11. *Styblinski M. A.* Statistical design optimization in *Computer aided design and design automation*, *Wai Kai Chen* Ed. Boca Raton: CRC Press, 2009, pp. 5-1-5-34.
12. *Sokolov B.V., Shevtsova T.G.* Metody opredeleniya dopuskov elektricheskikh tsepey i invariantov chuvstvitel'nosti [Methods for determining the tolerances of electrical circuits and sensitivity invariants], *Vestnik Kuzbasskogo gosudarstvennogo tekhnicheskogo universiteta* [Bulletin of the Kuzbass state technical university], 2010, No. 3, pp. 65-69.
13. *Gajda J., Sidor T.* Using Monte Carlo analysis for practical investigation of sensitivity of electronic converters in respect to component tolerances, *Electrical and Electronic Engineering*, 2012, No. 2, pp. 297-302. DOI: 10.5923/j.eee. 20120205.09. 2012.
14. *Pehl M., Graeb H.* Tolerance design of analog circuits using a branch-and-bound based approach, *Journal of Circuits, Systems and Computers*, 2013, Vol. 21, No. 8, pp. 1240022-1-1240022-17. DOI: 10.1142/S0218126612400221.
15. *Shilo G., Furmanova N., Kulyaba-Kharitonova T.* Software for tolerance design of electronic devices, *International Conference Advanced Computer Information Technologies (ACIT 2018)*, 2018, pp. 14-17.
16. *Khludenev A.* Tolerance design of active RC filters, *2021 International Seminar on Electron Devices Design and Production (SED 2021)*. IEEE, 2021, 9444367. **DOI:**10.1109/SED51197.2021.9444367.
17. *Fitzpatrick D.* Analog design and simulation using OrCAD Capture and PSpice, 2nd ed. Newnes, 2018, 438 p.
18. *Vlach J., Singhal K.* Computer methods for circuit analysis and design. New York: Van Nostrand Reinhold, 1983, 594 p.
19. *Mullins E., Mrabet A.* Analog front-end design for a narrowband power-line communications modem using the AFE031: Application Report SBOA130A. Texas Instruments, 2011, 35 p. Available at: https://www.ti.com/lit/pdf/sboa130.
20. *Wang B., Cao Z.* Design of active power filter for narrow-band power line communications, *2018 2nd International Conference on Material Engineering and Advanced Manufacturing Technology (MEAMT 2018)*, MATEC Web of Conferences, 2018, Vol. 189, 04012. DOI: 10.1051/matecconf/201818904012.

Статью рекомендовал к опубликованию д.т.н., профессор Ю.А. Кравченко.

Хлуденев Александр Владимирович – Оренбургский государственный университет; e-mail: avhludenev@yandex.ru; г. Оренбург, Россия, тел.: +73532372874; кафедра промышленной электроники и информационно-измерительной техники; к.т.н.; доцент.

Khludenev Alexander Vladimirovich – Orenburg State University; e-mail: avhludenev@yandex.ru; Orenburg, Russia; phone: +73532372874; industrial electronics and information measuring engineering department; cand. of eng. sc.; associate professor.

УДК 004.89 DOI 10.18522/2311-3103-2023-3-74-85

И.С. Берешполов, Ю.А. Кравченко, А.Г. Слепцов

АЛГОРИТМ КЛАСТЕРИЗАЦИИ ДАННЫХ ДЛЯ ЗАЩИТЫ КОНФИДЕНЦИАЛЬНОЙ ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ*

Статья посвящена решению научной задачи защиты конфиденциальной информации в сети Интернет на основе алгоритма кластеризации значительных объемов данных. Защита конфиденциальной информации компьютерной сети является актуальной темой для исследований, особенно в связи с растущим использованием информационных технологий и увеличением объема данных ценной информации, хранящейся в Интернете. С ростом информационной ответственности необходимость в эффективных методах информацион-

* Исследование выполнено за счет гранта Российского научного фонда № 22-21-00316, https://rscf.ru/project/22-21-00316/ в Южном федеральном университете.

ной безопасности компьютерных сетей стала критически важной. В данной научной статье авторы предлагают решение задачи защиты конфиденциальной информации компьютерных сетей на основе алгоритма кластеризации больших данных. Традиционные методы обнаружения вторжений имеют такие ограничения, как способность работать только с одно- или двумерными данными, а также имеют сильную зависимость от предварительных знаний. Авторы для устранения этих ограничений предлагают эвристический алгоритм обнаружения вторжений, который использует кластеризацию на основе облачной модели. Предлагаемый алгоритм использует преимущества как маркированных, так и немаркированных образцов для кластеризации данных, тем самым уменьшая зависимость от априорных знаний. Результаты вычислительного эксперимента, проведенного на предложенном алгоритме, сравнивались с несколькими каноническими алгоритмами обнаружения вторжений. Результаты показали, что предложенный алгоритм улучшил производительность системы обнаружения вторжений, повысил точность обнаружения, снизил частоту ложных тревог и усилил надежность системы. Метод динамического взвешивания, используемый в алгоритме, устранил сложность высокоуровневой обработки данных и позволил алгоритму самообучаться, что привело к формированию относительно стабильной облачной модели. Несмотря на значительное улучшение производительности предложенного алгоритма по сравнению с каноническими алгоритмами кластеризации, результаты исследования также показали, что у алгоритма есть некоторые ограничения, такие как высокий процент ложных срабатываний и чувствительность к данным с определенными видами распределения. Для устранения этих недостатков необходимо дальнейшее усовершенствование алгоритма. В целом, предложенный эвристический алгоритм обнаружения вторжений с кластеризацией на основе облачной модели представляет собой перспективное решение для защиты конфиденциальной информации компьютерных сетей.

