УДК 004.891.2 DOI: 10.26102/2310-6018/2020.28.1.039

Использование методов и алгоритмов анализа данных и машинного обучения в UEBA/DSS для поддержки принятия управленческих решений

П.А. Савенков, П.С. Трегубов
Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет», Тула, Российская Федерация

Резюме: Целью данного исследования является разработка математического и программного обеспечения обнаружения аномального поведения пользователей на основе анализа их поведенческих биометрических характеристик для создания новых способов предоставления аналитических данных анализирующей службе с описанием, почему выявленные действия считаются аномальными. Предметом исследования являются методы машинного обучения, применяемые в UBA/UEBA (User Behavioral Analytics/ User and Entity Behavioral Analytics), DLP (Data Leak Prevention), SIEM (Security information and event management) системах. Объект исследования - UBA/UEBA, DLP, SIEM системы. В данной статье осуществляется обзор применяемости методов машинного обучения в интеллектуальных UEBA/DSS системах. Одной из существенных проблем, в интеллектуальных UEBA/DSS системах, является получение полезной информации, из большого объема неструктурированных, несогласованных данных. Методы и алгоритмы интеллектуальной обработки данных и машинного обучения, применяемые в UEBA/DSS системах, позволяют решить задачи анализа данных различной направленности. Предлагается применение методов машинного обучения в реализации мобильной UEBA/DSS системы. Это позволит добиться высокого качества анализа данных и найти в них сложные зависимости. В ходе исследования был сформирован перечень наиболее значимых факторов, подаваемых на вход анализирующих методов.

Ключевые слова: Big Data, Data science, большие данные, программное обеспечение, информационная система машинное обучение, UEBA, DSS.

Для цитирования: Савенков П.А. Использование методов и алгоритмов анализа данных и машинного обучения в UEBA/DSS для помощи принятия управленческих решений. *Моделирование, оптимизация и информационные технологии.* 2020;8(1). Доступно по: https:.moit.vivt.ru/wp-content/uploads/2020/02/SavenkovTregubov_1_20_1.pdf DOI: 10.26102/2310-6018/2020.28.1.039

Using the methods and algorithms for data analysis and machine learning in UEBA/DSS to support management decision-making

P.A. Savenkov, P.S. Tregubov
Federal State Budgetary Educational Institution of Higher Education "Tula State University", Tula, Russian Federation

Abstract: The aim of this study is to develop mathematical and software for detecting abnormal user behavior based on an analysis of their behavioral biometric characteristics to create new ways to provide analytical data to the analyzing service with a description of why the identified actions are considered abnormal. The subject of the study is the machine learning methods used in UBA / UEBA (User Behavioral Analytics / User and Entity Behavioral Analytics), DLP (Data Leak Prevention), SIEM (Security information and event management) systems. Object of study - UBA / UEBA, DLP, SIEM

systems. This article provides an overview of the applicability of machine learning methods in intelligent UEBA / DSS systems. One of the significant problems in intelligent UEBA / DSS systems is obtaining useful information from a large amount of unstructured, inconsistent data. The methods and algorithms of intelligent data processing and machine learning used in UEBA / DSS systems make it possible to solve data analysis problems of various kinds. The application of machine learning methods in the implementation of a mobile UEBA / DSS system is proposed. This will allow to achieve high quality data analysis and find complex dependencies in them. During the study, a list of the most significant factors submitted to the input of the analyzing methods was formed. The application of machine learning methods in UEBA / DSS systems will allow you to make informed management decisions and reduce the time to obtain useful information.

Keywords: big Data, data science, software, machine learning information system, UEBA, DSS.

For citation: Savenkov P.A. Use methods and algorithms for data analysis and machine learning in UEBA/DSS to assist management decisions. *Modeling, Optimization and Information Technology*. 2020;8(1). Available from: https:.moit.vivt.ru/wp-content/uploads/2020/02/SavenkovTregubov_1_20_1.pdf DOI: 10.26102/2310-6018/2020.28.1.039 (In Russ).

Введение

В данный момент одной из существенных проблем в DSS (Decision Support System) системах является получение полезной информации из большого объема неструктурированных разнородных и несогласованных данных [1].

Принятие управленческих решений руководителей предприятий должно основываться на основе реальных данных, собираемых с анализируемого объекта.

