

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ИСТОЧНИКОВ ЦИФРОВОГО СЛЕДА ДЛЯ СИСТЕМ КЛАССА UEBA С КОЛИЧЕСТВЕННОЙ ОЦЕНКОЙ ИНФОРМАТИВНОСТИ И ВЫЧИСЛИТЕЛЬНОЙ СТОИМОСТИ

В.В. Мальцев¹, Г.И. Будовская¹

¹ФГБОУ ВО «Воронежский государственный лесотехнический университет
имени Г.Ф. Морозова»

Аннотация. Актуальность проблемы обнаружения инсайдерских угроз обусловлена ростом числа атак со стороны авторизованных пользователей, которые сложно выявлять традиционными средствами защиты. Системы класса User and Entity Behavior Analytics предлагают решение на основе анализа цифрового следа сотрудников, однако отсутствует методика выбора источников данных, обеспечивающая баланс между полнотой детектирования и вычислительными затратами. В статье представлена классификация источников цифрового следа, включающая пять групп: метаданные доступа, метаданные действий, сетевые потоки, контент и поведенческую биометрию. Предложен показатель – индекс информативности к сложности извлечения, позволяющий количественно оценить целесообразность использования каждого типа данных. Эксперименты на датасете CMU-CERT r4.2 показали, что максимальный индекс (3,56) достигается при использовании сетевых потоков, тогда как анализ контента писем дает прирост качества детектирования 8,3%, но характеризуется низким индексом (0,41) из-за высокой вычислительной стоимости.

Ключевые слова: UEBA, цифровой след, инсайдерские угрозы, анализ поведения, машинное обучение, информативность признаков, CMU-CERT.

COMPARATIVE ANALYSIS OF DIGITAL FOOTPRINT SOURCES FOR UEBA-CLASS SYSTEMS WITH QUANTITATIVE ASSESSMENT OF INFORMATIVENESS AND COMPUTATIONAL COST

V.V. Maltsev¹, G.I. Budovskaya¹

¹Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. The relevance of insider threat detection stems from the growing number of attacks by authorized users, which are difficult to identify using traditional security measures. User and Entity Behavior Analytics (UEBA) systems offer a solution based on analyzing employees' digital footprints; however, there is no methodology for selecting data sources that balances detection completeness and computational costs. The article presents a classification of digital footprint sources into five groups: access metadata, action

metadata, network flows, content, and behavioral biometrics. A proposed metric—the informativeness-to-extraction-complexity index—enables quantitative evaluation of the feasibility of using each data type. Experiments on the CMU-CERT r4.2 dataset showed that the maximum index (3.56) is achieved with network flows, while email content analysis provides an 8.3% improvement in detection quality but has a low index (0.41) due to high computational cost.

Keywords: UEBA, digital footprint, insider threats, behavior analysis, machine learning, feature informativeness, CMU-CERT.

Введение

Современные организации сталкиваются с растущей угрозой инсайдерских атак — вредоносных действий со стороны сотрудников, имеющих легитимный доступ к корпоративным ресурсам. Согласно отчету IBM за 2024 год, средняя стоимость инцидента с участием инсайдера достигла 4,99 млн долларов, а частота таких инцидентов за последние два года выросла на 44% [1]. Традиционные средства защиты, основанные на сигнатурном анализе и правилах корреляции, оказываются недостаточно эффективными, поскольку инсайдеры действуют в рамках штатных полномочий, маскируя свои действия под обычную активность [2].

В ответ на этот вызов появился класс систем User and Entity Behavior Analytics (UEBA), которые используют методы машинного обучения для построения профилей нормального поведения пользователей и выявления аномалий [3]. Однако практическое внедрение UEBA сталкивается с фундаментальной проблемой выбора источников данных. Современная ИТ-инфраструктура генерирует терабайты информации: логи доступа, действия с файлами, сетевой трафик, содержимое переписки [4]. Анализ контента потенциально наиболее информативен, но требует значительных вычислительных ресурсов и вызывает вопросы соблюдения законодательства о персональных данных [5]. Метаданные обрабатываются проще, но могут не содержать информации, достаточной для выявления сложных атак.

В литературе отсутствуют количественные критерии, позволяющие инженерам обоснованно выбирать источники данных для UEBA. Большинство работ фокусируются на алгоритмах детектирования, но не предлагают метрик для сравнения самих источников [6,7].

Цель работы — разработка методики количественного сравнения источников цифрового следа для задач UEBA, учитывающей как вклад каждого типа данных в качество обнаружения угроз, так и вычислительные затраты на их обработку.

