УДК 004.93'14

**Ю.И. Еременко, Ю.С. Олюнина**

Старооскольский технологический институт им. А.А. Угарова
(филиал НИТУ «МИСиС»), Старый Оскол, Россия

# ОБ ОПРЕДЕЛЕНИИ МЕТОДА ОБРАБОТКИ ПОТОКА ДАННЫХ С ЦЕЛЬЮ ВЫЯВЛЕНИЯ СКРЫТЫХ ХАРАКТЕРИСТИК КЛАВИАТУРНОГО ПОЧЕРКА

Данная статья посвящена вопросу обработки статистических данных, полученных при анализе клавиатурного почерка пользователя информационной системы. Рассмотрены основные характеристики клавиатурного почерка. Приведены обзор и сравнительный анализ методов обработки потока данных с целью выявления скрытых зависимостей клавиатурного почерка. Для решения данной задачи предлагается применение как классических статистических методов, например факторного анализа, так и методов искусственного интеллекта – искусственных нейронных сетей и иммунных сетей.

**Ключевые слова:** клавиатурный почерк, идентификация пользователя, биометрические характеристики, искусственные нейронные сети.

**Iu.I. Eremenko, Iu.S. Oliunina**

Stary Oskol Technological Institute named after A.A. Ugarov
(Branch of National University of Science and Technology "MISiS"),
Stary Oskol, Russian Federation

# ON METHOD SEARCH TO PROCESS DATA TO IDENTIFY HIDDEN CHARACTERISTICS OF KEYSTROKE PATTERN

This article scope is to process statistical data obtained by analyzing the information system user keystroke pattern. The main characteristics of the keystroke pattern are considered. The research gives an overview and comparative analysis of data flow processing methods if the purpose is to reveal hidden dependencies of the keystroke pattern. Solving this problem, it is proposed to use both classical statistical methods, for example, factor analysis, and such artificial intelligence methods as artificial neural networks and immune networks.

**Keywords:** keystroke pattern, user authentication, biometrics, artificial neural networks.

## Введение

Наиболее активно развивающейся угрозой безопасности информационных систем является угроза утечки данных, которая для современных компаний возрастает пропорционально интенсивности использования информационных технологий (ИТ). Как показывает статистика [1],

большинство угроз корпоративной сети являются именно внутренними, т.е. исходят от сотрудников компании.

Если воздействие внешних угроз можно контролировать с помощью таких аппаратно-технических мер, как защита каналов передачи данных, антивирусная защита внешних веб-ресурсов организации, контентная фильтрация трафика на наличие вредоносного программного обеспечения (ПО) и т.д., то отследить факты появления внутренних угроз гораздо сложнее. Обеспечение эффективного контроля доступа сотрудников к системе позволит повысить надежность защиты данных корпоративной системы и, как следствие, сократить финансовые потери [2].

Среди программно-аппаратных средств защиты перспективным направлением является использование динамических биометрических характеристик для идентификации, преимуществами которых являются возможность скрытой идентификации, неотделимость биометрической характеристики от владельца и крайняя сложность подделки.

Для решения данной задачи может быть применена идентификация пользователя по клавиатурному почерку. Однако использование биометрических методов является сложной задачей с точки зрения обработки полученных статистических данных в силу того, что они зависят от психофизиологического состояния человека и могут меняться с возрастом. В связи с этим актуален вопрос выбора метода обработки образцов почерка с целью установления личности их владельца.

## 1. Постановка задачи

В данной работе рассматриваются основные характеристики клавиатурного почерка как динамической биометрической характеристики, а также производится анализ существующих методов обработки потока данных с целью выявления характеристик клавиатурного почерка. Необходимо определить, какой из методов позволит наиболее эффективно осуществлять идентификацию пользователя по клавиатурному почерку.

