ВЫБОР МЕТОДА СЕГМЕНТАЦИИ ДАННЫХ ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ АУТЕНТИФИКАЦИИ ПРИ ИСПОЛЬЗОВАНИИ КЛАВИАТУРНОГО ПОЧЕРКА

М.О. Калмыков, Е.В. Рассказов, И.В. Горбунов
Научный руководитель: с.н.с., к.т.н. И.В. Горбунов
Томский государственный университет систем управления и радиоэлектроники,
Россия, г. Томск, пр. Ленина, 40, 634050
E-mail: kmo.azure@gmail.com, rev7.azure@gmail.com, giv@keva.tusur.ru

CHOSING METHOD OF THE KEYSTROKING DATA SEGMENTATION FOR INCREASING AUTHENTIFICATION ACCURACY

M.O. Kalmykov, E.V. Rasskazov, I.V. Gorbunov
Scientific Supervisor: Senior Reseacher, Ph.D. I.V. Gorbunov
Tomsk State University of Control Systems and Radioelectronics,
Russia, Tomsk, Lenin str., 40, 634050
E-mail: kmo.azure@gmail.com, rev7.azure@gmail.com, giv@keva.tusur.ru

Abstract. *The main advantages of the dynamic authentication type like the keystroking were described in the work. The necessity to apply data segmentation before building new and upgrading existing user keystroking profiles were proved. The main advantages of the data segmentation algorithms were shown, and the results of the analysis of the work applicability to the keystroking data were presented.*

Введение. Распространённые на текущий момент методы аутентификации успешно обходятся. Например, смс с кодом уведомления достаточно просто прочитать на любом современном телефоне. Табличные схемы достаточно надежны, но как правило сервисы предлагают скинуть доступ по ней с помощью все того же смс. Аппаратные ключи отделимы от человека и могут быть украдены. Использование отпечатков пальцев требует специального датчика, более того человек в ходе своей жизни часто оставляет свои отпечатки на поверхностях и в случае получения образа злоумышленниками отпечаток данного пальца больше не следует использовать для аутентификации в других системах [1].

Исходя из всех вышеперечисленных недостатков различных методов аутентификации, следует, что данные методы ненадёжны, и, поэтому, в качестве альтернативы был рассмотрена аутентификация на основе клавиатурного почерка (КП). Данный подход не требует специального оборудования для снятия проверяемых характеристик. Метод уникальный в своём роде, однако, существует вероятность ошибок первого и второго рода. Вследствие постоянной вялотекущей изменчивости характеристик КП человека требуется периодическое обновление профиля КП, что в свою очередь повышает требования к ограниченным вычислительным ресурсам, особенно при растущем количестве активных пользователей, а как следствие ведущем к быстрому росту количества данных о вводе каждого пользователя. Сегментация данных КП позволит уменьшить вероятность ошибок первого и второго рода, а также увеличит быстродействие системы при построении и обновлении профилей КП [2].

Экспериментальная часть. На данный момент существует огромное множество методов сегментации данных, ниже представлена наиболее широко используемая классификация данных

методов [3–5]: вероятностные – предполагается, что каждый рассматриваемый объект относится к одному из k классов (k-средние; EM-алгоритм; другие); методы, основанные на системах искусственного интеллекта - весьма условная группа, так как методов очень много и методически они весьма различны (c-средних; нейронная сеть Кохонена; генетический алгоритм; другие); теоретико-графовые (алгоритм выделения связных компонент; алгоритм минимального покрывающего дерева; другие); и другие (алгоритмы семейства KRAB; DBSCAN; и т.д.).

В качестве рассматриваемых алгоритмов были выбраны основные методы, а также их достоинства и недостатки указанны в таблице 1. В качестве критериев для сравнения были выбраны наиболее важные критерии для выбора метода сегментации данных КП пользователя, к ним относятся: вычислительная сложность алгоритма сегментации, размер формируемого кластера, а также его форма. В таблице использованы следующие обозначения: n – объём сегментируемых данных, k – количество кластеров, l – число итераций, m – общее количество нейронов в слое, p – размерность входного пространства.