1. Классификация источников цифрового следа

Анализ существующих подходов к построению UEBA-систем позволяет выделить пять групп источников, различающихся по природе происхождения и сложности извлечения признаков [4,8].

Первую группу составляют метаданные доступа. Сюда входят журналы событий операционных систем, логи VPN-подключений, данные прокси-серверов. Из этих источников извлекаются время входа и выхода, географическое местоположение, тип устройства, длительность сессии. На уровне абстракции метаданные доступа отвечают на вопросы «кто», «где» и «когда». Доступность этой группы высокая, поскольку соответствующие журналы генерируются всеми современными системами по умолчанию.

Вторую группу образуют метаданные действий. Источниками выступают журналы файловых серверов, DLP-системы, системы управления документами. Признаками являются тип действия (копирование, печать, удаление), имя объекта, размер, количество операций. На уровне абстракции метаданные действий отвечают на вопросы «кто», «что сделал» и «с каким объектом». Доступность этой группы средняя, требует включения аудита или установки агентов.

Третью группу представляют сетевые потоки. Информация собирается с помощью протоколов NetFlow, sFlow или извлекается из данных брандмауэров. Признаками служат IP-адреса назначения, порты, объем трафика, длительность соединений. На уровне абстракции сетевые потоки отвечают на вопросы «кто», «с кем общался» и «сколько передал». Доступность высокая, поддерживается большинством сетевых устройств.

Четвертую группу составляет контент. Сюда входит текст электронных писем, сообщения в мессенджерах, содержимое документов. Обработка требует методов NLP и позволяет извлекать N-граммы, стилометрические характеристики, тональность, психолингвистические маркеры [9]. На уровне абстракции контент отвечает на вопросы «о чем говорил» и «в каком контексте». Доступность низкая, требует DLP-систем с контентной фильтрацией и значительных ресурсов.

Пятую группу образует поведенческая биометрия. Данные собираются агентами на рабочих станциях и включают динамику набора текста, траектории движения мыши, частоту кликов. На уровне абстракции поведенческая биометрия отвечает на вопрос «как делал». Доступность низкая, требует установки специализированного ПО [10].

2. Методология оценки информативности

Для количественного сравнения выделенных групп предлагается интегральный показатель – индекс информативности к сложности извлечения (Informativeness-to-Difficulty Index, IDI). При разработке показателя учитывался опыт создания метрик для оценки качества информации, адаптированный к задачам анализа цифрового следа [11].

Индекс IDI вычисляется по формуле:

$$IDI = \frac{D \times U}{C \times T}$$

где D (Detection Gain) – прирост качества детектирования при добавлении данного источника к базовой модели, измеряемый в процентных пунктах увеличения F1-меры; U (Uniqueness) – уникальность признака, определяемая как 1 минус максимальная корреляция Пирсона с признаками из других групп; C (Computational Cost) – вычислительная стоимость обработки 1 Гбайт исходных данных в условных единицах времени процессора; T (Time) – временная задержка получения признака в секундах от момента генерации события до доступности признака для модели.

3. Экспериментальное исследование

Для эксперимента использован датасет CMU-CERT версии r4.2, разработанный отделом CERT Университета Карнеги-Меллон. Данный датасет является общепризнанным эталоном для тестирования систем обнаружения инсайдерских угроз и содержит синтезированные данные о поведении 4000 сотрудников в течение 18 месяцев, включая несколько сценариев инсайдерских атак [13].

В качестве базовой модели детектирования использован алгоритм изоляционного леса (Isolation Forest), рекомендованный для задач обнаружения аномалий [14]. Выбор обусловлен необходимостью исключить влияние сложности модели на результаты сравнения источников.

Эксперимент проводился в несколько этапов. Сначала выполнялось обучение базовой модели только на признаках из группы метаданных доступа с фиксацией F1-меры. Затем осуществлялось последовательное добавление признаков из остальных групп с оценкой прироста качества D. Далее рассчитывалась корреляционная матрица признаков для определения уникальности U. После этого измерялось время обработки 1 Гбайт исходных данных для каждой группы на конфигурации Intel Core i7-12700K с 32 ГБ RAM. В заключение вычислялся IDI для каждой группы источников.

Метаданные действий демонстрируют максимальный IDI (4,13) благодаря хорошему приросту качества при умеренной вычислительной стоимости и низкой латентности. Это объясняется тем, что информация о действиях с файлами непосредственно связана с типовыми сценариями инсайдерских атак. Сетевые потоки характеризуются наибольшим абсолютным приростом качества детектирования (10,7%). Несмотря на относительно невысокую уникальность (0,38), высокий IDI (3,56) подтверждает целесообразность включения данного источника.

