Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра интеллектуальных информационных технологий

Методы анализа поведения пользователя в задаче обнаружения внутренних вторжений с помощью нейросетей

Александров Валентин Валерьевич, 620 группа
Научный руководитель: к.ф.-м.н. Царёв Дмитрий Владимирович

Москва 2020

Содержание

1 Введение 3
1.1 Обзор моделей 5
1.2 Выводы по обзору 9
2 Постановка задачи 11
3 План решения задачи 12
Список использованных источников 13

1 Введение

Внутренние (инсайдерские) вторжения — это вредоносные для организации угрозы, которые исходят от людей внутри организации, таких как работники, бывшие работники, подрядчики или деловые партнеры, имеющих доступ к информации о методах безопасности внутри организации, данных и компьютерных системах.

Инсайдерские угрозы в данный момент является серьезной проблемой для многих организаций. Их обнаружение является очень трудной задачей в силу того, что инсайдеры для злонамеренных действий используют доверенный им доступ, из-за чего они могут легко обходить системы внешней защиты информации.

Существует ряд технологий, для решения проблем инсайдерских угроз:

— DLP (Data Loss Prevention) – предотвращение утечек с помощью анализа потоков данных, пересекающих периметр информационной системы

— SIEM (Security Information and Event Management) – анализ в реальном времени событий безопасности

— IAM (Identity and Access Management) – управление учетными данными пользователей, системами контроля и управления доступом

— UEBA (User and Entity Behavior Analytics) – анализ действий пользователя и принятие решения на основе исторических данных

DLP, SIEM и IAM системы способны бороться с инсайдерами только на самом последнем этапе утечки данных, когда инсайдер пытается провести эксфильтрацию данных. При этом, до этого момента инсайдер проходит стадии подготовки утечки, которая может длиться месяцами и в течение этого этапа инсайдер проявляет аномальное поведение. Поэтому существует отдельный класс решений UEBA, который направлен на анализ поведения пользователей и способны обнаруживать ранние признаки утечки информации.

Под поведенческой информацией понимается совокупность данных, которые описывают:

— Контекст – структурированные данные, описывающие атрибуты операций, которые пользователь выполняет с документами;

— Контент – неструктурированные данные в виде содержимого документов, ассоциированные с операциями пользователя.

Актуальность задачи

По отчетам Ponemon за 2020 год [1] стоимость ущерба от инсайдерских угроз составляет 11,45 миллионов долларов в среднем за атаку. При этом средняя стоимость атак выросла на 31% в два раза с 2018 года. По тому же отчету, частота инсайдерских инцидентов также заметно выросла. Также приводится, что из всех методов борьбы с инсайдерской угрозой UEBA системы лидируют по степени сокращения финансового ущерба. В среднем они сокращают 3,42 миллиона долларов. Это как раз связано с тем, что UEBA системы способны обнаруживать аномальное поведение пользователя на ранних этапах утечки информации. Это подтверждается в [2], где подчеркивается, что в силу того, что инсайдеры имеют мало препятствий, время с первых несанкционированных действий до обнаружения часто занимает месяцы и годы.

Согласно McKinsey [3] из всех опубликованных в VERIS Community Database публичных случаях утечки информации за период с 2012 по 2017 год, в 50% присутствовал элемент инсайдерской угрозы. При этом, 38% случаев из этих наличествует злонамеренные действия инсайдеров. В отчете Verizon за 2019 год [2] приводится, что 57% взломов базы данных организации включают в себя инсайдерские угрозы.

Отчеты Haystax [4][5] и Gartner [6] подтверждают растущие опасения организаций по этому поводу и тренд на рост частоты атак инсайдеров.

Согласно отчету Gartner [7] рынок самостоятельных UEBA-систем умирает, но в данный момент очень востребованы SIEM-системы, которые включают в себя UEBA.

По приведенным выше отчетам становится очевидной актуальность изучения и разработки моделей машинного обучения, способные определять ранние признаки аномального поведения корпоративных пользователей.

