

Применение метода k -средних для идентификации пользователя по клавиатурному почерку

Ю. И. Еременко, Ю. С. Олюнина

*Старооскольский технологический институт (филиал) ННТУ «МИСиС»
Старый Оскол, Россия*

Аннотация

Рассматривается проблема защиты от несанкционированного доступа к данным путем идентификации пользователей по биометрическим характеристикам, а именно по клавиатурному почерку. Для решения этой задачи авторами была проведена серия опытов для получения статистической выборки образцов клавиатурного почерка, которые используются в качестве биометрической характеристики для установления личности ее владельца. Для анализа данных и определения автора конкретного образца почерка в работе рассматривается метод k -средних, являющийся одним из наиболее простых и эффективных статистических методов классификации в случае, когда число кластеров заранее известно. Оценка эффективности предложенного метода для решения поставленной задачи оценивается с помощью коэффициентов ложного доступа и ложного отказа в доступе, которые являются основными характеристиками биометрических систем аутентификации. Полученные результаты позволяют сделать вывод о наличии ряда ограничений при использовании данного метода, возникающих в силу того, что данная задача является слабоформализуемой, зависящей от множества факторов, в том числе и не поддающихся математическому описанию, таких как нестабильность клавиатурного почерка, объясняемая изменением психофизиологического состояния пользователя, эргономичностью клавиатуры и т. д. Учитывая упомянутые особенности, дальнейшее решение задачи предлагается рассматривать в перспективе использования методов, основанных на интеллектуальной обработке данных, которые позволяют обнаруживать в потоке данных скрытые закономерности и зависимости.

Ключевые слова

идентификация пользователя, клавиатурный почерк, биометрические характеристики, метод k -средних

Для цитирования

Ерёменко Ю. И., Олюнина Ю. С. Применение метода k -средних для идентификации пользователя по клавиатурному почерку // Вестник ННТУ. Серия: Информационные технологии. 2019. Т. 17, № 2. С. 30–38. DOI 10.25205/1818-7900-2019-17-2-30-38

Using the K-means Method to Identify a User by Keyboard Handwriting

Yu. I. Eremenko, Yu. S. Olyunina

*Stary Oskol Technological Institute (branch) NUST "MISiS"
Stary Oskol, Russian Federation*

Abstract

This article discusses the problem of protection against unauthorized access to data by identifying users by biometric characteristics – keyboard handwriting. To do this, the authors conducted a series of experiments to obtain a set of samples of keyboard handwriting, which are used as a biometric characteristic to identify the owner. To analyze the data and determine the author of a particular handwriting sample, the authors consider the k -means method. This method is one of the simplest and most effective statistical classification methods when the number of clusters is known in advance. To evaluate the effectiveness of this method for solving this problem, the authors propose using the coefficients of false access and false access denial, which are the main characteristics of biometric authentication systems. The results suggest that there are some limitations when using this method. They are related to the fact that this problem is poorly formalized and depends on many factors that can not be mathematically described. For

example, the instability of the keyboard handwriting, which is due to changes in the psycho-physiological state of the user, the ergonomics of the keyboard and others. Given these features, the authors propose to solve the problem using methods based on intelligent data processing, which allow to detect hidden patterns and dependencies in the data flow.

Keywords

user identification, keyboard handwriting, biometric characteristics, the method k -means

For citation

Eremenko Yu. I., Olyunina Yu. S. Using the K -means Method to Identify a User by Keyboard Handwriting. *Vestnik NSU. Series: Information Technologies*, 2019, vol. 17, no. 2, p. 30–38. (in Russ.) DOI 10.25205/1818-7900-2019-17-2-30-38

Введение

Одной из наиболее активно возрастающих угроз безопасности информационных систем является проблема утечки данных. Для современных организаций она становится все более актуальной вместе с ростом интенсивности использования информационных технологий. Значительную опасность представляют угрозы, возникающие при обработке конфиденциальной информации в организациях, и в первую очередь это угроза несанкционированного доступа к данным¹.

Аналитики центра InfoWatch, занимающегося изучением вопросов защиты от утечек информации, отмечают, что лишь в 22 % случаев утечка информации происходила в результате хакерской активности, в большинстве случаев (75 %) информация утекала по вине внутреннего нарушителя². Отследить факты появления внутренних угроз гораздо сложнее, поскольку в отличие от внешних угроз их нельзя достаточно точно контролировать с помощью аппаратно-технических мер. На сегодняшний день одним из наиболее простых и распространенных средств защиты данных является парольная идентификация.