Анализируемым объектом, в данном случае, являться сотрудник предприятия, а управленческими решениями, решения, предпринимаемые руководством на основе полученных при помощи мобильного приложения поведенческих данных и их отклонений от эталонного профиля пользователя.

Однако, на основе выбранных с устройства пользователя данных достаточно сложно осуществить принятие какого-либо управленческого решения, так как данные являются разнородными, а их объемы крайне велики.

Для решения данной проблемы предлагается использовать методы машинного обучения и интеллектуальную обработку данных. Это позволит получать краткий перечень результирующих параметров, что позволит принимать взвешенные управленческие решения.

UEBA/DSS системы, основанные на методах машинного обучения, способны учитывать множество различных факторов во время работы и использовать результаты прошлых вычислений. Во время работы данные методы используют большие массивы неструктурированных разрозненных данных [2].

В UEBA/DSS системах достаточно часто применяется статистический и многомерный анализ, поиск и просмотр исходной информации, Data Mining. Однако использование первых двух методик оправдано лишь при наличии небольших объемов данных [3].

Задачу анализа многомерных, данных можно представить в виде треугольника (Рисунок 1) в основании которого объем исходных данных, а горизонтальная прямая, проведенная на разных уровнях, показывает, какой объем данных будут формировать соответствующие методы.

По мере роста объемов исходных данных меняется и метод их анализа. Для принятия решения, при наличии краткого перечня данных, достаточно использовать просмотр исходной информации. При среднем объеме исходных данных применяется

статистический и многомерный анализ, а при использовании больших данных используется Data Mining [3]. При использовании методов Data Mining, возможно получение максимально краткого объема информативных выходных значений, при достаточно большом количестве анализируемых входных параметров.

Методы и алгоритмы интеллектуальной обработки данных и машинного обучения, применяемые в UEBA/DSS системах, позволяют решить задачи анализа данных различной направленности. Среди таких методов можно выделить следующие: задача классификации, задача регрессии, задача ассоциации, задача кластеризации, последовательные шаблоны, анализ отклонений [4].

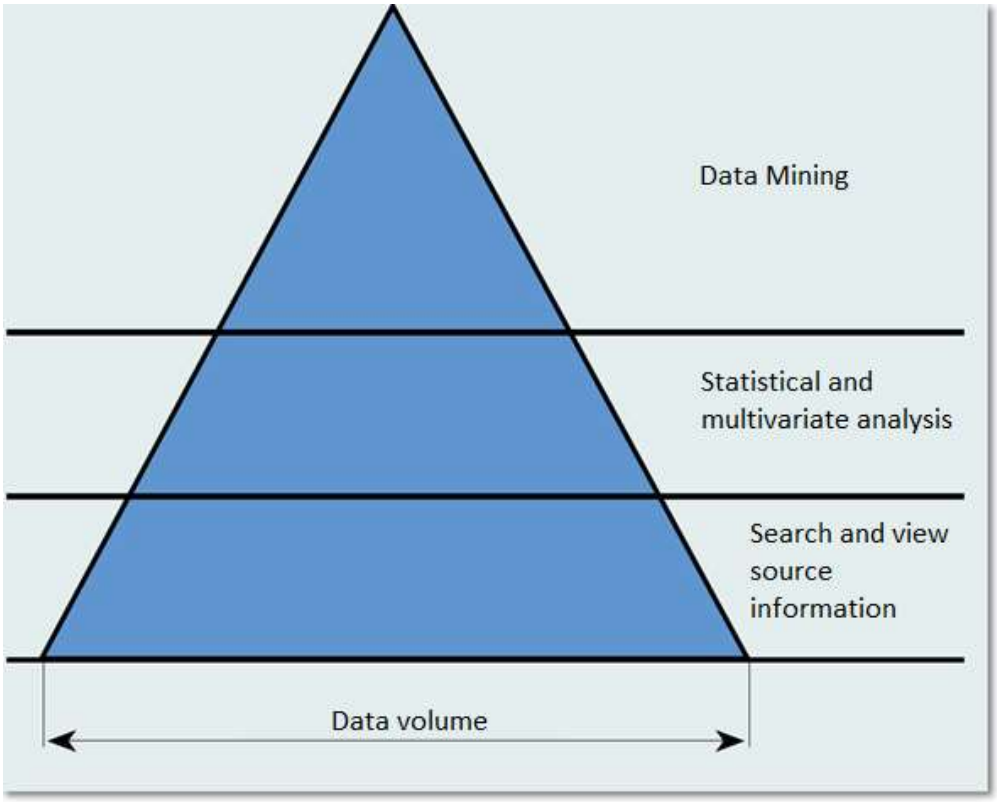

Рисунок 1 - Схема методов анализа данных
Figure 1 - Diagram of data analysis methods