Существующие UEBA-решения нацелены в первую очередь на анализ структурированной контекстной информации о поведении пользователя которые включают в себя обычно данные об операциях с файлами, электронной почты, подключении новых устройств и данные из системного журнала ОС. Это объ-

ясняется тем, что анализ контентной информации намного более затруднителен из-за своей неструктурированности и очень большого объёма. В то же время, анализ контента позволяет выявлять случаи, при которых поведения пользователя остаётся таким же, но меняется содержимое файлов, с которыми он работает. Анализ только структурированной информации такие случаи выявить не может, что позволяет утверждать о ценности контента для рассматриваемой задачи. Актуальность анализа текстовых данных также подвтерждается отчётом Gartner [7].

Рассмотрим существующие методы построения моделей машинного обучения для задачи автоматического обнаружения инсайдерских угроз на основе поведенческих данных. Также рассмотрим доступные публичные наборы данных, которые потенциально могут быть использованы для обучения моделей.

1.1 Обзор моделей

Синтетический набор данных разработанный подразделением CERT Carnegie Melon University пользуется очень большой популярностью в исследованиях на данную тему. Поэтому, по умолчанию, во всех приведенных статьях ниже используется набор данных CERT версии 4.2. Также по умолчанию содержимое писем, файлов и вебсайтов из данного набора данных не используется.

В работе [8] используется скрытая марковская модель (Hidden Markov Model - **HMM**). Скрытая марковская модель представляет собой обычную марковскую модель, в которой модель выводит некоторый символ каждый раз перед переходом в следующее состояние. Модель называется скрытой, поскольку мы наблюдаем только выходы модели, а последовательность переходов состояний скрыта от нас.
В данной работе все действия пользователей кодируются числами. Затем действия собираются отдельно для каждого пользователя и сортируются по времени. На первых пяти неделях модель обучается (используется предположение, что за это время инсайдерского поведения не было), для последующих недель модель предсказывает вероятность аномальности поведения и только потом обучается. Поведение за неделю считается аномальным, если оно превышает заданный порог.
Эксперименты на CERT 4.2 после подбора гиперпараметров показали AUC ROC 0.83.

В [9] также применяется метод обучения без учителя и задача обнаружения инсайдерских угроз ставится как задача обнаружения аномалий. В этой работе сравниваются два классических метода нахождения аномалий: Isolation Forest и One Class SVM. Работа проводилась на наборе данных CERT. Моделям на вход подавались данные агрегированные по разным временным периодам - дням, месяцам,

полугодиям и годам. Также в качестве дополнительного признака используется trust score (показатель доверия), который означает оценку аномальности пользователя для предыдущего периода. Авторы показывают, что этот признак заметно улучшает точность предсказаний моделей, особенно в случаях моделей, в которых данные агрегируются по малым периодам.

В статье [10] применяется рекуррентная нейронная сеть LSTM для поиска аномалий в поведении пользователей. В работе проведены эксперименты, которые показывают, что LSTM действительно обучается паттернам поведения пользователей и находить аномальное поведение. Наилучший результат при подборе гиперпараметров: точность - 0.84 и полнота - 0.60.
В статье приводится предположение, что некоторые случаи, в которых модель ошибочно не распознала аномальное поведение, можно разрешить с помощью анализа контента.

В [11] было испробовано множество различных семейств алгоритмов машинного обучения. В этой статье приводится к заключению, что алгоритмы случайного леса и бустинга показывают наилучшие результаты. Важно отметить, что они в качестве признаков также использовали сентимент-анализ текста электронной почты и содержания посещенных сайтов.

При применении моделей машинного обучения качество моделей очень сильно зависит от признаков, которые были вручную сгенерированы исследователями. Однако в последнее время наблюдается очень большой интерес к нейронным сетям, в том числе из-за того, что они способны автоматически выучивать хорошее признаковое представление данных. Поэтому и в данной теме множество последних работ использует нейросети.