В работе [1] отмечаются основные недостатки данного метода: неоднозначность идентификации оператора ключевой системы, возможность обмана системы защиты, например путем кражи или взлома пароля, «невозможность обнаружения подмены законного авторизованного пользователя. В данном случае злоумышленник может нанести вред обрабатываемой КС информации, когда оператор оставляет без присмотра КС с пройденной процедурой авторизации» [1].

В качестве перспективного варианта решения проблемы возможно применение биометрических систем идентификации пользователя. Актуальность выбора данного метода описана в работах [2; 3]. Такой подход имеет ряд преимуществ по сравнению с другими средствами идентификации, описанными в работах [1; 3–5] и др. На сегодняшний день биометрическая идентификация личности является одним из перспективных направлений в сфере информационной безопасности и характеризуется такими преимуществами, как неотделимость биометрической характеристики от владельца и крайняя сложность подделки [2]. Кроме того, в отличие от парольных средств защиты, которые человек может непреднамеренно или по своему умыслу сообщить злоумышленнику, или от различных карт и ключей, которые могут быть утеряны или украдены, биометрические характеристики позволяют однозначно идентифицировать человека [6].

В рамках данной работы предложено использование клавиатурного почерка, поскольку из всех биометрических методов защиты информации данный метод идентификации является самым простым для внедрения и наименее затратным [4], не требует установки специальных аппаратных средств, не нуждается в сопровождении, является прозрачным для конечного пользователя [5], т. е. не причиняет ему неудобств и позволяет проводить скрытую аутентификацию.

Анализ работ по данной теме показал, что для классификации образцов клавиатурного почерка с целью установления личности его владельца и защиты данных компьютерной сис-

¹ Аналитический центр InfoWatch. URL: www.infowatch.ru/analytics/ (дата обращения 21.11.2018).

² Там же.

темы от несанкционированного доступа наибольшее распространение получила обработка данных с использованием методов математической статистики и теории вероятностей либо методы, используемые в совокупности с дополнительными механизмами защиты (парольной защитой либо использованием дополнительных аппаратных средств). Необходимость двухфакторной идентификации обусловлена тем, что значения коэффициентов ложного доступа и ложного отказа в доступе, являющихся основными критериями надежности биометрических систем, достаточно высоки. Использование методов математической статистики «основано на анализе усредненных значений временных параметров, характеризующих манеру работы пользователя с клавиатурой, поэтому результаты, полученные при их применении, не позволяют произвести процедуру идентификации достаточно эффективно» [6].

В рамках данной работы рассмотрено использование альтернативного метода обработки данных, такого как метод k -средних, который является одним из наиболее простых и эффективных методов машинного обучения для решения задач классификации, характеризуется простотой реализации и возможностью распространения полученных решений на новые наблюдения [7]. Кроме того, данный метод не требует вычисления и хранения матрицы расстояний и позволяет распараллеливать процесс вычислений.

Постановка задачи

Задачей данного исследования является необходимость оценить эффективность использования метода k -средних для решения задачи идентификации пользователя по клавиатурному почерку и сделать дальнейшие выводы о возможности его применения или необходимости модификации с целью получения максимального значения целевой функции, характеризующей определение вероятности принадлежности образца почерка конкретному пользователю [8].

Описание предмета исследования

Клавиатурный почерк представляет собой набор характеристик работы пользователя на клавиатуре и характеризуется множеством параметров, описанных в работах [3; 5].

На основании описанных характеристик была произведена серия опытов по их регистрации. Для их проведения были использованы фразы-панграммы, включающие в себя все буквы алфавита, в частности, такие как: «съешь ещё этих мягких французских булок, да выпей же чаю»; «в чащах юга жил-был цитрус... – да, но фальшивый экземпляр» и т. п.

Результаты проведенных экспериментов представлены в виде графиков. Для более наглядного представления результатов на графиках отображена только часть набранных фраз, так, например, на рис. 1 представлена динамика написания слова «каждый», которая является схожей во всех проведенных сериях опытов. Это утверждение верно и при наборе каждой фразы целиком. На данном графике по оси ординат обозначено время удержания нажатой клавиши, на оси абсцисс – соответствующая буква. Опыт проводился одним и тем же пользователем 10 раз.