Информационная безопасность; конфиденциальная информация; кластеризация; облачная модель; эвристический алгоритм.

I.S. Bereshpolov, Yu.A. Kravchenko, A.G. Sleptsov

DATA CLUSTERING ALGORITHM FOR PROTECTING CONFIDENTIAL INFORMATION ON THE INTERNET

The article is devoted to solving the scientific problem of protecting confidential information in the Internet based on the algorithm for clustering significant amounts of data. The protection of a computer network confidential information is a hot topic for research, especially in connection with the growing use of information technology and the increase in data of valuable information stored in the Internet. With the growth of information responsibility, the need for effective methods of computer networks information security has become critical. In this scientific article, the authors propose a solution to the problem of protecting computer networks confidential information based on the big data clustering algorithm. Traditional intrusion detection methods have limitations such as the ability to work only with one- or two-dimensional data, and also have a strong reliance on prior knowledge. To eliminate these limitations, the authors propose a heuristic intrusion detection algorithm that uses clustering based on a cloud model. The proposed algorithm takes advantage of both labeled and unlabeled samples for data clustering, thereby reducing reliance on a priori knowledge. The results of a computational experiment carried out on the proposed algorithm were compared with several canonical intrusion detection algorithms. The results showed that the proposed algorithm improved the performance of the intrusion detection system, increased the accuracy of detection, reduced the false alarm rate, and enhanced the reliability of the system. The dynamic weighting method used in the algorithm removed the complexity of high-level data processing and allowed the algorithm to learn itself, resulting in a relatively stable cloud model. Despite the significant improvement in the performance of the proposed algorithm compared to the canonical clustering algorithms, the results of the study also showed that the algorithm has some limitations, such as a high false positive rate and sensitivity to data with certain types of distribution. To eliminate these shortcomings, further improvement of the algorithm is required. In general, the proposed heuristic clustering intrusion detection algorithm based on the cloud model is a promising solution for protecting computer networks confidential information.

Information security; confidential information; clustering; cloud model; heuristic algorithm.

Введение. Благодаря своему постоянному развитию, интернет-технологии стали применяться в различных областях и оказали значительное влияние на жизнь людей, а также привели к глобальному росту объемов обрабатываемых и хранимых данных. Очевидно, что в подобных условиях выросла актуальность проблемы повышения эффективности средств обеспечения информационной безопасности [1–3]. Одними из основных задач, решаемых в сфере информационной безопасности, являются классификация и кластеризация данных. В настоящее время разработано множество различных методов и алгоритмов для решения данных задач.

Наиболее уязвимыми с точки зрения обеспечения информационной безопасности являются процессы передачи информации в компьютерной сети. При этом, основными проблемами сетевой информационной безопасности являются следующие:

1) уязвимость протоколов TCP/IP;

2) слабая защищенность сетевой структуры;

3) повышенный риск хищения информации;

4) слабая осведомленность о мерах защиты информации [2–7].

Компьютерная сетевая система имеет повышенные риски для информационной безопасности, в основном потому, что состоит из множества локальных сетей, это увеличивает размер такой сети и делает ее уязвимой для атак злоумышленника, которому достаточно только передать хост, после чего он может действовать, проводя атаку для хищения ценной информации. Часто используемое бесплатное программное обеспечение имеет сниженные возможности для шифрования ценной информации, что также делает систему более уязвимой с точки зрения достаточности информационной безопасности.

Помимо этого, некоторые пользователи считают, что брандмауэр доставляет много хлопот и влияет на использование ими некоторого программного обеспечения, поэтому они решают закрыть брандмауэр, в случае отсутствия аутентификации прокси-сервера брандмауэра и соединения через двухточечный протокол канального уровня PPP брандмауэр становится бесполезным, а потенциальные угрозы безопасности могут возникнуть в любой момент.

Есть много факторов, которые угрожают безопасности информации компьютерной сети, среди которых взлом является одним из наиболее распространенных, хакеры представляют большую угрозу для современных компьютерных сетевых систем [2–4, 7–9]. Если сеть атакована хакерами и сервер поврежден, он не может нормально обслуживать пользователей, в результате сеть парализована, что приводит к негативным последствиям.

Система предотвращения вторжений (Intrusion Prevention System, IPS) сочетает в себе функции обнаружения атак, она может контролировать сетевой трафик, своевременно прерывать, корректировать или изолировать источники потенциальной опасности, является активной и эффективной системой информационной защиты [8–10]. Система предотвращения вторжений различает злонамеренную активность и предполагаемые планы атак, постоянно анализируя оперативную сетевую информацию. Механизм IPS проверяет сетевой трафик и последовательно анализирует его с помощью набора входящих сигнатурных данных для выявления планов атак. IPS может удалить подобную вредоносную активность, а затем заблокировать весь будущий трафик с IP-адреса или порта злоумышленника. Реальный трафик может продолжать передаваться без каких-либо явных ограничений и помех. На рис. 1 показано развертывание приложения предотвращения вторжений.

Обычно IPS регистрирует данные об обнаруженных событиях, отправляет сообщение в службу безопасности и составляет необходимые отчеты [11, 12]. IPS может естественным образом получать обновления функций сдерживания вредоносных атак для обеспечения информационной безопасности, чтобы постоянно отслеживать и устранять угрозы, связанные с Интернетом, что может помочь защитить организацию.