## 2. Характеристики клавиатурного почерка

Идентификация пользователя по клавиатурному почерку является наименее затратным способом, поскольку «не требует установки специальных аппаратных средств, не нуждается в сопровождении, яв-

ляется прозрачным для конечного пользователя, т.е. не причиняет ему неудобств, и позволяет проводить скрытую аутентификацию» [3]. Рассматривая клавиатурный почерк (КП) как биометрическую характеристику, следует отметить его основные параметры: время удержания нажатой клавиши и интервалы времени между нажатиями. Однако на сегодняшний день выделены и другие параметры КП, описанные в работе [3]: общее время набора парольной фразы, частота возникновения ошибок при наборе, факт использования дополнительных клавиш (использование числовой клавиатуры), особенности ввода заглавных букв (использование клавиши Shift или CapsLock) и т.д.

Кроме того, в работе [4] автор рассматривает следующие характеристики КП:

1) число символов – чистый размер текста без учета символов, удаленных при помощи BackSpace;

2) общее время – считается от момента нажатия первой клавиши до момента нажатия последней;

3) min пауза – минимальная пауза между нажатиями;

4) max пауза – максимальная пауза между нажатиями;

5) интервалы между нажатиями клавиш – средняя пауза между нажатиями;

6) среднее время удержания – резкость нажатия, показывает среднее время между нажатием и отпусканием клавиши;

7) скорость spm – количество набранных символов в минуту;

8) скорость нетто – чистая скорость набора текста, считается для всех неудаленных символов текста;

9) скорость wpm – количество символов в минуту; в англоязычных странах скорость считается именно в этих единицах, причем длина «слова» всегда равна 5 символам, иначе говоря, это скорость нетто, деленная на 5;

10) скорость брутто – скорость набора с учетом удаленных символов, позволяет оценить потери скорости в связи с неправильным вводом;

11) скорость брутто+ – скорость набора с учетом удаленных символов и нажатий BackSpace, позволяет оценить потери скорости, связанные с неправильным вводом и его исправлением;

12) скорость брутто* – при расчете этой скорости не учитывают нажатия ошибочно введенных символов, клавиши BackSpace, а также

время, затраченное на эти нажатия; позволяет оценить скорость при наборе данного текста, если бы ошибок не было вовсе;

13) потери от исправлений – показывает в процентном соотношении, насколько падает скорость из-за времени, потраченного на совершение ошибок и их исправление;

14) степень аритмичности при наборе – степень неравномерности набора в процентах, среднее отклонение паузы между нажатиями от среднего значения (в скобках отображают значение аритмии для участков текста, набранных без ошибок);

15) число исправлений – число символов, удаленных при помощи клавиши BackSpace;

16) процент серий исправлений – каждая группа подряд удаленных символов считается за одно исправление;

17) max без исправлений – размер максимального фрагмента текста, набранного без исправлений (в скобках обозначают значение в процентах по отношению к общему размеру текста);

18) число перекрытий между клавишами – показывает число перекрытий клавиш (кнопка не опущена, но нажата уже другая).

С помощью бесплатного программного обеспечения Typing Statistics авторами были получены параметры набора текста. Графики распределения их значений представлены на рис. 1.

Также были получены числовые значения данных параметров. Они представлены на рис. 2.