Кроме того, на рис. 1 видно, что числовые характеристики почерка каждого человека остаются схожими вне зависимости от количества опытов, но в то же время почерк отдельного человека очевидно отличим от других.

На основании полученных результатов можно сделать вывод о том, что клавиатурный почерк является достаточно информативной биометрической характеристикой, и его использование возможно для определения личности владельца почерка.

Однако клавиатурный почерк как динамическая (поведенческая) характеристика пользователя подвержен влиянию множества факторов, что делает его весьма нестабильным. Основные сложности при работе с клавиатурным почерком связаны с большим разнообразием навыков набора текста у пользователей. Процедура идентификации может быть неэффективной или невозможной для лиц, не имеющих стойких навыков работы с клавиатурой. Из тео-

рии машинописи и делопроизводства установлено, что срок, необходимый для формирования устоявшегося клавиатурного почерка, составляет не менее 6 месяцев. Кроме того, на характеристики набора текста влияют факторы, обусловленные психологическим состоянием человека: сонливость, тревога, усталость. Не менее значимыми являются и другие объективные причины, например травма кисти или пальца, использование устройств ввода нестандартного размера, обладающих другой эргономичностью [1]. Все эти факторы снижают достоверность идентификации. В связи с этим выбор метода обработки параметров клавиатурного почерка является сложной задачей, требующей тщательного анализа используемых методов обработки данных.