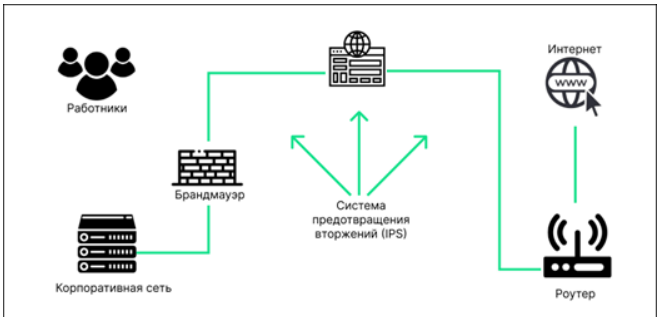

Рис. 1. Развертывание приложения предотвращения вторжений (IPS)

Эффективность систем обеспечения информационной безопасности может быть повышена на основе применения в логике их работы методов машинного обучения. Данные методы позволяют, в отличие от переборных подходов, находить наборы квазиоптимальных решений для обеспечения информационной безопасности на основе применения процедур диверсификации пространства поиска при попадании в «локальные ямы» Рассмотрим основные особенности и преимущества методов машинного обучения.

1. Методы машинного обучения в задачах обеспечения информационной безопасности. Машинное обучение (МО) определяется как дисциплина искусственного интеллекта (ИИ), которая предоставляет машинам возможность автоматически учиться на данных и прошлом опыте, чтобы выявлять закономерности и делать прогнозы с минимальным вмешательством человека. Методы машинного обучения позволяют компьютерам работать автономно без явного программирования. Приложения машинного обучения получают новые данные и могут самостоятельно учиться, эволюционировать и адаптироваться [12–15]. Машинное обучение извлекает полезную информацию из значительных объемов данных, используя алгоритмы для выявления скрытых зависимостей и закономерностей, а также обучения в итеративном режиме. Алгоритмы машинного обучения используют методы вычислений, чтобы учиться непосредственно на данных вместо того, чтобы применять аналитические выражения, которые могут служить математической моделью.

С развитием вычислительной мощности компьютеров машинное обучение проникло во многие области, такие как распознавание образов, интеллектуальный анализ данных и компьютерная графика. Методы машинного обучения имеют собственную классификацию. В зависимости от того, имеют ли данные обучающей выборки, используемые в процессе обучения, информацию о метках, машинное обучение можно разделить на неконтролируемое обучение (обучение без учителя), контролируемое обучение (обучение с учителем) и частично контролируемое обучение (обучение с частичным привлечением учителя) [8–10, 15, 16]. Приведем описание методов в соответствии с представленной классификацией.

Обучение без учителя – это своего рода информация о неклассифицированных данных для анализа и распознавания, в то же время знания, связанные с кластером, могут применяться к неконтролируемому обучению, чтобы анализировать данные выборки и прогнозировать информацию о категории выборки [16]. При неконтролируемом обучении набор известных образцов $X = (x_1, x_2, \dots, x_n)$. Выборка независима и одинаково распределена, поэтому метод исследования неконтролируемого обучения заключается в определении $(n \times d)$ матрицы, строки которой представляют выборки:

$$X = (x_i^T)_{i\in[n]}^T.$$

Цель неконтролируемого обучения состоит в том, чтобы обнаружить различную структурную информацию и законы, содержащиеся в матрице. Неконтролируемое обучение не выполняет предварительную подготовку на обучающих выборках, также отсутствует доступная информация о контроле, и невозможно установить библиотеку функций выборок [16–18]. Если классификатор продолжает принимать большое количество краевых тестовых выборок, это может повлиять на точность классификации, что приведет к неправильной классификации.

Обучение с учителем – это традиционный метод машинного обучения, в котором используются предварительные знания, предоставляемые системой (такие как информация об отмеченном классе выборки, информация о парных ограничениях и априорная вероятность) [16–18]. Изучается известный набор обучающих выборок, настраиваются параметры классификатора и устанавливается модель обучения выборки, затем классификация неизвестных выборок осуществляется в соответствии с моделью выборки. В обучении с учителем набор образцов X связан метками с классом образца:

$$Y = (y_1, y_2, \dots, y_n).$$

Как известно, y_i – метка класса, соответствующая образцу x_i, а пара данных (x_i, y_i) представляет собой обучающий набор образцов, необходимых для построения обучаемого [19]. Обучение под наблюдением с поиском отображения между известным обучающим набором x и меткой y, соответствующим образом строит требуемого обучаемого.

Размеченные данные часто трудно получить при обучении с учителем, а также необходимо создать библиотеку функций. Это ведет к тому, что функции новых данных могут не соответствовать функциям в библиотеке, в результате это может привести к неправильной классификации.

Обучение с частичным привлечением учителя – это метод является средним приближением между обучением без учителя и обучением с учителем, набор данных, используемый в процессе обучения, обычно содержит небольшое количество отмеченной информации [19], через эти образцы идентификационной информации метод реализует изучение неизвестных образцов.

При обучении с частичным привлечением учителя весь набор данных $X = (x_1, x_2, \dots, x_n)$ делится на две части: набор известных размеченных данных $X = (x_1, x_2, \dots, x_l)$, соответствующая маркировка $Y = (y_1, y_2, \dots, y_l)$ и наборы данных с неизвестными метками:

$$X_u = (x_{l+1}, x_{l+2}, \dots, x_{l+u}).$$

Основное содержание, которое необходимо изучить при обучении с частичным привлечением учителя, заключается в том, как всесторонне использовать отмеченные образцы и образцы без маркировки.