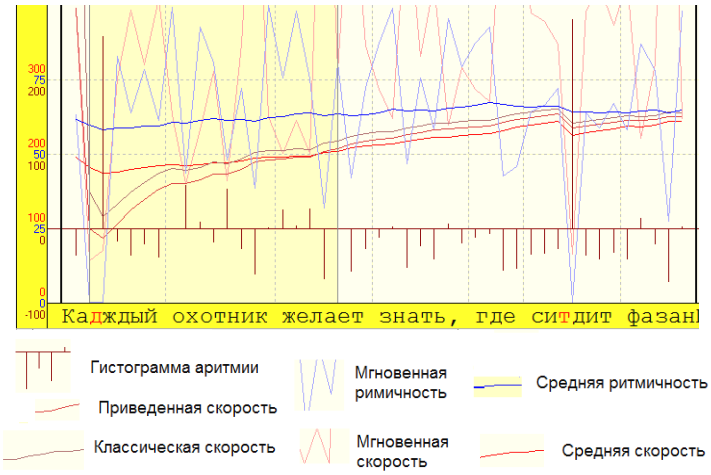


Рис. 1. Графики распределения характеристик КП

| Параметр | Значение |
|---|---|
| Символов | 28 (28) |
| Общее время | 2,729с |
| Средняя пауза | 101,076 мс |
| Среднее время удер... | 110,838 мс |
| Скорость spm | 620,45 (599,05) |
| Скорость wpm | 123,12 (118,72) |
| Скорость нетто | 615,60 (593,61) |
| Скорость брутто* | 615,60 (593,61) |
| Потери от исправле... | 0,00% |
| Аритмия | 30,35% (30,35)% |
| Исправлений | 0 (0,00%) |
| Серий исправлений | 0 (0,00%) |
| max без исправлений | 28 (100,00%) |

Рис. 2. Числовые значения параметров КП

На основании вышеизложенного можно сделать вывод, что выбор метода обработки потока данных и выявления скрытых характеристик в нем для идентификации пользователя является сложной многофакторной задачей, требующей тщательного анализа.

### 3. Анализ методов обработки статистических данных

В силу того что количество характеристик ПК достаточно велико, а в режиме скрытой идентификации пользователь будет вводить символы непрерывно, первоочередной задачей является возможное уменьшение размерности исходной выборки методами уменьшения размерности.

Их использование позволяет:

– сократить вычислительные затраты при обработке данных;

– бороться с переобучением;

– сжимать данные для более эффективного хранения информации;

– визуализировать данные, проектируя выборку на двух-, трехмерное пространство;

– извлекать новые признаки.

К способам решения данной задачи относится метод главных компонент – один из наиболее распространенных классических методов, применяемых для обработки больших объемов статистических данных. Идея метода заключается в поиске в исходном пространстве гиперплоскости заданной размерности с последующей проекцией выборки на данную гиперплоскость. При этом выбирается та гиперплоскость, ошибка проецирования данных на которую является минимальной в смысле суммы квадратов отклонений [5].

Однако использование метода главных компонент имеет ряд ограничений. Во-первых, если выборка данных образует скрытую поверхность, которая является нелинейной, метод главных компонент может приводить к неадекватным результатам. Во-вторых, существует возможность определения скрытых компонент только с точностью до афинного преобразования. В-третьих, существует сложность в определении момента остановки процедуры выделения факторов, так как в процессе последовательного выделения факторов они включают в себя все меньше и меньше изменчивости.

Еще одним возможным способом решения задачи является факторный анализ как методика комплексного и системного изучения и измерения факторов на величину результативных показателей [6]. Факторный анализ используется с целью сокращения данных, определения структуры взаимосвязей между переменными, т.е. классификации переменных, отбора факторов, определяющих исследуемые результативные показатели, определения формы зависимости между факторами и результативными показателями и ее моделирования.

Однако использование факторного анализа также имеет ряд ограничений, в частности:

– при использовании факторного анализа исходят из того, что факторы изменяются независимо друг от друга;

– факторный анализ неизбежно сопровождается потерей исходной информации о связях между переменными, и эта потеря часто весьма ощутима – от 30 до 50 %;

– ценность представляет решение, когда группы переменных, которые соответствуют разным факторам, незначительно коррелируют друг с другом;

– не все данные могут быть подвергнуты факторному анализу.

Кроме того, рассматривая статистические методы, следует отметить, что их использование не позволяет в достаточной степени эффективно выявлять зависимости, скрытые в полученных данных.

Для решения данной задачи перспективным является применение методов искусственного интеллекта, так как их важнейшей особенностью является способность решать слабоформализуемые задачи, добиваясь при этом результатов, по эффективности сравнимых с решениями, получаемыми человеком-экспертом. Наиболее подходящими для

решения задачи выявления скрытых характеристик клавиатурного почерка являются иммунные системы и нейронные сети.

Иммунные системы обладают важными характеристиками, такими как уникальность, автономность, распределенное обнаружение и устойчивость к шуму, способность к самоорганизации, способность распознавать образы, которые могут быть использованы для различения чужеродных клеток, поступающих в организм от клеток тела, способность к обучению и принятию решений в незнакомой ситуации [7]. Искусственные иммунные сети используются в области обнаружения аномалий и неисправностей, в системах компьютерной и интернет-безопасности.