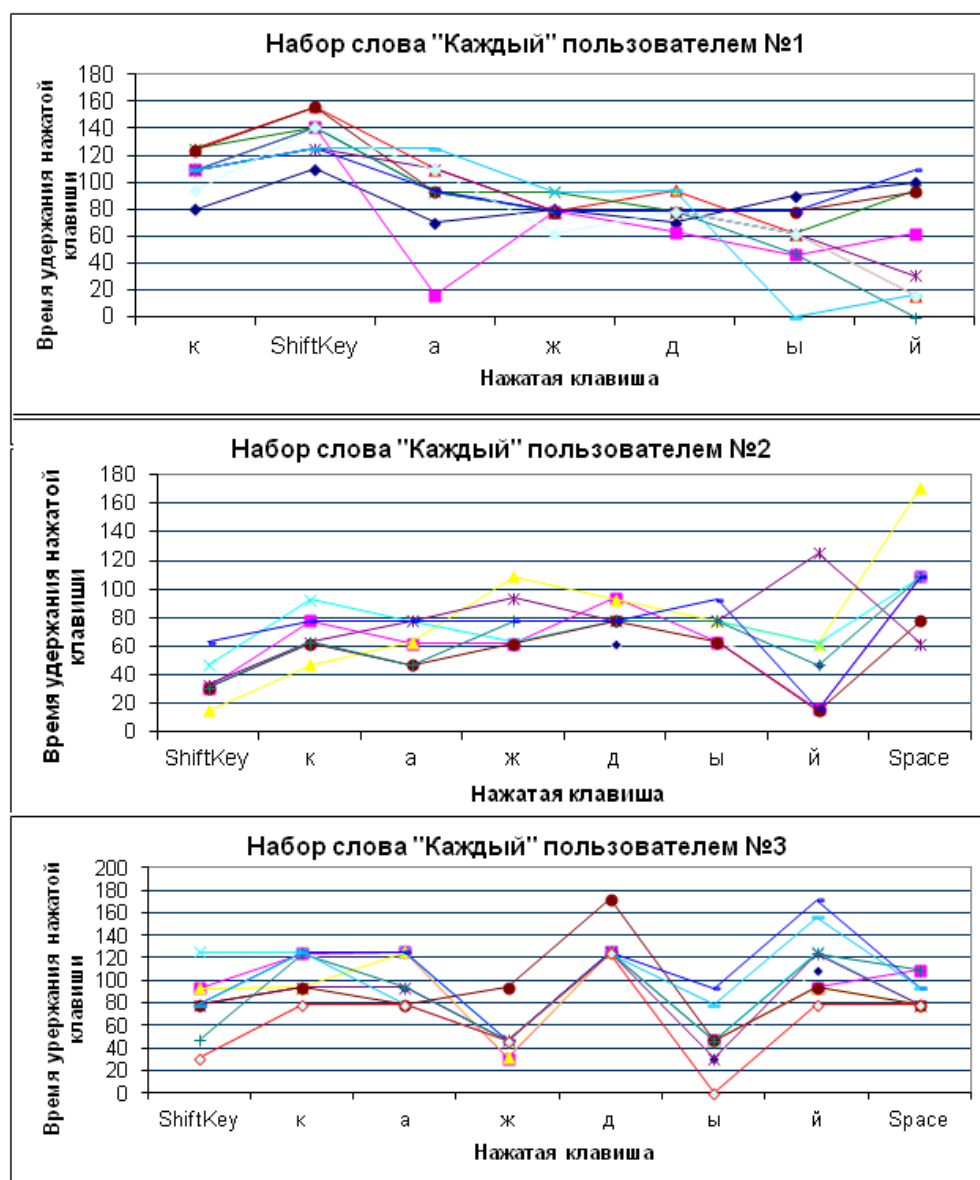


Рис. 1. Образцы клавиатурного почерка пользователей № 1–3

Fig. 1. Samples of Users' Keyboard Handwriting № 1–3

Поставленная задача по своей сути является задачей классификации, так как образцы клавиатурного почерка представляют собой конечное множество объектов, которое нужно соотнести с некоторым множеством классов. Для ее решения возможно использование таких методов, как регрессионный анализ, метод k -средних, метод главных компонент и т. д., которые позволят отсеять наименее информативные признаки клавиатурного почерка и, как следствие, предварительно уменьшить размерность обучающей выборки. При использовании регрессионного анализа для решения поставленной задачи авторами был выявлен ряд ограничений, влияющих на достоверность идентификации, что подробно рассмотрено в [8].

В данной работе предложено использование метода k -средних, поскольку он эффективен в случае, когда число кластеров заранее известно, достаточно прост в реализации, поэтому его можно использовать для анализа больших объемов данных. Кроме того, данный метод может учитывать весовые коэффициенты, что является несомненным преимуществом, поскольку одной из задач исследовательской работы авторов является оценка значимости параметров клавиатурного почерка.

Метод k -средних основан на вычислении расстояния от объекта до k ближайших соседей, причем число k задается индивидуально для каждого случая. Под «соседями» в данном случае будем понимать объекты, близкие к исследуемому. Определение принадлежности объекта к тому или иному классу (кластеру) производится на основе вывода о том, к какому классу относятся его ближайшие соседи [9]. На практике вопрос выбора оптимального значения параметра k является достаточно сложным, поскольку от правильности его выбора зависит правильность решения задачи о принадлежности объекта к некоторому классу. Так, если взять k слишком малым, возможна ситуация, при которой единственным ближайшим соседом окажется объект с неверно определенным классом, что приведет к принятию неверного решения о принадлежности данного объекта. Если же взять k , напротив, слишком большим, например равным количеству объектов, то расстояние до центра не будет иметь вообще никакого значения.

В рамках поставленной задачи классификация данных производилась путем разбиения множества объектов $X = 100$ на $k = 10$ кластеров, поскольку требуется идентификация 10 человек, предоставивших образцы почерка. В качестве меры близости объекта к центру кластера можно использовать расстояние Евклида, манхеттенское расстояние, формулу Минковского и др. В данном случае используется Евклидово расстояние, поскольку реализация метода k -средних осуществлялась с использованием программной среды Statistica, в вычислительный алгоритм которой заложена именно эта мера:

$$\rho(x, y) = |x - y| = \sqrt{\sum_{p=1}^n (x_p - y_p)^2},$$

где R^n – пространство множества объектов; x, y – координаты объекта в пространстве.

Применение данного метода было реализовано в программной среде Statistica. Для получения статистических данных в эксперименте приняли участие 10 пользователей, каждый из которых предоставил по 10 образцов клавиатурного почерка. Таким образом, в целом была получена выборка данных объемом 100 образцов клавиатурного почерка. В качестве текста для набора были использованы фразы-панграммы, примеры которых были приведены ранее.

Выбор такого варианта текста для набора основан на том, что данные фразы позволяют максимально эффективно оценить все параметры клавиатурного почерка. В работах [2; 3] отмечалось, что каждый образец почерка характеризуется 17 параметрами. Таким образом, вся статистическая выборка представляет собой матрицу X размером 100×17 .

Классификация данных производилась путем разбиения множества объектов $X = 100$ на $k = 10$ кластеров, поскольку поставленная задача требует идентификации 10 человек, предоставивших образцы почерка.