Контент писем обеспечивает существенный прирост качества и высокую уникальность (0,71), поскольку текстовая информация слабо коррелирует с метаданными. Однако крайне низкий IDI (0,41) обусловлен высокой вычислительной стоимостью обработки текстов на естественном языке и значительной латентностью. Поведенческая биометрия дает минимальный прирост качества (2,5%) при наличии всех остальных источников. Умеренный IDI (0,50) позволяет рассматривать данный источник как дополнительный, но не первоочередной.

Полученные результаты согласуются с практическими наблюдениями специалистов по безопасности и дополняют их количественными оценками [15].

4. Практические рекомендации

На основе проведенного анализа можно сформулировать следующие рекомендации по выбору источников цифрового следа.

Для типовой организации с ограниченными вычислительными ресурсами оптимальным является фокус на метаданных доступа, метаданных действий и сетевых потоках. Данная конфигурация обеспечивает F1-меру 0,76 при умеренных затратах. Для организаций с высокими требованиями к безопасности целесообразно дополнительное подключение анализа контента, но с применением двухуровневой схемы: сначала фильтрация подозрительных событий по метаданным, затем углубленный контент-анализ только для отобранных случаев. Поведенческая биометрия рекомендуется для сценариев, где критически важна аутентификация пользователя в реальном времени, но не как основной источник для UEBA. При проектировании архитектуры систем UEBA следует учитывать весовые коэффициенты, пропорциональные IDI, для взвешивания вклада различных модулей в интегральную оценку риска.

Заключение

В работе предложена методика количественного сравнения источников цифрового следа для задач обнаружения инсайдерских угроз с использованием систем UEBA. Разработана классификация источников, включающая пять групп с различным уровнем абстракции и доступности. Предложен интегральный показатель IDI, учитывающий прирост качества детектирования, уникальность признаков, вычислительную стоимость и латентность. Экспериментально на датасете CMU-CERT r4.2 показано, что максимальную эффективность обеспечивают метаданные действий (IDI=4,13) и сетевые потоки (IDI=3,56). Установлено, что анализ контента писем, несмотря на высокую информативность, характеризуется низкой экономической эффективностью из-за высокой вычислительной стоимости. Сформулированы практические рекомендации по выбору источников данных в зависимости от условий эксплуатации. Полученные количественные оценки могут быть использованы при проектировании архитектур гибридных нейросетевых моделей для UEBA.

Список литературы

1. IBM Security. Отчет о стоимости утечек данных за 2024 год // Корпорация IBM. – 2024. – 76 с. – URL: <https://www.ibm.com/security/data-breach> (дата обращения: 15.02.2026).
2. Падиев П. Анализ злонамеренных инсайдерских угроз целостности данных: Докторская диссертация. – Австралия: Университет Чарльза Стерта, 2024. – 210 с.
3. Шелухин О.И. Анализ методов выявления редкой аномальной активности пользователей в информационных системах // Международный вестник Дона. – 2025. – № 1. – URL: <http://www.ivdon.ru/ru/magazine/archive/n1y2026/10690> (дата обращения: 15.02.2026).
4. Останина Е.А., Поколодина Е.В. Цифровой след: понятие, источники, методы сбора и анализа // Открытое образование. – 2024. – Т. 28, № 3. – С. 24–33.
5. Нассир Н.Ф.М., Рауф У.Ф.А., Зайнол З., Гани К.А. Методы прогнозирования инсайдерских угроз: систематический обзор // Технологические горизонты: Весенние новости прикладных наук и технологий. – Cham: Springer Nature, 2024. – С. 119-126.

