработки неопределенности. Неопределенность возникает во многих формах в современных базах данных: неточность, неспецифичность, непоследовательность, расплывчатость и т.д. Нечеткие множества используют неуверенность в попытке сделать сложную систему управляемой. Таким образом, нечеткие множества представляют собой мощный подход не только к неполным, зашумленным или неточным данным, но также могут быть полезны при разработке неопределенных моделей данных, которые обеспечивают более высокую и точную производительность, чем традиционные системы.

Методы кластеризации

Кластеризация (обучение без учителя) может рассматриваться как обобщение классификации. Техника кластеризации может быть использована для обнаружения вторжений и неизвестных атак [4].

Известно много методов кластеризации. Выбор конкретного метода зависит от типа желаемого выхода. В общем, их можно разделить на две категории, основанные на структуре кластера, которые они образуют: иерархические и неиерархические.

1) Неиерархические методы

Неиерархические методы делят набор данных из N - объектов на M - кластеров, с перекрытием или без него. Каждый объект является членом кластера, к которому он наиболее близок. Однако порог сходства должен быть определен.

2) Иерархические методы

Иерархические методы образуют набор вложенных кластеров, каждая пара которых вкладывается в больший кластер. Процесс продолжается до тех пор, пока не останется только один кластер. Иерархические методы могут быть агломеративные и дивизимные.

В агломеративных методах [5] каждый объект соотносится с отдельным кластером. Кластеры объединяются до получения единственного кластера, включающего в себя все объекты.

В дивизимных методах [6] все рассматриваемые объекты определяются в один кластер на начальном этапе. На основании выбранной меры сходства выполняется разделение кластера до тех пор, пока каждый объект не будет определен в отдельный кластер. Примером является метод COBWEB.

а) Метод ближайшего соседа

Метод ближайшего соседа является одним из известных методов, относящихся к иерархическим агломеративным методам [7]. В данном методе информация о расстоянии между ближними, дальними и центральными объектами кластеров объединяется.

б) Минимальное покрывающее дерево

Данный метод выполняет кластеризацию объектов «сверху вниз» [8], в котором вначале все объекты определяются в единственный кластер. Кластеры разбиваются на два до тех пор, пока каждый объект не окажется в отдельном кластере. При этом происходит оценка расстояния между ними, которое должно быть максимальным.

в) Метод COBWEB

В этом методе реализуется вероятностное представление кластеров [5]. Объект, принадлежащий

кластеру, определяется с помощью вероятности появления конкретных значений параметра у объекта.

г) Метод k-means

Цель метода заключается в минимизации расстояний между объектами и центром кластера, к которому они относятся. Для метода k-means функцию минимизации расстояний можно представить [9]:

$$J = \sum_{k=1}^{M}\sum_{i=1}^{N} d^2(x_i, c_k),$$

где $x_i \in X$ – объект, а $c_j \in C$ – центр кластера, $|X| = N$, $|C| = M$.

д) Метод Fuzzy C-Means

Данный метод относится к нечетким методам, в которых делается предположение, что каждый элемент с определенной степенью принадлежит каждому кластеру [10]. Минимизация происходит на основе функции [10]:

$$J = \sum_{k=1}^{M}\sum_{i=1}^{N} u_{ij}^{m} d(x_i, c_k),$$

где $x_i \in X$ – объект, а $c_j \in C$ – центр кластера, m – мера нечеткости, $|X| = N$, $|C| = M$.

е) Метод EM

Данный метод основан на широко известном подходе максимизации ожиданий (Expectation Maximization), который предполагает, что кластеры формируются из результатов наблюдений, имеющих нормальное распределение [9]. Алгоритм состоит из двух основных шагов: E-шаг ожидания и M-шаг максимизации.

Ассоциативные правила

Типичным примером обучения без учителя являются ассоциативные правила. Это выражения вида $[X \to Y, c, s]$, где X, Y – подмножества множества значений на множестве атрибутов, $s = Support(X \bigcup Y)$, $c = \dfrac{Support(X \bigcup Y)}{Support(X)}$ – процент записей в базе, которые содержат $X \bigcup Y$ [3]. На этапе обучения добавляются новые правила при обнаружении аномалий. Обучение прекращается, пока по исходным данным уже нельзя будет получить новые данные.

Методы машинного обучения (Machine learning, ML) традиционно нашли применение в области информационной безопасности для обнаружения аномалий в работе информационной системы. Основная идея проста и применяется для различных видов информационных систем, таких как сети, операционные системы, прикладное программное обеспечение, и так далее [1]. Работа механизма обнаружения аномалии (AD) с использованием методов ML как правило, делится на две фазы: фаза обучения и фаза обнаружения. В фазе обучения механизм AD узнает нормальное поведение информационной системы. Результатом этапа обучения является множество профилей, характеризующих поведение системы. Профили используются для логического разделения поведения системы таким образом, чтобы поддерживать выделение аномалий на этапе обнаружения. Например, поведение операционной системы может быть смоделировано в виде набора профилей, где каждый профиль связан с действиями одного пользователя. Такая модель может быть использована для обнаружения аномальных действий пользователя. Кроме того, если количество пользователей больше и большинство пользователей демонстрируют аналогичное поведение, один профиль может быть связан с группой пользователей (например как роли), которые обладают подобными поведениями.