2. Постановка задачи кластеризации для систем обеспечения информационной безопасности. Класс или кластер – это набор объектов данных, объекты данных в одном кластере похожи друг на друга, в отличие от объектов в других кластерах. С точки зрения машинного обучения кластерный анализ – это тип обучения без учителя [20]. Перед выполнением кластерного анализа неизвестно, на сколько категорий разделятся входные данные. Группировка в кластерах производится с учетом сходства оцениваемых признаков между данными. Выполняется правило максимизации сходства между данными одного кластера, и минимизации сходства между данными разных кластеров.

Методы кластеризации для решения разных задач, также различны, но все они основаны на определенной последовательности этапов. Большинство методов кластеризации состоят из четырех этапов: выбор или извлечение признаков, разра-

ботка или выбор алгоритма кластеризации, подтверждение кластера и интерпретация результатов. Это является процессом преобразования данных в ценные знания. В узком смысле кластеризация включает разработку и выбор алгоритмов решения задачи, процесс подтверждения кластеризации и интерпретацию результатов.

Алгоритм k-средних является одним из классических алгоритмов, работающих на основе определения центроида. Его первым шагом является определение количества кластеров (K). Второй шаг – случайным образом сгенерировать K начальных центров кластеров, т. е. $C = \{c_1, c_2, \dots, c_k\}$. Третий шаг – присвоить каждый объект данных кластеру с наибольшим сходством. Четвертый шаг – оценка соответствия центров кластеров и их перемещение в случае необходимости [20]. Алгоритм k -средних имеет преимущества легкого понимания, простой реализации и высокой скорости сходимости.

Случайный выбор начальных центров кластеров приводит к множеству начальных центров кластеров в одной группе, особенно когда данные сложные. Более того, сложно найти оптимальный центр кластера за ограниченное количество итераций. Следовательно, алгоритм k-средних легко сходится к локальному оптимуму, что приводит к неудовлетворительным результатам кластеризации.

Таким образом, алгоритм k-средних имеет преимущества простого использования, быстрой сходимости и низких затрат памяти, в то же время, есть весомые недостатки, такие как зависимость производительности алгоритма от инициализированных прототипов кластера, что приводит к нестабильности работы и чувствительности к шумовым выбросам. Рассмотрим разработку полууправляемого алгоритма кластеризации данных для защиты конфиденциальной информации в сети Интернет, который позволяет улучшить показатели кластеризации по сравнению с каноническими алгоритмами.

3. Разработка алгоритма кластеризации данных для защиты конфиденциальной информации в сети Интернет. В алгоритме обнаружения вторжений существует проблема деления порога, значение порога напрямую влияет на результат обнаружения, и на практике он не может гибко реагировать на ситуации вторжения. Кроме того, применение классификаторов общей облачной модели для обнаружения вторжений часто реализуется с помощью генератора правил ассоциации, который работает медленно, учитывая, что свойства сетевых данных не являются исчерпывающими. В практических приложениях подобный генератор правил не может справиться со сложными изменениями сетевой среды. Для повышения качества работы алгоритмов кластеризации, авторы используют комбинированный вариант алгоритма обучения с частичным привлечением учителя с облачной моделью. Такой подход не требует пороговой обработки значений после первоначальной кластеризации, что позволяет извлекать соответствующие данные непосредственно из небольшого количества идентификационной информации и создавать классификатор облачной модели, используя метод динамического взвешивания. При этом, использование гибкого обнаружения данных в реальном масштабе времени делает метод взвешивания более точным. Происходящий в дальнейшем процесс обучения корректирует настройки облачной модели, повышая ее способность адаптироваться к изменяющейся сетевой среде.

Относительная близость облаков отражает степень сходства между облаками и полностью выражает случайность и неоднозначность оценки языковых понятий. Предположим, есть два облака $A1(Ex1, En1, He1)$ и $A2(Ex2, En2, He2)$ в дискурсивном пространстве U, при этом определено, что $D_{1,2} = |Ex1 - Ex2|$, тогда $D_{1,2}$ отражает относительную близость двух облаков.

В рамках разработки полууправляемого алгоритма кластеризации данных для защиты конфиденциальной информации опишем процедуру *взвешенного обнаружения вторжений*. Генератор обратного облака получает цифровые характери-

стики облака из реального обучающего набора, формирует правила суждения, реализует нормальное распределение. На практике этот алгоритм требует большого количества обучающих данных и времени обучения [21–22]. Цифровые собственные значения облака, полученные из обучающих данных, не отражают реальной ситуации на момент вторжения, а расчет весов атрибутов слишком субъективен, так же сложно определить пороговое значение при обнаружении вторжений.

Сначала используем для кластеризации набора данных алгоритм обучения с частичным привлечением учителя, затем результаты кластеризации располагаются в порядке возрастания размера кластеров, при этом кластеры предварительно отсеиваются в соответствии с информацией тега.

Поскольку относительная близость облака обладает большей объективностью, авторы ссылаются на эту характеристику для настройки весов атрибутов, предполагая, что «нормальное» облако имеет значение A_1 при обнаружении вторжений, «аномальное» облако имеет значение A_2, тогда при построении модели облака для каждого атрибута измерения размер $D_{1,2}$ отражает степень различия между «нормальными» облаками и «аномальными» облаками. Динамическое взвешивание может в полной мере использовать неявную информацию самих данных, а метод взвешивания является более научным.