Достигнутые результаты

На рис. 2 представлено окно описания параметров в программе Statistica.

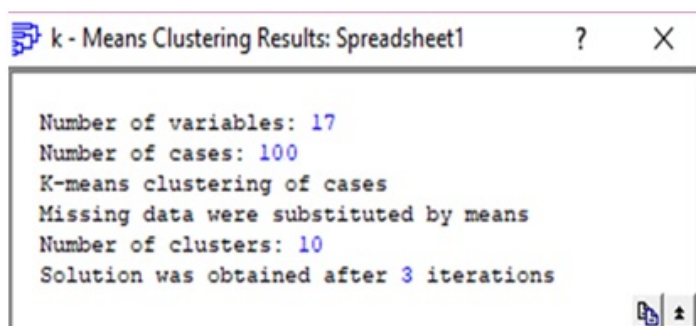


Рис. 2. Окно описания параметров анализа данных в программе Statistica

Fig. 2. A Window Describing the Parameters of Data Analysis in the Program Statistica

Одним из важнейших показателей эффективности применения метода идентификации является достоверность аутентификации пользователей [4]. При принятии решения о личности владельца почерка система идентификации может выдать как неправильное решение, приняв нелегального пользователя за легального, так и противоположное решение, отказав авторизованному пользователю в доступе. На основании этих факторов оценка эффективности применения данного метода производилась путем вычисления значений коэффициента ложного доступа FAR и коэффициента ложного отказа в доступе FRR. Коэффициент FAR (False Acceptance Rate) возникает в случае, когда зарегистрированный пользователь аутентифицируется как не имеющий доступа к системе, а коэффициент FRR (False Rejection Rate) характеризует ситуацию, при которой пользователь, не имеющий прав доступа, аутентифицируется как зарегистрированный [4].

При определении принадлежности образца почерка (Case_No) определенному кластеру (Cluster) в программе Statistica учитывалось его расстояние от центра данного кластера, вычисленного на основании данных о координатах элементов, ближайших к рассматриваемому. Окно результатов оценивания представлено в формате таблицы Word.

Результаты классификации

Classification Result

№ образца почерка (Case_No)	№ кластера (Cluster)	Расстояние до центра кластера (Distance)
4	4	0,06
5	4	0,06
21	4	0,06
22	4	0,06
23	4	0,06
24	4	0,05
25	4	0,06
26	4	0,06
27	4	0,07
28	4	0,07
30	4	0,06

Таблицы, аналогичные представленной, получены для каждого из 10 выделенных кластеров. На основании полученных результатов вычислялось среднее значение коэффициентов FAR и FRR, которые являются основными характеристиками биометрических систем аутентификации. Их значения составили 28 и 30 % соответственно. Полученные результаты являются неудовлетворительными, поскольку стандартным порогом доступа биометрических систем является значение 90 % [1].

Оценивая результаты, следует отметить, что использование данного метода для решения поставленной задачи имеет ряд ограничений, к которым можно отнести:

- необходимость хранения всей выборки данных для проведения классификации, что требует выделения большого объема памяти;
- использование данных без возможности дообучения, что критично для решения поставленной задачи, поскольку клавиатурный почерк представляет собой динамическую биометрическую характеристику, т. е. может изменяться с течением времени;
- близкую расположенность центров кластеров друг к другу в силу небольшого размаха значений выборки (в некоторых случаях разница в тысячные доли), что приводит к увеличению значений FRR и FAR. На рис. 3 видно, что центры кластеров, характеризующих пользователей 2, 5, 6, 9 практически совпадают, и результаты классификации данных, соответствующих этим пользователям, показали самые высокие значения ошибки.