Информация в созданном профиле зависит от наблюдаемых системных событий, типа информации, собранной из этих событий, и от используемого алгоритма обучения. После того, как профили созданы, механизм AD работает в режиме обнаружения. В режиме обнаружения новое наблюдаемое событие сравнивается с созданными профилями для соответствия. Различные критерии, такие как статистические тесты отклонения, тесты обнаружения выбросов, апостериорные вероятности и т.д., могут быть использованы для определения соответствия события к созданным профилям [1]. Если событие не соответствует профилям (на основе некоторого порога), то оно классифицируется как аномальное. В противном случае информация от событий либо объединяется в профили, либо отбрасывается. Система AD имеет преимущество над системой на основе сигнатур в том, что она может обнаружить потенциально новые атаки, в то время как система на основе сигнатур может работать только для тех атак, для которых были созданы подписи нападений. Недостатком системы AD является число ложных тревог, которые она выдает, когда профили, созданные на этапе обучения, не полностью соответствуют нормальному поведению.

Выделяют ошибки следующих видов: а) нормальное поведение системы или пользователя ошибочно принимается за злоумышленное (false positives); б) попытка злоумышленного проникновения в систему принимается за нормальную активность (false negatives). Более опасной является вторая ситуация [11].

III. ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ К ПРОБЛЕМАМ ИБ

Одна из проблем связана с парадигмой обнаружения аномалий в сети. Некоторые авторы [12, 13] ставят под сомнение применение модели Деннинга для обнаружения аномалий в одной системе, предложенной в 1987 году [14] в неизменном виде к обнаружению аномалий в сети, и предлагают заменить его новой парадигмой в свете изменений в операционной среде.

К системам IDS (Intrusion detection system) можно предъявить различные требования. Важным требованием

является то, что система должна обнаружить значительный процент вторжений, при этом сохраняя частоту ложных тревог на приемлемом уровне. Работа демонстрирует, что для разумного набора предположений частота ложных тревог является ограничивающим фактором для производительности системы IDS. Это связано с явлением base-rate fallacy, которое применительно к системам IDS означает, что для достижения существенных значений уровня обнаружения P(Intrusion|Alarm), требуется обеспечить низкий уровень ложных тревог (возможно, в некоторых случаях недостижимо). Анализ отчетов производительности IDS показывает, что, по крайней мере для некоторых видов обнаружения вторжений, достичь таких низких частот ложных тревог является проблематичным [13].

Одной из больших проблем является то, что злонамеренная активность находится среди терабайтов нормальных данных как иголка в стоге сена, в результате чего методы из учебников по машинному обучению, часто не предназначенных для таких сценариев, производят массу ложных срабатываний. Еще одной проблемой является неизбежное отсутствие доступа к данным атак нулевого дня для тренировки методов машинного обучения.

Хотя применение машинного обучения для ИБ имеет аналоги использования в других областях, оно также имеет существенные различия. Главным различием применения машинного обучения в области ИБ является то, что в других областях никто не пытается обмануть систему. В случае применения машинного обучения для обнаружения атак или вредоносных программ противник постоянно развивает свои технологии обхода системы безопасности. Согласно Saxe [15], точность систем, которые вначале способны обнаружить более 80% атак, в течение двух лет может снизиться до 60% или меньше из-за эволюции вредоносных программ или тактики атак.

ПРОБЛЕМЫ МАШИННОГО ОБУЧЕНИЯ ВО ВРАЖДЕБНОЙ СРЕДЕ

Исследование методов машинного обучения во враждебной среде (adversarial machine learning) находится на пересечении машинного обучения и ИБ [16, 17]. Оно направлено на безопасное применение методов машинного обучения во враждебной среде, таких, как фильтрация спама, обнаружение вредоносных программ и биометрическое распознавание.

Проблема возникает из-за того, что методы машинного обучения были изначально разработаны для стационарных сред, в которых предполагается, что данные для обучения и тестирования генерируются из одного и того же распределения. В присутствии интеллектуальных и адаптивных противников, однако, эта гипотеза, вероятно, будет нарушена, по крайней мере в некоторой степени (в зависимости от противника). На самом деле злонамеренный противник может тщательно манипулировать входными данными, эксплуатируя конкретные уязвимости алгоритмов обучения для компрометации всей системы безопасности.