Шаги процедуры обнаружения вторжений, основанной на алгоритме обучения с частичным привлечением учителя при кластеризации данных, следующие:

1) используйте алгоритм обучения с частичным привлечением учителя для кластеризации набора данных S;

2) расположите результаты кластеризации в порядке возрастания размеров кластеров;

3) в сочетании с информацией о метках данных, начальные «нормальные» кластеры и «аномальные» кластеры отсеиваются как C_n и C_a соответственно, остальные данные размещаются в C_r;

4) для каждого измерения данных в C_n соответствующее цифровое собственное значение облака $(Exl_i, Enl_i, Hel_i), i = 1, \dots, d$ получается с использованием обратного генератора облаков;

5) для каждого измерения данных в C_a, используйте генератор обратного облака, чтобы получить соответствующее цифровое собственное значение облака $(Ex2_i, En2_i, He2_i), i = 1, \dots, d$;

6) используйте следующую формулу для расчета веса каждого атрибута:

$$w_i = \frac{|Exl_i - Ex2_i|}{\sum_{j=1}^{d} |Exl_i - Ex2_i|};$$

7) возьмите объект данных x из C_r в соответствии с -условием прямого генератора облаков с использованием формулы, рассчитываются модели распределения «аномальных» и «нормальных» облаков:

$$\mu_j = \sum_{i=1}^{d} w_i \cdot \left[\frac{-(x - Exj_i}{2 \cdot Enj_i}\right], j = 1,2;$$

если $\mu_1 > \mu_2$, то x принадлежит к «нормальному» кластеру, обозначим его C_n, после возврата к шагу (4) для обновления «нормальной» облачной модели перейдите к шагу (6) для пересчета веса каждого атрибута, в противном случае, назначьте x из C_a и вернитесь к шагу (5). После обновления «аномальной» облачной модели перейдите к шагу (6), чтобы пересчитать вес каждого атрибута, пока не закончится классификация всех данных.

Опишем результаты вычислительного эксперимента, подтверждающего эффективность предложенного полууправляемого алгоритма кластеризации данных для защиты конфиденциальной информации в сети Интернет.

4. Результаты вычислительного эксперимента. Для проведения вычислительного эксперимента было выбрано 500 записей идентификационных данных, 10000 записей в качестве тестовых данных, среди них 758 DoS атак, 15 атак R2L, 42 атаки U2R и 92 атаки зондирования.

В вычислительном эксперименте использовалась экспериментальная модель на базе MATLAB и набор данных KDDCUP99 (набор данных для 3-го международного конкурса средств обнаружения знаний и интеллектуального анализа данных, который проводился совместно с KDD-99 5-ой международной конференцией по обнаружению знаний и интеллектуальному анализу данных). Эта база содержит стандартный набор данных для аудита, который включает широкий спектр вторжений, смоделированных в сетевой среде, и содержит атрибуты символьных данных, которые не могут быть распознаны MATLAB. Поэтому необходимо перенумеровать значения атрибутов символьного типа и использовать набор натуральных чисел для перенумерации значений, взяв в качестве примера protocol_type – tcp, udp и icmp замены на натуральные числа 1, 2 и 3, соответственно. Таким образом, исходные данные станут числовым типом.

Существует два типа числовых переменных: один представляет собой непрерывную переменную характеристики атрибута, а другой – переменную дискретной характеристики атрибута. Для непрерывных признаковых переменных атрибутов атрибутивные характеристики могут иметь разные метрики, если перед экспериментом данные не были предварительно обработаны.

Для вышеуказанных 10000 записей были использованы алгоритм кластеризации *k*-средних и предложенный полууправляемый алгоритм. Были выполнены тесты данных обнаружения вторжений. Рис. 2 и 3 показывают среднюю частоту обнаружения и частоту ложных срабатываний для алгоритма кластеризации *k*-средних и предложенного полууправляемого алгоритма.

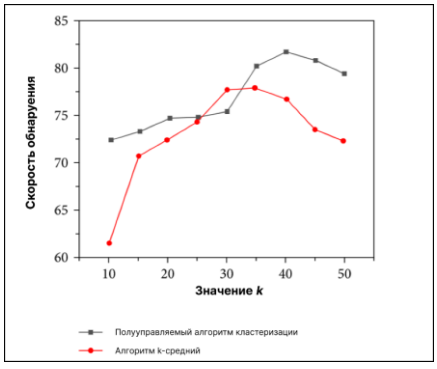

Рис. 2. Сравнение результатов скорости обнаружения

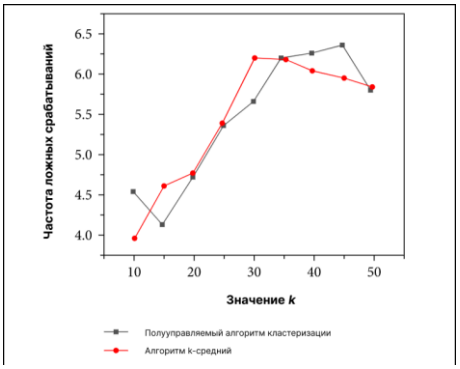

Рис. 3. Сравнение частоты ложных срабатываний

Из рис. 2 и 3 видно, что скорость предложенного алгоритма кластеризации выше, чем у алгоритма k-средних, а также улучшены характеристики по частоте ложных срабатываний алгоритма. Таким образом, предложенный алгоритм кластеризации данных для защиты конфиденциальной информации в сети Интернет повышает устойчивость системы и эффективность решения задачи кластеризации.

Рис. 4 и 5 показывают среднюю частоту обнаружения и частоту ложных срабатываний при различных k-значениях.