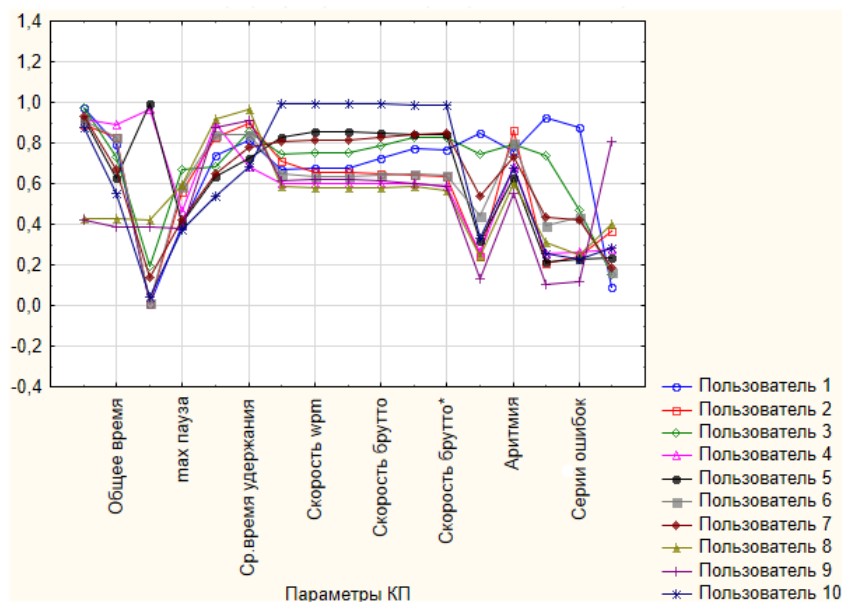


Рис. 3. Распределение центров кластеров

Fig. 3. Distribution of Cluster Centers

На основании описанных выше недостатков данного метода для повышения эффективности идентификации пользователя по клавиатурному почерку предложено использовать интеллектуальные методы, в частности искусственные нейронные сети. Применение метода k -средних показало, что центры кластеров находятся слишком близко друг к другу, поэтому плохо различимы. Такая ситуация может привести к тому, что значения, соответствующие разным пользователям, могут быть идентифицированы как принадлежащие одному человеку, и наоборот, что может привести к увеличению значений FRR и FAR. Точности метода не хватает для учета таких ограничений, в то же время нейронные сети обладают более высокой способностью к разграничению данных, что позволяет выдвинуть предположение о том, что их применение позволит получить более высокий результат. Кроме того, использование дан-

ного метода подразумевает использование статистики без возможности дообучения, что не позволяет учитывать факт изменения клавиатурного почерка пользователя со временем или под влиянием каких-либо внешних факторов. В то же время одним из важнейших преимуществ нейронных сетей является способность обучаться на основе данных окружающей среды и в результате обучения повышать свою производительность посредством интерактивного процесса корректировки синаптических весов и порогов. Кроме того, ИНС обладают такими преимуществами, как возможность выявлять скрытые закономерности и зависимости в потоке данных, а также обеспечивать более высокое быстродействие по сравнению с другими методами за счет распараллеливания процесса обработки данных.

Заключение

В работе рассматривается задача идентификации пользователя по клавиатурному почерку. После анализа существующих методов решения данной задачи авторами предложено использование метода k -средних, являющегося одним из наиболее простых в реализации методов машинного обучения для решения задач классификации. В ходе проведенных экспериментов получены результаты, представленные в виде значения ошибок первого и второго рода, которые являются стандартными критериями оценок надежности всех биометрических систем. Анализ полученных результатов показал, что пороговые значения доступа ниже, чем стандартные пороговые значения для биометрических систем. В связи с этим авторами предложено использование альтернативных методов для обработки статистических данных.

Список литературы / References

1. **Савинов А. Н.** Методы, модели и алгоритмы распознавания клавиатурного почерка в ключевых системах: Дис. ... канд. техн. наук. СПб., 2013. 128 с.
Savinov A. N. Methods, models and algorithms for recognition of keyboard handwriting in key systems. Dis. ... Cand. Tech. Sci. St. Petersburg, 2013, 128 p. (in Russ.)
2. **Еременко Ю. И., Олюнина Ю. С.** Об идентификации клавиатурного почерка пользователей // Сборник материалов XXVIII Междунар. науч.-практ. конф. «Перспективы развития информационных технологий». 2016. С. 145–151.
Eremenko Yu. I., Olyunina Yu. S. About the identification of keyboard handwriting users. In: Collection of materials of the XXVIII International Scientific and Practical Conference “Prospects for the development of information technology”, 2016, p. 145–151. (in Russ.)
3. **Еременко Ю. И., Олюнина Ю. С.** Об обработке потока данных с целью выявления скрытых характеристик клавиатурного почерка // Современные сложные системы управления. HTCS'2017: Материалы XII Междунар. науч.-практ. конф.: В 2 ч. Липецк: Изд-во Липецк. гос. техн. ун-та, 2017. Ч. 2. С. 31–36.
Eremenko Yu. I., Olyunina Yu. S. About the processing of the data stream to identify latent characteristics of the keyboard handwriting. In: *Sovremennyye slozhnyye sistemy upravleniya*. Lipetsk, Lipetsk State Technical University Publ., 2017, p. 31–36. (in Russ.)
4. **Шарипов Р. Р.** Разработка полигауссова алгоритма аутентификации пользователей в телекоммуникационных системах и сетях по клавиатурному почерку: Дис. ... канд. техн. наук: 05.12.13. Казань, 2006. 135 с.
Sharipov R. R. Development of poly-Gaussian algorithm for user authentication in telecommunication systems and networks using keyboard handwriting. Dis. ... Cand. Tech. Sci. Kazan, 2006, 135 p. (in Russ.)
5. **Скуратов С. В.** Использование клавиатурного почерка для аутентификации в компьютерных информационных системах // Безопасность информационных технологий. 2010. № 2. С. 35–38.