Материалы и методы

Для реализуемой UEBA системы с функционалом DSS, основанной на анализе поведенческих биометрических характеристик персонала предприятия, в связи с большим объемом входных анализируемых данных, предлагается использовать методы машинного обучения и интеллектуальную обработку данных. Это позволит уменьшить количество результирующих параметров.

Для сбора исходных данных используются мобильные устройства сотрудников предприятия с ОС Android. Для поведенческого анализа, предлагается использовать следующие методы:

- метод k ближайших соседей;
  - используемые приложения;
  - координаты GPS (история перемещения сотрудника);
  - посещаемые сайты;

  - набираемый текст;
  - получаемый текст.
- нейронные сети;
  - звонки, диктофон;
  - камера, изображение.

Программное обеспечение указывает на определенные отклонения поведенческих характеристик пользователя, предлагает осуществить ряд действий администратору. В некоторых случаях администратор системы принимает решение о блокировке пользователя.

Нейронные сети применяются для анализа таких данных как записанные звонки, записанный звук с диктофона и фотографии [5]. Для нахождения в них отклонений, происходит предварительное обучение сети.

Для нахождения отклонений от эталонного профиля пользователя в таких данных как, история перемещений сотрудника (GPS), набираемый текст, получаемый текст применяется метод k ближайших соседей. Использование данного метода позволяет уменьшить нагрузку при анализе данных, а также сократить количество итераций при обучении. В процессе обучения данный метод только хранит тренировочные данные. Классификация осуществляется при получении на входе метода новых немаркированных данных. В данном случае происходит проверка полученных от пользователя данных и поиск их принадлежности к определенной группе пользователей или к определенному пользователю.

Сравнение характеристик пользователя осуществляется путем поиска Евклидова расстояния до всех записей из полученной выборки [6].

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2} \quad (1)$$

Затем производится отбор k записей, для которых евклидово расстояние от текущей записи до новой будет минимальным. Далее для каждого пользователя осуществляется подсчет суммы обратных квадратов расстояний между записями этого класса и новой записью. Новой записи присуждается класс, для которого сумма обратных квадратов получается наибольшей. На Рисунке 2 показано графическое представление алгоритма. В данном случае метод knn использует такие параметры как:

- $d(p,q)$ – расстояние между точками;
- p_n – координата точки p по оси n;
- q_n – координата точки q по оси n.

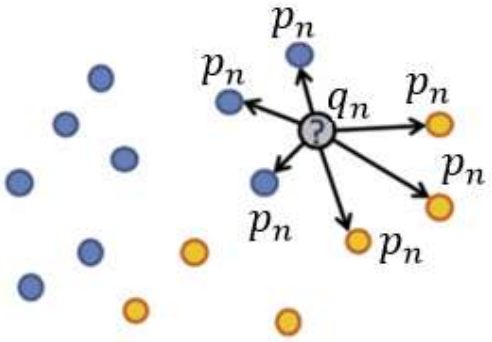

Рисунок 2 - Метод knn
Figure 2 – knn method

Если идентификатор пользователя или группы, присвоенный методом к анализируемой записи, соответствует идентификатору пользователя или группы, полученному при начальной авторизации в системе, считается, что полученные характеристики соответствуют эталонным и отклонения от эталонного профиля не были найдены.

В случае если идентификатор полученный при начальной авторизации в системе не соответствует идентификатору, присвоенному методом к новой сформированной записи, то считается, что полученные характеристики отличаются от эталонных или принадлежат другому пользователю или группе пользователей.

На основе «просеянных» данных программное обеспечение предлагает осуществить ряд действий администратору указывая на определенные отклонения пользователя от эталонного профиля. В некоторых случаях администратор системы принимает решение о блокировке пользователя [7]. На Рисунке 3 представлена базовая схема связей в системе.