6. Сапонова А.В. Интеграция опросных данных и цифровых следов в социологических исследованиях: возможности и ограничения: дис. ... канд. наук. – М., 2025. – 152 с.
7. Ма В., Растоги Н. ДАНТЕ: прогнозирование инсайдерской угрозы с помощью LSTM на основе системных журналов // Материалы 19-й Международной конференции IEEE 2020 по доверию, безопасности и конфиденциальности в вычислительной технике и коммуникациях (TrustCom). – IEEE, 2020. – С. 1151-1156.
8. Модель обнаружения внутренних угроз на основе CNN-LSTM // Международная конференция 2025 года по электрической автоматизации и искусственному интеллекту (ICEAAI). – IEEE, 2025. – С. 985-990.
9. Уланов К.А. Метрики качества данных // Молодой ученый. – 2024. – № 20 (519). – С. 17–19.
10. Barracuda Networks. Исследование показывает глобальное снижение затрат на утечку данных // Блог Barracuda. – 2025. – 2 сентября. – URL: <https://blog.barracuda.com/2025/09/02/survey-global-decline-data-breach-costs> (дата обращения: 15.02.2026).
11. Чон Д., Чон У., Ким Т., Но А., Чхве Дж. Теоретико–информационная оценочная метрика для сильного отлучения // Препринт arXiv arXiv: 2405.17878. – 2024.
12. Робертс С. У. Тесты контрольных карт, основанные на геометрических скользящих средних // Технометрия. – 1959. – Том 1, № 3. – С. 239-250.
13. Линдауэр Б. Набор данных для проверки инсайдерских угроз // Университет Карнеги-Меллона. – 2020. – URL: https://kithub.cmu.edu/articles/dataset/Insider_Threat_Test_Dataset/12841247/1 (дата обращения: 15.02.2026).
14. Лю Ф.Т., Тин К.М., Чжоу З.Х. Изолированный лес // Восьмая международная конференция IEEE по интеллектуальному анализу данных 2008 года. – IEEE, 2008. – С. 413-422.
15. Лин Д. Объяснена система оценки поведения пользователей и организаций в аналитике поведения // Блог Exabeam. – 2017. – URL: <https://www.exabeam.com/blog/ueba/a-user-and-entity-behavior-analytics-scoring-system-explained/> (дата обращения: 15.02.2026).

References

1. IBM Security. Cost of a Data Breach Report 2024 // IBM Corporation. – 2024. – 76 p. – URL: <https://www.ibm.com/security/data-breach> (дата обращения: 15.02.2026).
2. Padiet P. Analysis of Malicious Insider Threats to Data Integrity: Doctoral Thesis. – Australia: Charles Sturt University, 2024. – 210 p.
3. Shelukhin O.I. Analysis of methods for detecting rare abnormal user activity in information systems // Инженерный вестник Дона. – 2025. – № 1. – URL: <http://www.ivdon.ru/ru/magazine/archive/n1y2026/10690> (дата обращения: 15.02.2026).
4. Останина Е.А., Поколодина Е.В. Цифровой след: понятие, источники, методы сбора и анализа // Открытое образование. – 2024. – Т. 28, № 3. – С. 24–33.
5. Nassir N.F.M., Rauf U.F.A., Zainol Z., Ghani K.A. Insider Threat Prediction Techniques: A Systematic Review // Tech Horizons: SpringerBriefs in Applied Sciences and Technology. – Cham: Springer Nature, 2024. – P. 119–126.
6. Сапонова А.В. Интеграция опросных данных и цифровых следов в социологических исследованиях: возможности и ограничения: дис. ... канд. наук. – М., 2025. – 152 с.
7. Ma Q., Rastogi N. DANTE: predicting insider threat using LSTM on system logs // Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). – IEEE, 2020. – P. 1151–1156.
8. CNN-LSTM-Based Insider Threat Detection Model // 2025 International Conference on Electrical Automation and Artificial Intelligence (ICEAAI). – IEEE, 2025. – P. 985–990.
9. Уланов К.А. Метрики качества данных // Молодой ученый. – 2024. – № 20 (519). – С. 17–19.
10. Barracuda Networks. Survey sees global decline in data breach costs // Barracuda Blog. – 2025. – 2 Sept. – URL: <https://blog.barracuda.com/2025/09/02/survey-global-decline-data-breach-costs> (дата обращения: 15.02.2026).
11. Jeon D., Jeung W., Kim T., No A., Choi J. An Information Theoretic Evaluation Metric for Strong Unlearning // arXiv preprint arXiv:2405.17878. – 2024.
12. Roberts S.W. Control chart tests based on geometric moving averages // Technometrics. – 1959. – Vol. 1, No. 3. – P. 239–250.

13. Lindauer B. Insider Threat Test Dataset // Carnegie Mellon University. – 2020. – URL: https://kithub.cmu.edu/articles/dataset/Insider_Threat_Test_Dataset/12841247/1 (дата обращения: 15.02.2026).
14. Liu F.T., Ting K.M., Zhou Z.H. Isolation Forest // 2008 Eighth IEEE International Conference on Data Mining. – IEEE, 2008. – P. 413–422.
15. Lin D. A User and Entity Behavior Analytics Scoring System Explained // Exabeam Blog. – 2017. – URL: <https://www.exabeam.com/blog/ueba/a-user-and-entity-behavior-analytics-scoring-system-explained/> (дата обращения: 15.02.2026).