- Skuratov S. V.** Using keyboard handwriting for authentication in computer information systems. *Besopasnost informacionnykh tekhnologij* 2010, no. 2, p. 35–38. (in Russ.)
6. **Епифанцев Б. Н., Ложников П. С., Сулавко А. Е.** Альтернативные сценарии авторизации при идентификации пользователей по динамике подсознательных движений // Вопросы защиты информации. 2013. № 2. С. 28–35.
- Epifantsev B. N., Lozhnikov P. S., Sulavko A. E.** Alternative authorization scenarios for identifying users by the dynamics of subconscious movements. *Voprosi sashiti informatsii*, 2013, no. 2, p. 28–35. (in Russ.)
7. **Агеев М. С., Добров Б. В.** Метод эффективного расчета матрицы ближайших соседей для полнотекстовых документов // Вестник Санкт-Петербургского университета. Серия 10: Прикладная математика, информатика, вопросы управления. 2011. Вып. 3. С. 72–84.
- Ageev M. S., Doprov B. V.** The method of effective calculation of the matrix of nearest neighbors for full-text documents. *Vestnik St. Petersburg University. Series 10*, 2011, no. 3, p. 72–84. (in Russ.)
8. **Еременко Ю. И., Олюнина Ю. С.** Об определении наиболее значимых параметров клавиатурного почерка с помощью регрессионного анализа // Системы управления и информационные технологии. 2018. № 2 (72). С. 28–31.
- Eremenko Yu. I., Olyunina Yu. S.** About determination of the most significant parameters of keyboard handwriting using regression analysis. *Sistemy upravleniya i informatsionnye tekhnologii*, 2018, no. 2 (72), p. 28–31. (in Russ.)
9. **Стрюков Р. К., Шашкин А. И.** О модификации метода ближайших соседей // Вестник ВГУ. Серия: Системный анализ и информационные технологии. 2015. № 1. С. 114–120.
- Strukov R. K., Shashkin A. I.** On the modification of the nearest neighbors method. *Vestnik VSU. Seriya: Sistemy analiz i informatsionnye tekhnologii*, 2015, no. 1, p. 114–120. (in Russ.)

Материал поступил в редколлегию

Received
20.03.2019

Сведения об авторах / Information about the Authors

Еременко Юрий Иванович, доктор технических наук, профессор, заведующий кафедрой АИСУ, Старооскольский технологический институт (филиал) НИТУ «МИСиС» (мкр. Макаренко, 42, Старый Оскол, 309516, Белгородская область, Россия)

Yuriy I. Eremenko, Doctor of Technical Sciences, Professor, Head of the Department of Stary Oskol Technological Institute branch of the National Research Technological Institute “Moscow Institute of Steel and Alloys” (42 Makarenko md., Stary Oskol, 309516, Belgorod Region, Russian Federation)
erem49@mail.ru

Олюнина Юлия Сергеевна, аспирант, ассистент кафедры АИСУ, Старооскольский технологический институт (филиал) НИТУ «МИСиС» (мкр. Макаренко, 42, Старый Оскол, Белгородская область, 309516, Россия)

Yuliya S. Olyunina, graduate student, assistant at the department of AISU, Stary Oskol Technological Institute branch of the National Research Technological Institute “Moscow Institute of Steel and Alloys” (42 Makarenko md., Stary Oskol, 309516, Belgorod Region, Russian Federation)
julijasergeevna@mail.ru