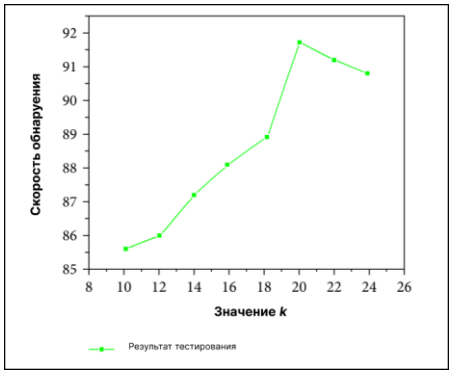

Рис. 4. Скорость обнаружения при разных k-значениях

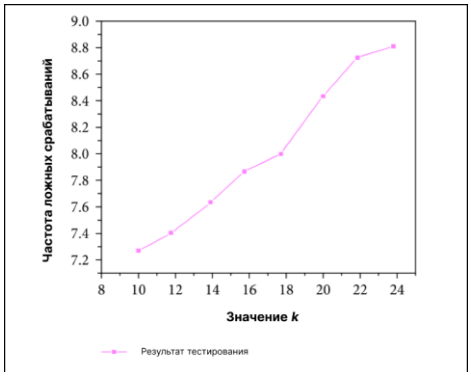

Рис. 5. Частота ложных срабатываний при разных k-значениях

Из экспериментальных результатов, представленных на рис. 4 и 5, видно, что при k постепенном увеличении значения ложная тревога также увеличивается. Однако, когда k=20, скорость обнаружения максимальна. Из этого можно сделать вывод, что при значении k=20 предложенный алгоритм кластеризации, основанный на облачной модели, может обеспечить лучший результат обнаружения вторжений, его уровень обнаружения достигает 91,76%, а уровень ложных тревог составляет 8,54%.

В предложенном авторами алгоритме кластеризации на основе облачной модели частота обнаружения и частота ложных срабатываний при различных значениях k сравниваются с алгоритмом -средних, как показано на рис. 6 и 7.

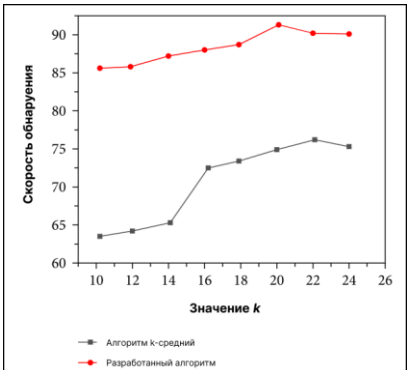

Рис. 6. Сравнение результатов скорости обнаружения при разных значениях k

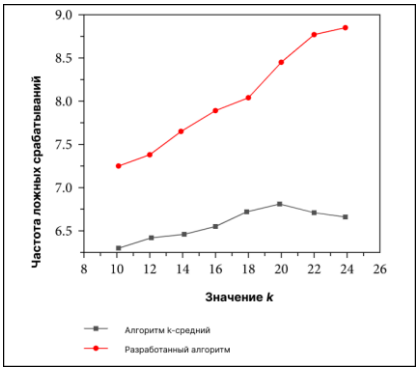

Рис. 7. Сравнение результатов частоты ложных срабатываний при разных значениях k

Как видно из приведенных рисунков, в случае разных значений k предложенный алгоритм значительно эффективней алгоритма k-средних по частоте обнаружения. Частота ложных срабатываний несколько выше, чем у k-средних.

В целом, скорость обнаружения атак у предложенного алгоритма значительно выше, чем у канонических методов, а частота ложных срабатываний незначительно выше, чем у k-средних, это подтверждает повышение эффективности решения поставленной задачи для предложенного алгоритма кластеризации данных для защиты конфиденциальной информации в сети Интернет.

Заключение. В данной статье представлена разработка алгоритма обнаружения вторжений, который использует кластеризацию на основе облачной модели. Предлагаемый алгоритм использует преимущества как маркированных, так и немаркированных образцов для кластеризации данных, тем самым уменьшая зависимость от априорных знаний.

Для оценки эффективности предложенного алгоритма разработано программное приложение и проведен вычислительный эксперимент. Результаты вычислительного эксперимента, проведенного на предложенном алгоритме, сравнивались с несколькими каноническими алгоритмами обнаружения вторжений. Результаты показали, что предложенный алгоритм улучшил производительность системы обнаружения вторжений, повысил точность обнаружения, снизил частоту ложных тревог и усилил надежность системы. Метод динамического взвешивания, используемый в алгоритме, устранил сложность высокоуровневой обработки данных и позволил алгоритму самообучаться, что привело к формированию относительно стабильной облачной модели.