Круг задач, решаемых искусственными нейронными сетями, также достаточно широк. В частности, это распознавание и классификация, кластеризация, прогнозирование, идентификация, аппроксимация и интерполяция и т.д.

Применение искусственных нейронных сетей целесообразно в разных областях, и на сегодняшний день существует множество их видов, таких как конкурентные нейронные сети, карты Кохонена, АРТ-2 нейронная сеть, сеть Хопфилда, сеть Элмана, вероятностная нейронная сеть.

В работах [8, 9] отмечаются такие преимущества нейронных сетей, как «способность выявлять скрытые закономерности развития ситуации и зависимости между входными и выходными данными, используя способность обучения на множестве примеров», «возможность построения нелинейных зависимостей, возможность применения для широкого круга задач», «способность работать при наличии большого числа неинформативных, шумовых входных сигналов» и т.д.

Таким образом, для решения задачи идентификации пользователя по клавиатурному почерку предложено использовать аппарат искусственных сетей, так как он по сравнению с классическими статистическими методами не использует усредненные значения временных параметров, характеризующих манеру работы пользователя с клавиатурой. Кроме того, по сравнению с другими методами искусственного интеллекта нейронные сети обладают способностью выявлять скрытые закономерности зависимости в потоке данных, обучаться на множестве примеров, работать при сильном зашумлении входных сигналов, обеспечивать более высокое быстродействие за счет распараллеливания процесса обработки данных и т.д.

## Заключение

Идентификация пользователя по клавиатурному почерку позволит осуществлять контроль доступа к системе в непрерывном скрытом режиме без использования дополнительных аппаратных средств. Задачу обработки статистических данных, полученных в результате анализа КП, предложено решать с помощью аппарата искусственных нейронных сетей, которые позволяют выявлять скрытые зависимости и зависимости во входных данных, а также способны обучаться на множестве примеров.

## Список литературы

1. Аналитический центр InfoWatch [Электронный ресурс]. – URL: www.infowatch.ru/analytics (дата обращения: 16.04.2016).

2. Еременко Ю.И., Олюнина Ю.С. Идентификация пользователя по его клавиатурному почерку // Современные проблемы горно-металлургического комплекса. Наука и производство: материалы XII Всерос. науч.-практ. конф. с междунар. участием / НИТУ «МИСиС». – Старый Оскол, 2015. – С. 147–151.

3. Шарипов Р.Р. Разработка полигауссового алгоритма аутентификации пользователей в телекоммуникационных системах и сетях по клавиатурному почерку: дис. … канд. техн. наук: 05.12.13. – Казань, 2006. – 135 с.

4. Тушканов Е.В. Разработка методов и алгоритмов повышения защищенности информации на основе анализа клавиатурного почерка: автореф. дис. ... канд. техн. наук: 05.13.19 / С.-Петерб. нац. исслед. ун-т информ. технологий, механики и оптики. – СПб., 2016. – 19 с.

5. Ветров Д.П., Кропотов Д.А., Осокин А.А. Автоматическое определение количества компонент в EM-алгоритме восстановления смеси нормальных распределений // Журнал вычислительной математики и математической физики. – 2010. – Т. 50, № 4. – С. 770–783.

6. Такахаси С. Факторный анализ. – М.: ДМК Пресс, 2015. – 146 с.

7. Брюховецкий А.А., Скатков А.В. Применение моделей искусственных иммунных систем для решения задач многомерной оптимизации // Оптимізація виробничих процесів. – 2010. – № 7. – С. 119–122.

8. Костюченко Е.Ю. Идентификация непрерывных биометрических сигналов на основе нейронных сетей: автореф. дис. … канд. техн. наук: 05.13.01. – Томск, 2010. – 24 с.

9. Хайкин С. Нейронные сети: полный курс. – 2-е изд. – М.: Вильямс, 2006. – 1104 с.