В работе [12] используется информация о социальном графе для нахождения аномалий в нем и поведенческая информация пользователей для психологического профилирования. Интересно отметить, что в данной работе в качестве данных использовались данные из популярной онлайн многопользовательской игры World of Warcraft(WoW). Они собрали данные о социальном графе игроков внутри игры и его изменениях за шесть месяцев. Авторы с помощью своего методы пытались предсказать то, что игрок в скором времени покинет гильдию (социальную группу внутри игры). В пользу необычного выбора набора данных приводится, что данных много, содержат в себе зловредные поведения, публичны и не ограничены правилами конфиденциальности. Авторы утверждают, что предложенный ими метод можно применять также и на реальных предприятиях.

В статье [13] используется следующий подход: сначала LSTM выучивает поведение пользователя по его действиям и извлекает временные признаки, затем извлеченные признаки подаются на вход CNN классификатору.

Сверточные нейронные сети CNN - специальная разновидность архитектур нейронных сетей, в которой слои, выполняющие свертку, чередуются со слоями субдискретизации. На данный момент, CNN является одним из лучших алгоритмов по распознаванию и классификации изображений.

Поведения пользователя рассматривается как последовательность действий и действия одного пользователя соответствует одному "предложению как в обычных NLP задачах. Все действия пользователя перед подачей на вход модели one-hot кодируются. После подачи каждого действия каждый скрытый слой LSTM выдает вектор, который выражает текущее состояние сети в пространстве малой размерности. Выходы последнего слоя собираются в одну матрицу, которая затем приводится к фиксированному размеру с помощью отбрасывания лишних векторов в случае длинных последовательностей действий и заполнения нулями в случае коротких последовательностей. Полученная матрица подается на вход CNN сети, которая в свою очередь предсказывает аномальность поведения. Авторы пишут, что их метод показал $AUCROC = 0.9449$ на отложенной выборке CERT.

В другой работе [14] используется обратный подход. В ней CNN с одномерными сверточными слоями сначала пытается извлечь признаки, затем они подаются на вход LSTM для классификации. Значимое отличие от предыдущей работы заключается в том, что отображение поведения пользователей происходит с помощью представления каждого отдельного действия вектором малой размерностью. Значения этого вектора характеризуются соседними действиями, которые обычно встречаются вместе с ним. Авторы не указывают точно какой метод они использовали, но по описанию это очень похоже на популярный подход word2vec [15]. Также авторы использовали технику SMOTE [16] для семплирования объектов малого класса и исправления сильного дисбаланса классов в наборе данных. SMOTE генерирует синтетические данные, которые похожи на k ближайших соседей малого класса.

Обе работы [14] и [13] ставят свою задачу как бинарную классификацию, в которой алгоритму необходимо определить аномальность поведения пользователя за некоторый промежуток времени (в обеих статьях рассматривают данные по дням)

В работе [17] исследуется применение механизма **Attention** (внимание) для задачи обнаружения инсайдеров. Этот механизм позволяет сети обращать особое внимание для некоторых важных действий. Attention-слой в данной сети собирают

выходы LSTM-слоя в один вектор, присваивая различный вес каждому выходу в зависимости от его важности. Как показали эксперименты в [17], добавление Attention увеличивает AUC ROC для LSTM и RNN сетей при использовании набора данных CERT.

Для поиска аномалий также возможно применение нейронных сетей из области распознавания изображений. Вдохновленные недавними успехами в применении нейросетей для анализа изображений для классификации вредоносных программ, [18] применили этот подход для задачи обнаружении инсайдерской угрозы. В этой работе из набора данных CERT было вручную отобрано 20 признаков, и для каждого пользователя отдельно по этим признаком составлялись изображения 32 на 32 пикселей, которые подавались предобученной на ImageNet популярным нейросетевым моделям VGG16 и MobileNet. В результате было получено 99.32 точность и полнота на отложенной выборке.

Также, как показано в статье [17], техника attention в применении к поведенческой информации пользователя, способна значительно улучшить качество работы модели, позволяя модели давать разный вес элементам последовательности в зависимости от важности этого элемента. Это открывает потенциал для работы с семейством моделей BERT [19], чей успех в NLP задачах приписывается слоям-трансформерам, которые, в свою очередь, используют attention. Поэтому с помощью модели BERT можно анализировать последовательность действий пользователя.