Несмотря на значительное улучшение производительности предложенного алгоритма по сравнению с каноническими алгоритмами кластеризации, результаты исследования также показали, что у алгоритма есть некоторые ограничения, такие как высокий процент ложных срабатываний и чувствительность к данным с определенными видами распределения. Для устранения этих недостатков необходимо дальнейшее усовершенствование алгоритма. В целом, предложенный эвристический алгоритм обнаружения вторжений с кластеризацией на основе облачной модели представляет собой перспективное решение для защиты конфиденциальной информации компьютерных сетей.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Kravchenko Y.A., Bova V.V., Kursitys I.O*. Models for Supporting of Problem-Oriented Knowledge Search and Processing // Intelligent Information Technologies for Industry. – 2016. – Vol. 1. – P. 287-295.
2. *Липинский А.П.* Обеспечение конфиденциальности информации, получаемой при производстве следственных действий // Вестник Удмуртского университета. Серия Экономика и право. – 2021. – Т. 31, № 5. – С. 856-860.
3. *Бабиева Н.А.* Информационная безопасность личности и вопросы защиты конфиденциальной информации // Сборники конференций НИЦ Социосфера. – 2016. – № 31. – С. 66-68.
4. *Кравченко Ю.А., Нацкевич А.Н., Курситыс И.О.* Бустинг биоинспирированных алгоритмов для решения задачи кластеризации // Международная конференция по мягким вычислениям и измерениям. – 2018. – Т. 1. – С. 777-780.
5. *Лобчикова А.С.* Защита конфиденциальной информации при ее передаче по открытым каналам связи // Новая наука: Проблемы и перспективы. – 2017. – Т. 2, № 3. – С. 147-149.
6. *Садуллаев У.Б.* Проблемы защиты конфиденциальной информации // Ростовский научный журнал. – 2019. – № 1. – С. 196-203.
7. *Дуля И.С.* Применение методов глубокого обучения к задаче кластеризации временных рядов // Аллея науки. – 2021. – Т. 1, № 5 (56). – С. 974-978.
8. *Бова В.В., Кулиев Э.В., Щеглов С.Н.* Метод семантической кластеризации распределенных ресурсов знаний с динамическими компонентами на основе контентной фильтрации // Информатика, вычислительная техника и инженерное образование. – 2019. – № 1 (34). – С. 30-42.
9. *Козлова О.А.* Методы кластеризации в задачах оценки технического состояния телекоммуникационного оборудования // Международная конференция по мягким вычислениям и измерениям. – 2014. – Т. 1. – С. 95-96.
10. *Соловьев А.С.* Оценки, анализ, кластеризация и управление в иерархических структурах // Экономика и социум. – 2021. – № 4-2 (83). – С. 404-419.
11. *Бойко Е.А.* Кластеризация социальных сетей с помощью алгоритма кластеризации BSP // Восточно-Европейский журнал передовых технологий. – 2012. – Т. 3, № 11 (57). – С. 34-36.

12. *Giordano J., O'Reilly M., Taylor H., Dogra N.* Confidentiality and autonomy: The challenge(s) of offering research participants a choice of disclosing their identity // Qualitative Health Research. – 2007. – 17. – P. 264-275.
13. *He Z., Cai Z., and Yu J.* Latent-data privacy preserving with customized data utility for social network data // IEEE Transactions on Vehicular Technology. – 2017. – Vol. PP, No. 99. – P. 1-10.
14. *Omran M.G.H., Engelbrecht A.P., Salman A.* An overview of clustering methods // Intelligent Data Analysis. – 2007. – Vol. 11, No. 6. – P. 583-605.
15. *Crotty B.H., Mostaghimi A.* Confidentiality in the digital age // BMJ. – 2014. – Vol. 348.
16. *Алексеев Д.М., Минюк А.Н., Шумилин А.С.* Защита конфиденциальной информации в облачной медицинской информационной системе // Инновационная наука. – 2020. – № 6.
17. *Egoshin N.S. et al.* A Model of Threats to the Confidentiality of Information Processed in Cyberspace Based on the Information Flows Model // Symmetry. – 2020. – Vol. 12, No. 11. – P. 1840.
18. *Livraga G., Viviani M.* Data confidentiality and information credibility in on-line ecosystems // Proceedings of the 11th International Conference on Management of Digital EcoSystems. – 2019. – P. 191-198.
19. *Ge J., Liu J.* Security assessment algorithm of navigation control system based on big data // Journal of coastal research. – 2019. – Vol. 93. – P. 1026-1033.
20. *Deng H.* Multicriteria analysis with fuzzy pairwise comparison // International Journal of Approximate Reasoning. – 1999. – 21 (3). – P. 215-231
21. *Danilowicz C., Nguyen N.* Consensus Methods for Solving Inconsistency of Replicated Data in Distributed Systems // Distributed and Parallel Databases. – 2003. – Vol. 14. – P. 53-69.
22. *Paixao M.P., Silva L. Elias G.* Clustering Large-Scale Distributed Software Component Repositories // Proc. the Fourth Int'l Conf. Advances in Databases Knowledge and Data Applications. – 2012. – P. 124-129.

REFERENCES

1. *Kravchenko Y.A., Bova V.V., Kursitys I.O*. Models for Supporting of Problem-Oriented Knowledge Search and Processing, *Intelligent Information Technologies for Industry*, 2016, Vol. 1, pp. 287-295.
2. *Lipinskiy A.P.* Obespechenie konfidentsial'nosti informatsii, poluchaemoy pri proizvodstve sledstvennykh deystviy [Ensuring the confidentiality of information obtained in the course of investigative actions], *Vestnik Udmurtskogo universiteta. Seriya Ekonomika i pravo* [Bulletin of the Udmurt University. Series Economics and Law], 2021, Vol. 31, No. 5, pp. 856-860.
3. *Babieva N.A.* Informatsionnaya bezopasnost' lichnosti i voprosy zashchity konfidentsial'noy informatsii [Information security of the individual and issues of confidential information protection], *Sborniki konferentsiy NITS Sotsiosfera* [Collections of conferences of the Research Center Sociosphere], 2016, No. 31, pp. 66-68.
4. *Kravchenko Yu.A., Natskevich A.N., Kursitys I.O.* Busting bioinspirirovannykh algoritmov dlya resheniya zadachi klasterizatsii [Boosting bioinspired algorithms for solving the clustering problem], *Mezhdunarodnaya konferentsiya po myagkim vychisleniyam i izmereniyam* [International Conference on Soft Computing and Measurements], 2018, Vol. 1, pp. 777-780.
5. *Lobchikova A.S.* Zashchita konfidentsial'noy informatsii pri ee peredache po otkrytym kanalam svyazi [Protecting confidential information during its transmission over open communication channels], *Novaya nauka: Problemy i perspektivy* [New Science: Problems and Perspectives], 2017, Vol. 2, No. 3, pp. 147-149.
6. *Sadullaev U.B.* Problemy zashchity konfidentsial'noy informatsii [Problems of protection of confidential information], *Rostovskiy nauchnyy zhurnal* [Rostov scientific journal], 2019, No. 1, pp. 196-203.
7. *Dulya I.S.* Primenenie metodov glubokogo obucheniya k zadache klasterizatsii vremennykh ryadov [Application of deep learning methods to the problem of time series clustering], *Alleya nauki* [Alley of Science], 2021, Vol. 1, No. 5 (56), pp. 974-978.
8. *Bova V.V., Kuliev E.V., Shcheglov S.N.* Metod semanticheskoy klasterizatsii raspredelennykh resursov znaniy s dinamicheskimi komponentami na osnove kontentnoy fil'tratsii [The method of semantic clustering of distributed knowledge resources with dynamic components based on content filtering], *Informatika, vychislitel'naya tekhnika i inzhenernoe obrazovanie* [Informatics, Computer Science and Engineering Education], 2019, No. 1 (34), pp. 30-42.