## References

1. Analiticheskii tsentr InfoWatch [Analytic center InfoWatch], available at: www.infowatch.ru/analytics (accessed: 16 April 2016).

2. Eremenko Iu.I., Oliunina Iu.S. Identifikatsiia pol'zovatelia po ego klaviaturnomu pocherku [User identification by keyboard handwriting]. *Sovremennye problemy gorno-metallurgicheskogo kompleksa. Nauka i proizvodstvo: materialy XII Vserossiiskoi nauchno-prakticheskoi konferentsii s mezhdunarodnym uchastiem*. Staryi Oskol STI NITU «MISiS», 2015, pp. 147-151.

3. Sharipov R.R. Razrabotka poligaussovogo algoritma autentifikatsii pol'zovatelei v telekommunikatsionnykh sistemakh i setiakh po klaviaturnomu pocherku [Design of the poly-Gaussian algorithm of authentication by keyboard handwriting of telecommunication networks and systems user]: Ph. D. thesis. Kazan', 2006, 135 p.

4. Tushkanov E.V. Razrabotka metodov i algoritmov povysheniia zashchishchennosti informatsii na osnove analiza klaviaturnogo pocherka [The development of methods and algorithms to increase information security based on keyboard handwriting analysis]: Abstract of Ph. D. thesis. Saint Petersburg, Sankt-Peterburgskii natsional'nyi issledovatel'skii universitet informatiki tekhnologii, mekhaniki i optiki, 2016, 19 p.

5. Vetrov D.P., Kropotov D.A., Osokin A.A. Avtomaticheskoe opredelenie kolichestva komponent v EM-algoritme vosstanovleniia smesi normal'nykh raspredelenii [Automatic determination of the number of components in the EM algorithm of restoration of a mixture of normal distributions]. *Computational Mathematics and Mathematical Physics*, 2010, vol. 50, iss. 4, pp. 770-783.

6. Takakhasi S. Faktornyi analiz [Factor analysis]. Moscow, DMK Press, 2015, 146 p.

7. Briukhovetskii A.A., Skatkov A.V. Primenenie modelei iskusstvennykh immunnykh sistem dlia resheniia zadach mnogomernoi optimizatsii [Application of artificial immune systems for solution multiattributive optimization problems]. *Optimizatsiia proizvodstvennykh protsessov*, 2010, no. 7, pp. 119-122.

8. Kostiuchenko E.Iu. Identifikatsiia nepreryvnykh biometricheskikh signalov na osnove neironnykh setei [Iidentification of continuous biometric signals based on neural networks]: Abstract of Ph. D. thesis. Tomsk, 2010, 24 p.

9. Khaikin S. Neironnye seti: polnyi kurs [Neural networks, full course]. Moscow, Vil'iams, 2006, 1104 p.

Получено 20.06.2017

## Об авторах

**Еременко Юрий Иванович** (Старый Оскол, Россия) – доктор технических наук, профессор, декан факультета автоматизации и информационных технологий, заведующий кафедрой автоматизированных и информационных систем управления, Старооскольский технологический институт им. А.А. Угарова (филиал НИТУ «МИСиС») (e-mail: erem49@mail.ru).

**Олюнина Юлия Сергеевна** (Старый Оскол, Россия) – аспирант кафедры «Автоматизированные и информационные системы управления», Старооскольский технологический институт им. А.А. Угарова (филиал НИТУ «МИСиС») (e-mail: julijasergeevna@mail.ru).

## About the authors

**Iurii I. Eremenko** (Stary Oskol, Russian Federation) – Doctor of Technical Sciences, Professor, Dean of the Automation and Information Technologies, Head of the Department of Automation and Information Control Systems, Stary Oskol technological institute named after A.A. Ugarov (Branch of National University of Science and Technology "MISiS") (e-mail: erem49@mail.ru).

**Iuliia S. Oliunina** (Stary Oskol, Russian Federation) – Postgraduate Student, Department of Automation and Information Control Systems, Stary Oskol Technological Institute named after A.A. Ugarov (Branch of National University of Science and Technology "MISiS") (e-mail: julijasergeevna@mail.ru).