Чтобы понять свойства безопасности алгоритмов обучения во враждебной среде, следует рассмотреть следующие основные вопросы [18–20]:

- идентифицировать потенциальные уязвимости алгоритмов машинного обучения в процессе обучения и классификации;
- разработать соответствующие атаки, которые соответствуют идентифицированным угрозам, и оценить их влияние на целевую систему;
- предложить контрмеры для улучшения безопасности алгоритмов машинного обучения против рассматриваемых атак.

В литературе встречается описание следующих атак на методы машинного обучения.

Атаки уклонения

Атаки уклонения [21] являются наиболее распространенными типами атак на алгоритмы машинного обучения во враждебной среде. Ярким примером атак уклонения является спам на основе изображения, в котором содержание спама встраивается в прилагаемое изображение для уклонения от текстового анализа, выполняемого антиспамфильтром. Другим примером атак уклонения является спуфинг против биометрических систем распознавания [22, 23].

Атаки отравления

Алгоритмы машинного обучения часто повторно обучаются на данных, собранных во время работы. Например, системы обнаружения вторжений (IDS) часто переобучаются на множестве образцов, собранных во время работы в сети. В этом случае злоумышленник может отравить обучающие данные путем введения тщательно разработанных образцов для компрометации процесса обучения. Примеры атак отравления алгоритмов машинного обучения могут быть найдены в [18, 19, 23, 24].

Атаки на алгоритмы кластеризации

Алгоритмы кластеризации применяются в системах безопасности для обнаружения опасных или незаконных действий. Например, кластеризация вредоносных программ и компьютерных вирусов направлена на выявление и классификацию различных семей существующих вредоносных программ, и генерацию конкретных сигнатур для их обнаружения антивирусами или на основе сигнатур систем обнаружения вторжений как Snort. Алгоритмы кластеризации не были изначально разработаны для борьбы с попытками преднамеренных атак, которые предназначены для подрыва самого процесса кластеризации. Можно ли безопасно применять кластеризацию в таких условиях, остается пока невыясненным. Предварительные работы, которые сообщают о некоторых уязвимостях кластеризации, можно найти в [25, 26].

Отметим, что встречаются также примеры атак на методы распознавания, не применяемые непосредственно в системах информационной безопасности. Например, в [27] доказывается, что для каждой обученной нейронной

сети можно найти такие примеры входных данных, которые будут давать на выходе отрицательный ответ, в то время как эти примеры данных ничтожно отличаются от примеров, дающих положительный ответ.

В области машинного обучения во враждебной среде были предложены ряд защитных механизмов против атак уклонения, атак отравления и атак на конфиденциальности, в том числе:

- определение безопасных алгоритмов обучения [28];
- использование нескольких систем классификаторов [29, 30];
- использование рандомизации или дезинформации для заблуждения злоумышленника при приобретении знаний о системе [30];
- разработка методов обучения, сохраняющих конфиденциальность [31].

ЗАКЛЮЧЕНИЕ

Машинное обучение, визуализация данных и технологии масштабируемого хранения данных создают платформу для улучшения состояния ИБ. Машинное обучение позволяет обнаружить полезные корреляции и разработать эффективные правила обнаружения атак. Машинное обучение также поможет системе обнаружить не замеченные ранее признаки атак. Ранние прототипы показали отличные результаты и доказали, что системы машинного обучения прекрасно справляются с уточнением правил обнаружения атак.

ЛИТЕРАТУРА

[1] M.I.Jordan, T.M.Mitchell, Machine learning: Trends, perspectives, and prospects, Science, Vol. 349, No. 6245, pp. 255–260, 2015. DOI:10.1126/science.aaa8415

[2] N.Idika, A.P.Mathur, A Survey of Malware Detection Techniques, Technical report, Software Engineering Research Center, 2007.

[3] D.Barbara, S.Jajodia, Applications of Data Mining in Computer Security, Kluwer, 2002.

[4] S.Dua, X.Du, “Data Mining and Machine Learning in Cybersecurity, Auerbach Publications, 2011.

[5] D.H.Fisher, Knowledge acquisition via incremental conceptual clustering, Machine Learning, Vol. 2, No. 2, pp. 139–172, 1987.

[6] Д.А.Вятченин, “Нечеткие методы автоматической классификации: Монография,” Мн.: УП «Технопринт», 2004.

[7] M.A.Maloof, "Machine Learning and Data Mining for Computer Security: Methods and Applications, Springer-Verlag: New York, 2005.

[8] Н.Г.Загоруйко, “Прикладные методы анализа данных и знаний”, Новосибирск: ИМ СО РАН, 1999.

[9] A.K.Jain, M.N.Murty, and P.J.Flynn, Data Clustering, ACM Computing Surveys, Vol. 31, pp. 264–323, 1999.