В [20] представлен подход представления данных для последующего анализа, который заключаются в использовании как *последовательных* данных (последовательность действий конкретного пользователя за некоторый период времени), так и *численных* данных. Численные данные в данной работе делятся на пользовательские и данные о действиях. Пользовательские данные в наборе данных CERT представлены информацией о роли пользователя в подразделении, его отделе и психометрике. Данные о действиях получаются подсчетом количества совершенных действий одной категории за рассматриваемый промежуток времени. Контентные данные набора данных CERT не были рассмотрены по силу синтетической природы набора данных.

Архитектуры рекуррентных сетей изначально предполагают работу только с последовательными данными. Чтобы совместить последовательные и численные данные можно использовать подход предложенный в [21]. В данной статье решалась задача автоматической аннотации изображений, которая состоит в выделением некоторых участков изображения с различными объектами и генерация текстового описания объеков в этих участках. Для генерации текста использовалась архитекту-

ра RNN, в которой начальное состояние скрытых слоев зависит от информационного вектора о соответствующем участке изображения. Таким образом, все состояния рекуррентной сети обусловлены содержанием изображения. Этот подход успешно решал поставленную задачу. Нетрудно расширить этот подход для произвольного числового вектора $\overrightarrow{x}$, который также называется *условным*. Мы можем преобразовать условный вектор так, чтобы он совпадал по размеру с вектором скрытого состояния рекуррентной сети $\overrightarrow{h}$. Для этого достаточно применить следующее простое преобразование:

$$\overrightarrow{h_0} = \mathbf{W}\overrightarrow{x} + \overrightarrow{b}$$

где $\mathbf{W}$ и $\overrightarrow{b}$ являются обучаемыми параметрами. Затем, полученным вектором можно инициализировать скрытое состояние рекуррентой сети. Процесс повторяется для каждой обрабатываемой рекуррентной сетью последовательностью.

В статье [22] на наборе данных сравниваются модели с различными рассматриваемыми периодами для каждого пользователя: день, неделя и пользовательская сессия. Для данного периода собирались числовые данные о частоте событий и их статистические признаки, такие как среднее, медиана и стандартное отклонение. В результате исследования модели, которые были обучены на полных пользовательских сессиях дали лучший результат.

LSTM также используется в Unsupervised-режиме в работе [23]. В данной работе рассматривается обучение LSTM-автокодировщика. Также авторы в данной работе разбивают всех пользователей в наборе данных CERT по 8 непересекающимся сообществам. Авторы предлагают находить потенциальных инсайдеров по тому, насколько они выделяются внутри своего сообщества. Эксперименты показали, что таким способом можно успешно найти всех инсайдеров среди пяти наиболее аномальных пользователей из каждого сообщества. Эксперименты проводились на наборе данных CERT v6.2.

В работе [24] используется похожий Unsupervised-подход с LSTM. Модель была обучена на задаче предсказания следующего действия в последовательности. Для определения аномальности используется отклонения действий пользователя от предсказанных моделью. Авторы называют главными преумествами своего подхода интерпретируемость результатов модели и возможность онлайн-обучения.

1.2 Выводы по обзору

Обзор литературы по существующим решениям показал, что тема обнаружения инсайдерских угроз является актуальной, и в данный момент привлекает большой интерес исследователей. Синтетический набор данных CERT является текущим де-фактом стандартом для обучения и валидации моделей. По обзору моделей стано-

вится понятно, что тренд на популярность нейросетевых моделей не обошел и задачу обнаружения инсайдерских угроз. Существуют способы, которые ставят задачу как поиск аномалий (метод обучения без учителя) и как классическую классификацию (метод обучения с учителем). Исходя из изученных работ, задача бинарной классификации, основанная на наборе данных CERT, является наиболее актуальной. Неформально эта задача представляет собой определить аномальность поведения пользователя за некоторый промежуток времени (день, неделя или пользовательская сессия).