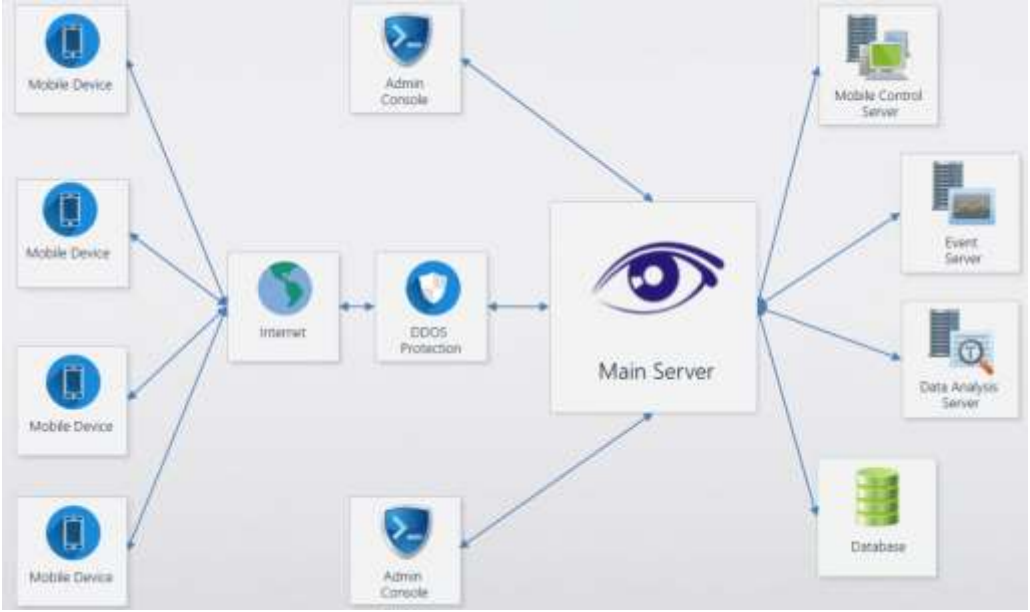

Рисунок 3 - Схема связей системы
Figure 3 - System connection diagram

На каждое мобильное устройство, подключенное к системе, устанавливается мобильное приложение – клиент. После установки приложения, на мобильное устройство сотрудника, администратор системы назначает пользователю перечень собираемых параметров. Набор анализируемых параметров, которые будут собираться на устройстве и анализироваться на сервере, отличается в зависимости от пользователя/группы пользователей. Формирование перечня анализируемых параметров и групп пользователей осуществляется администратором системы.

Мобильное приложение (Mobile Device) запускается при старте мобильного устройства как сервис. Мобильное устройство запрашивает перечень команд с сервера раз в N минут. После получения команд идет их обработка, получение соответствующей информации и отправка данных на сервер (Main Server). Команды (Event Server) имеют различные приоритеты выполнения. Так же команды имеют различные статусы, такие как однократное выполнение и циклическое выполнение с таймером. После отправки данных на сервер происходит их прием главным сервером (Main Server), обработка,

последующий анализ (Data Analisis server), и запись в базу данных (DataBase) исходных и результирующих параметров. В случае ошибки данная команда выполняется повторно.

Архитектура построена таким образом, что со стороны клиентских устройств невозможно получить информацию из базы данных, что исключает утечку информации о пользователях. Мобильные устройства получают лишь перечень команд, которые должны выполнить и ответить серверу.

Панель администратора (Admin Console) подключена напрямую к главному серверу и имеет определенные возможности, такие как:

- управление группой или определенным пользователем (Mobile Control Server);
- добавление к выполнению новых команд (Event Server);
- генерирование отчетов.

Прямой доступ к серверу обеспечивает постоянный доступ к управлению на случай ddos атаки, в обход фильтрации ddos атак (DDOS Protection).

Результаты

В ходе исследования был сформирован перечень наиболее значимых факторов, подаваемых на вход анализирующих методов. С ростом количества признаков экспоненциально увеличивалось количество объектов, которые должны находиться в обучающей выборке для покрытия всевозможных ситуаций. Была построена базовая архитектура клиент серверного комплекса программного обеспечения, обеспечивающая высокую стабильность в обработке данных. Уменьшение количества входных параметров позволило экспоненциально уменьшить объем обучающей выборки для метода knn, так как всевозможное количество комбинаций признаков XN, где N это количество признаков, X количество состояний признака.