[10] В.П.Боровиков, STATISTICA. Искусство анализа данных на компьютере: Для профессионалов”, СПб.: Питер, 2003.

[11] S.Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In ACM Conference on Computer and Communications Security, pp.1–7, 1999.

[12] C.Gates, C.Taylor, Challenging the Anomaly Detection Paradigm: A Provocative Discussion, Proceedings of the Workshop on New Security Paradigms, pp. 21–29, 2007.

[13] S.Axelsson, The base-rate fallacy and the difficulty of intrusion detection, ACM Transactions on Information and System Security, Vol. 3, No. 3, pp.186–205, Aug. 2000 [doi>10.1145/357830.357849]

[14] D.E.Denning, An Intrusion-Detection Model, IEEE Transactions on Software Engineering - Special issue on computer security and privacy, Vol. 13, Issue 2, February 1987, pp. 222–232.

[15] J.Saxe, Why Security Data Science Matters and How Its Different: Pitfalls and Promises of Data Science Based Breach Detection and Threat Intelligence. BlackHat USA 2015 Briefings. https://www.blackhat.com/us-15/speakers/Joshua-Saxe.html

[16] P.Laskov, R.Lippmann, Machine Learning in Adversarial Environments, Machine Learning Journal, 81, 2010.

[17] A.D.Joseph, P.Laskov, F.Roli, J. Doug Tygar, and B. Nelson (eds) Machine Learning Methods for Computer Security Report from Dagstuhl Perpectives Workshop 12371.

[18] B. Biggio, G. Fumera, and F. Roli. "Security evaluation of pattern classifiers under attack". IEEE Transactions on Knowledge and Data Engineering, 26(4):984–996, 2014.

[19] B.Biggio, I.Corona, B.Nelson, B. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli. Security evaluation of support vector machines in adversarial environments. In Y.Ma and G.Guo (eds), Support Vector Machines Applications, pp. 105–153. Springer, 2014.

[20] B.Biggio, G.Fumera, and F.Roli, “Pattern recognition systems under attack: Design issues and research challenges, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 28, No. 7, 1460002, 2014. DOI: 10.1142/S0218001414600027.

[21] B. Biggio, I.Corona, D.Maiorca, B.Nelson, N.Srndic, P.Laskov, G.Giacinto, and F.Roli. Evasion attacks against machine learning at test time. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Part III, LNCS Vol. 8190, pp. 387– 402. Springer Berlin Heidelberg, 2013.

[22] R.N.Rodrigues, L.L.Ling, and V.Govindaraju. Robustness of multimodal biometric fusion methods against spoof attacks, Journal of Visual Languages and Computing, Vol. 20, No. 3, pp. 169–179, 2009.

[23] B.Biggio, L.Didaci, G.Fumera, and F.Roli, Poisoning attacks to compromise face templates, In 6th IAPR Int’l Conf. on Biometrics (ICB 2013), pp. 1–7, 2013.

[24] B.Biggio, B.Nelson, and P.Laskov, Poisoning attacks against support vector machines. Proc. of the 29th Int’l Conf. on Machine Learning (ICML), 2012. arxiv.org/vc/arxiv/papers/1206/1206.6389v1.pdf

[25] B.Biggio, I.Pillai, S. R. Bulò, D.Ariu, M.Pelillo, and F.Roli. Is data clustering in adversarial settings secure? Proc. of the ACM Workshop on Artificial Intelligence and Security (AISec’13), pp. 87-98, 2013.

[26] D.B.Skillicorn. Adversarial knowledge discovery, IEEE Intelligent Systems, Vol. 24, pp. 54–61, 2009.

[27] C.Szegedy, W.Zaremba, I.Sutskever, J.Bruna, D.Erhan, I.Goodfellow, R.Fergus, Intriguing properties of neural networks, Proc. of the International Conference on Learning Representations (ICLR), 2014. http://arxiv.org/abs/1312.6199

[28] O.Dekel, O.Shamir, L.Xiao, Learning to classify with missing and corrupted features, Machine Learning, Vol.81, No.2, pp. 149–178, 2010.

[29] B.Biggio, G.Fumera, and F.Roli. Multiple classifier systems for robust classifier design in adversarial environments. International Journal of Machine Learning and Cybernetics, Vol. 1, No. 1, pp. 27–41, 2010.

[30] B.Biggio, G.Fumera, and F.Roli, Adversarial pattern classification using multiple classifiers and randomisation. In 12th Joint IAPR International Workshop on Structural and Syntactic Pattern Recognition (SSPR 2008), LNCS Vol. 5342, pp. 500–509, 2008.

[31] B.I.P. Rubinstein, P.L.Bartlett, L. Huang, and N.Taft. Learning in a large function space: Privacy- preserving mechanisms for svm learning. Journal of Privacy and Confidentiality, Vol. 4, No. 1, pp. 65–100, 2012.