2 Постановка задачи

Исследование и разработка методов обнаружения аномального поведения пользователей с использованием нейросетей на основе собираемой поведенческой информации следующих типов:

а) контекст пользовательских операций

б) контент файлов

3 План решения задачи

а) Воспрозвести результаты статьи [13], так как эта работа имеет наилучшие представленные результаты на наборе данных CERT. На данный момент статья уже воспроизведена, для улучшения качества классификации были подобраны гиперпараметры, добавлены слои BatchNormalization [25] и Dropout [26], добавлена ребалансировка весов в функции потерь. На основе полученной модели проводятся все последующие эксперименты.

б) Так как все современные работы не используют контентные данные набора данных CERT, следует провести эксперименты с ними. Этот этап выполнен. Для обработки контентных данных использовалась обработка алгоритмом LDA для тематического представления текстов. Добавление контентных данных не улучшило качество классификации.

в) Добавить агрегированные статистики по последовательностям, как описано в [20]. Сложность данного этапа заключается в том, что для обучения LSTM-модели невозможно напрямую добавить такой вид данных. Чтобы обойти это, можно использовать подход, описанный в [21], который заключается в инициализации скрытого состояния LSTM на основе произвольного входного вектора.

г) Добавить Attention-слой в LSTM-кодировщик. Как показано в [13], это позволит улучшить качество классификации. Необходимо провести эксперимент, сможет ли это улучшить в случае использования сверточного классификатора

д) Использовать BERT [19]. На данный момент модели из семейства BERT дают наилучшее качество практически во всех задачах обработки естественных языков. Можно попробовать адаптировать BERT для анализа последовательностей пользовательских действий.

Список использованных источников

1. Ponemon Report: 2018 Cost of Insider Threats – Global Organizations. — 2018. — apr. https://www.observeit.com/ponemon-report-cost-of-insider-threats/.

2. 2019 Insider Threat Report. https://enterprise.verizon.com/resources/reports/insider-threat-report/.

3. Insider threat: The human element of cyberrisk | McKinsey. https://www.mckinsey.com/business-functions/risk/our-insights/insider-threat-the-human-element-of-cyberrisk.

4. *Veriato.* Insider Threat Report 2018. https://www.veriato.com/resources/whitepapers/insider-threat-report-2018.

5. *Company, Haystax A. Fishtech Group.* Insider Threat Report - 2019. https://info.haystax.com/insider-threat-report-2019.

6. Emerging Insider Threat Detection Solutions. — 2018. — apr. https://blogs.gartner.com/avivah-litan/2018/04/05/insider-threat-detection-replaces-dying-dlp/.

7. Gartner Report: "Market Trends: UEBA Providers Must Embrace Enterprise Specialization". — 2019. — jun. https://www.observeit.com/gartner-report-market-trends-ueba-providers-must-embrace-enterprise-specializat

8. *Rashid, Tabish.* A New Take on Detecting Insider Threats: Exploring the Use of Hidden Markov Models / Tabish Rashid, Ioannis Agrafiotis, Jason R.C. Nurse // Proceedings of the 2016 International Workshop on Managing Insider Security Threats - MIST '16. — Vienna, Austria: ACM Press, 2016. — Pp. 47–56. http://dl.acm.org/citation.cfm?doid=2995959.2995964.

9. *Aldairi, Maryam.* A Trust Aware Unsupervised Learning Approach for Insider Threat Detection / Maryam Aldairi, Leila Karimi, James Joshi. — 2019. — jul. — Pp. 89–98.

10. *Lu, Jiuming.* Insider Threat Detection with Long Short-Term Memory / Jiuming Lu, Raymond K. Wong // Proceedings of the Australasian Computer Science Week Multiconference on - ACSW 2019. — Sydney, NSW, Australia: ACM Press, 2019. — Pp. 1–10. http://dl.acm.org/citation.cfm?doid=3290688.3290692.

11. *Noever, David.* Classifier Suites for Insider Threat Detection / David Noever // *arXiv:1901.10948 [cs, stat].* — 2019. — jan. — arXiv: 1901.10948. http://arxiv.org/abs/1901.10948.