Благодаря уменьшению количества входных параметров удалось добиться достаточно высокой корректности идентификации пользователя по полученным характеристикам. Корректность идентификации пользователей представлен на Рисунке 4.

Рисунок 4 - Диаграмма корректности идентификации
Figure 4 - Identity Correction Chart

В среднем данные пользователей были корректно идентифицированы в 92% случаев и некорректно в 8% случаев. В связи с уменьшением количества выходных данных и повышением корректности идентификации пользователей по их

характеристикам, было уменьшено среднее время получения полезных данных из системы поддержки принятия управленческих решений.

Заключение

Благодаря применению метода машинного обучения Knn для класстеризации данных, в разрабатываемой программной системе, удалось добиться повышения информативности результирующих данных перемещений сотрудников. Применение данных методов позволило находить аномалии в перемещениях каждого пользователя системы. Применение данных методов позволит принимать взвешенные управленческие решения относительно сотрудника и применять к нему соответствующие дисциплинарные санкции. Так же использование машинного обучения, в частности использованных методов позволит сократить время на анализ выбранной информации о перемещениях сотрудника. Актуальность применения методов машинного обучения в системах поддержки принятия решений в настоящее время очень высока и будет увеличиваться со временем по мере развития новых информационных технологий.

ЛИТЕРАТУРА

1. Cai L., Zhu Y. The challenges of data quality and data quality assessment in the big data era. *Data science journal*.2015;14.
2. Cao J. et al. Big data: A parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce .*PloS one*. 2016;11(6).
3. Chen H., Chiang R. H. L., Storey V. C. Business intelligence and analytics: From big data to big impact .*MIS quarterly*. 2012;36(4).
4. Dutt A., Ismail M. A., Herawan T. A systematic review on educational data mining .*IEEE Access.* – 2017;5.
5. Ivutin A. N., Savenkov P. A., Veselova A. V. Neural network for analysis of additional authentication behavioral biometrie characteristics .*2018 7th Mediterranean Conference on Embedded Computing (MECO)*. 2018;(7).
6. Wang J., Neskovic P., Cooper L. N. Improving nearest neighbor rule with a simple adaptive distance measure .*Pattern Recognition Letters*.2007;28(2):207-213.
7. Yan Z. et al. *Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach .2012 16th international symposium on wearable computers*. – Ieee, 2012;16.

REFERENCES

1. Cai L., Zhu Y. The challenges of data quality and data quality assessment in the big data era .*Data science journal*.2015;14.
2. Cao J. et al. Big data: A parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce .*PloS one*. 2016;11(6).
3. Chen H., Chiang R. H. L., Storey V. C. Business intelligence and analytics: From big data to big impact .*MIS quarterly*. 2012;36(4).
4. Dutt A., Ismail M. A., Herawan T. A systematic review on educational data mining .*IEEE Access.* – 2017;5.
5. Ivutin A. N., Savenkov P. A., Veselova A. V. Neural network for analysis of additional authentication behavioral biometrie characteristics .*2018 7th Mediterranean Conference on Embedded Computing (MECO)*. 2018;(7).

6. Wang J., Neskovic P., Cooper L. N. Improving nearest neighbor rule with a simple adaptive distance measure .*Pattern Recognition Letters*.2007;28(2):207-213.
7. Yan Z. et al. *Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach .2012 16th international symposium on wearable computers*. – Ieee, 2012;16.

ИНОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Савенков Павел Анатольевич, аспирант, кафедра вычислительной техники, ФГБОУ ВО "Тульский государственный университет" Институт прикладной математики и компьютерных наук, Тула, Российская Федерация.
e-mail: pavel@savenkov.net

Pavel A. Savenkov, Postgraduate Student, Department of Computer Engineering, Tula State University Institute of Applied Mathematics and Computer Science, Tula, Russian Federation.

Трегубов Павел Сергеевич, магистрант, кафедра вычислительной техники, ФГБОУ ВО "Тульский государственный университет" Институт прикладной математики и компьютерных наук, Тула, Российская Федерация.
e-mail: tregubov.1997@yandex.ru

Pavel S. Tregubov, Undergraduate Student, Department of Computer Engineering, Tula State University Institute of Applied Mathematics and Computer Science, Tula, Russian Federation.