9. *Kozlova O.A.* Metody klasterizatsii v zadachakh otsenki tekhnicheskogo sostoyaniya telekommunikatsionnogo oborudovaniya [Clustering methods in the problems of assessing the technical condition of telecommunication equipment], *Mezhdunarodnaya konferentsiya po myagkim vychisleniyam i izmereniyam* [International Conference on Soft Computing and Measurements], 2014, Vol. 1, pp. 95-96.
10. *Solov'ev A.S.* Otsenki, analiz, klasterizatsiya i upravlenie v ierarkhicheskikh strukturakh [Evaluation, analysis, clustering and management in hierarchical structures], *Ekonomika i sotsium* [Economy and Society], 2021, No. 4-2 (83), pp. 404-419.
11. *Boyko E.A.* Klasterizatsiya sotsial'nykh setey s pomoshch'yu algoritma klasterizatsii BSP [Clustering of social networks using the BSP clustering algorithm], *Vostochno-Evropeyskiy zhurnal peredovykh tekhnologiy* [Eastern European Journal of Advanced Technologies], 2012, Vol. 3, No. 11 (57), pp. 34-36.
12. *Giordano J., O'Reilly M., Taylor H., Dogra N.* Confidentiality and autonomy: The challenge(s) of offering research participants a choice of disclosing their identity, *Qualitative Health Research*, 2007, 17, pp. 264-275.
13. *He Z., Cai Z., and Yu J*. Latent-data privacy preserving with customized data utility for social network data, *IEEE Transactions on Vehicular Technology*, 2017, Vol. PP, No. 99, pp. 1-10.
14. *Omran M.G.H., Engelbrecht A.P., Salman A.* An overview of clustering methods, *Intelligent Data Analysis*, 2007, Vol. 11, No. 6, pp. 583-605.
15. *Crotty B.H., Mostaghimi A.* Confidentiality in the digital age, *BMJ*, 2014, Vol. 348.
16. *Alekseev D.M., Minyuk A.N., Shumilin A.S.* Zashchita konfidentsial'noy informatsii v oblachnoy meditsinskoy informatsionnoy sisteme [Protection of confidential information in a cloud-based medical information system], *Innovatsionnaya nauka* [Innovative Science], 2020, No. 6.
17. *Egoshin N.S. et al.* A Model of Threats to the Confidentiality of Information Processed in Cyberspace Based on the Information Flows Model, *Symmetry*, 2020, Vol. 12, No. 11, pp. 1840.
18. *Livraga G., Viviani M.* Data confidentiality and information credibility in on-line ecosystems, *Proceedings of the 11th International Conference on Management of Digital EcoSystems*, 2019, pp. 191-198.
19. *Ge J., Liu J.* Security assessment algorithm of navigation control system based on big data, *Journal of coastal research*, 2019, Vol. 93, pp. 1026-1033.
20. *Deng H.* Multicriteria analysis with fuzzy pairwise comparison, *International Journal of Approximate Reasoning*, 1999, 21 (3), pp. 215-231
21. *Danilowicz C., Nguyen N.* Consensus Methods for Solving Inconsistency of Replicated Data in Distributed Systems, *Distributed and Parallel Databases*, 2003, Vol. 14, pp. 53-69.
22. *Paixao M.P., Silva L. Elias G.* Clustering Large-Scale Distributed Software Component Repositories, *Proc. the Fourth Int'l Conf. Advances in Databases Knowledge and Data Applications*, 2012, pp. 124-129.

Статью рекомендовал к опубликованию к.т.н., доцент С.Г. Буланов.

Берешполов Игорь Сергеевич – Южный федеральный университет; e-mail: bereshpolov@sfedu.ru; г. Таганрог, Россия; тел.: 88634371651; кафедра систем автоматизированного проектирования, аспирант.

Кравченко Юрий Алексеевич – e-mail: yakravchenko@sfedu.ru; кафедра систем автоматизированного проектирования; д.т.н.; доцент.

Слепцов Алексей Геннадьевич – e-mail: alslepcov@sfedu.ru; кафедра систем автоматизированного проектирования; аспирант.

Bereshpolov Igor Sergeevich – Southern Federal University; e-mail: bereshpolov@sfedu.ru; Taganrog, Russia; phone: +78634371651; the department of computer aided design, postgraduate.

Kravchenko Yury Alekseevich – e-mail: yakravchenko@sfedu.ru; the department of computer aided design; dr. of eng. sc.; associate professor.

Sleptsov Aleksey Gennadievich – e-mail: alslepcov@sfedu.ru; the department of computer aided design; postgraduate.