12. Proactive Insider Threat Detection through Graph Learning and Psychological Context / Oliver Brdiczka, Juan Liu, Bob Price et al. // 2012 IEEE Symposium on Security and Privacy Workshops. — 2012. — may. — Pp. 142–149. — ISSN: null.

13. Insider Threat Detection with Deep Neural Network / Fangfang Yuan, Yanan Cao, Yanmin Shang et al. // Computational Science – ICCS 2018 / Ed. by Yong Shi, Haohuan Fu, Yingjie Tian et al. — Cham: Springer International Publishing, 2018. — Vol. 10860. — Pp. 43–54. http://link.springer.com/10.1007/978-3-319-93698-7_4.

14. Insider Threats Detection using CNN-LSTM Model / Ahmed Saaudi, Zaid Al-Ibadi, Yan Tong, Csilla Farkas. — 2019. — apr.

15. Efficient Estimation of Word Representations in Vector Space / Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean // *arXiv:1301.3781 [cs]*. — 2013. — sep. — arXiv: 1301.3781. http://arxiv.org/abs/1301.3781.

16. SMOTE: Synthetic Minority Over-sampling Technique / N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer // *Journal of Artificial Intelligence Research*. — 2002. — jun. — Vol. 16. — Pp. 321–357. https://www.jair.org/index.php/jair/article/view/10302.

17. Attention-Based LSTM for Insider Threat Detection / Fangfang Yuan, Yanmin Shang, Yanbing Liu et al. — 2019. — nov. — Pp. 192–201.

18. *G, Gayathri R.* Image-Based Feature Representation for Insider Threat Classification / Gayathri R. G, Atul Sajjanhar, Yong Xiang // *arXiv:1911.05879 [cs]*. — 2019. — nov. — arXiv: 1911.05879. http://arxiv.org/abs/1911.05879.

19. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // *arXiv:1810.04805 [cs]*. — 2019. — may. — arXiv: 1810.04805. http://arxiv.org/abs/1810.04805.

20. *Le, Duc C.* Evaluating Insider Threat Detection Workflow Using Supervised and Unsupervised Learning / Duc C. Le, A. Nur Zincir-Heywood // 2018 IEEE Security and Privacy Workshops (SPW). — San Francisco, CA: IEEE, 2018. — may. — Pp. 270–275. https://ieeexplore.ieee.org/document/8424659/.

21. *Karpathy, Andrej.* Deep Visual-Semantic Alignments for Generating Image Descriptions / Andrej Karpathy, Li Fei-Fei // *arXiv:1412.2306 [cs]*. — 2015. — apr. — arXiv: 1412.2306. http://arxiv.org/abs/1412.2306.

22. *Le, Duc C.* Analyzing Data Granularity Levels for Insider Threat Detection Using Machine Learning / Duc C. Le, Nur Zincir-Heywood, Malcolm I. Heywood // *IEEE Transactions on Network and Service Management*. — 2020. — mar. — Vol. 17, no. 1. — Pp. 30–44. https://ieeexplore.ieee.org/document/8962316/.

23. *Paul, Sudipta.* LAC : LSTM AUTOENCODER with Community for Insider Threat Detection / Sudipta Paul, Subhankar Mishra. — P. 9.

24. Deep Learning for Unsupervised Insider Threat Detection in Structured Cybersecurity Data Streams / Aaron Tuor, Samuel Kaplan, Brian Hutchinson at al. // *arXiv:1710.00811 [cs, stat]*. — 2017. — dec. — arXiv: 1710.00811. http:

//arxiv.org/abs/1710.00811.

25. *Ioffe, Sergey.* Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift / Sergey Ioffe, Christian Szegedy // *arXiv:1502.03167 [cs].* — 2015. — mar. — arXiv: 1502.03167. http://arxiv.org/abs/1502.03167.

26. Dropout: A Simple Way to Prevent Neural Networks from Over[U+FB01]tting / Nitish Srivastava, Geo[U+FB00]rey Hinton, Alex Krizhevsky at al. — P. 30.