Таблица 1

Выявление основных достоинств и недостатком выбранных алгоритмов кластеризации

Алгоритм	Вычислительная сложность	Размеры кластеров	Форма кластера
Ближайший сосед	$O(n^2)$	Произвольные	Произвольная
Наиболее удалённый сосед	$O(n^2)$	Произвольные	Произвольная
k-средние	$O(n * k * l)$	Примерно равные	Гиперсфера
c-средние	$O(n * k * l)$	Примерно равные	Гиперсфера
Самоорганизующиеся карты Кохонена	$O(l * n * p) + O(l * m * n) + O(m^2)$	Примерно равные	Гиперсфера
Метод Уорда	$O(n * k * \log n)$	Примерно равные	Гиперсфера

В результате анализа основных достоинств и недостатков алгоритмов кластеризации, рассмотренных в рамках применимости данных алгоритмов к сегментации данных КП, был выбран метод Уорда. Достоинствами данного алгоритма являются: высокая скорость работы метода, при чём, как показала практика, данный алгоритм «разгоняется» в процессе своей работы (чем меньше векторов остаётся разнести по кластерам, тем растёт его скорость работы), также к достоинствам алгоритма относится, что алгоритм старается создать кластера одинакового размера, и кластера имеют форму гиперсферы, что также является достоинством, применимо к сегментации данных КП.

При реализации данного модуля сегментации данных, на вход данного алгоритма, в качестве сегментируемых данных, подавались данные пользователей, которые входили в систему безопасного облачного хранения аутентификационных данных «Lockout» [6]. Векторы данных, подаваемые на вход модуля, имеют вид:

$$L_i = \{a_1, c_1, a_2, c_2, \dots, a_n, c_n, full, min, max, click\}$$

где: i – номер вектора КП пользователя, $i = 0, 1, 2, \dots, n$, L_i – вектор КП пользователя, a_j – время нажатия на клавиатуру, c_j – время удержания клавиши, Время a_j и c_j считаются относительно a_1, $full$ – полное время набора кодового слова, min – время между нажатием на близлежащие клавиши, max – время между

нажатием на самые отдалённые клавиши, *click* – количество нажатий на экран/клавиатуру после набора кодового слова и до момента нажатия на кнопку «Вход». При расчёте переменных *min* и *max*, клавиатура пользовательского устройства представлялась в виде координатный плоскости, где каждая клавиша имеет координату. Так, например, при разборе кодовой фразы «НИР3», наибольшее расстояние будет между буквой «Р» и цифрой «3», а минимальное, между «И» и «Р».

В качестве выходных данных, является набор векторов пользователей с соответствующими номерами кластеров. На рис. 1 представлена модель чёрного ящика кластеризации данных, по методу Уорда, КП пользователей при входах в систему.

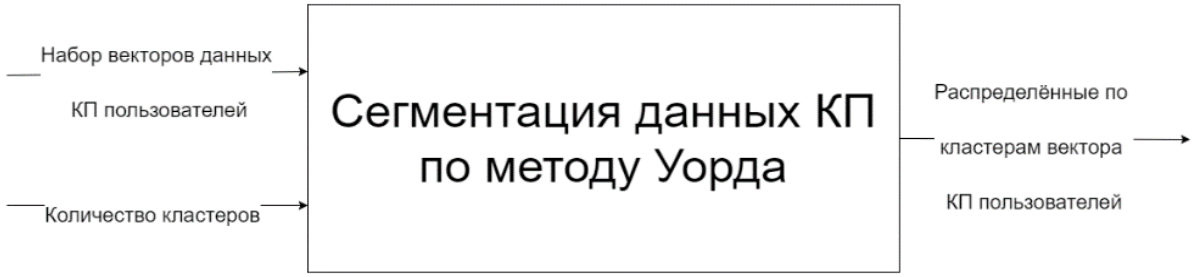

Рис. 1. Модель чёрного ящика кластеризации данных КП, по методу Уорда

Заключение. В результате практического применения данного алгоритма, реализованного в качестве модуля сегментации данных КП в системе безопасного хранения аутентификационных данных «Lockout», было выявлено, что при внедрении данного модуля сегментации, время работы алгоритма формирования профилей пользователей с их портативных устройств, уменьшилось в 2,437 раза.

СПИСОК ЛИТЕРАТУРЫ

1. Обмануть сканер отпечатков можно при помощи обычного струйного принтера [Электронный ресурс]. – Режим доступа: https://xakep.ru/2016/03/09/2d-printed-fringerprints/. – 13.09.2016.
2. Агурьянов И. Клавиатурный почерк как средство аутентификации [Электронный ресурс]. – Режим доступа: http://www.securitylab.ru/blog/personal/aguryanov/29985.php. – 10.10.2016.
3. Liu H., Motoda H. Instance selection and construction for data mining. – Dordrecht: Springer Science + Business Media, 2001. – 432 p.
4. Gan G., Ma C., Wu J. Data Clustering Theory, Algorithms, and Applications. – Philadelphia: SIAM, 2007. – 488 p.
5. Fasulo D. An Analysis of Recent Work on Clustering Algorithms // Technical report. – 1999. – No. 01-03-02. – P. 1–23.
6. Lockout [Электронный ресурс]. – Режим доступа: https://play.google.com/store/apps/details?id=com.lockout.keys&hl=ru. – 20.12.2016.