Для цитирования: Пащенко Д.В., Бальзанникова Е.А., Сергина И.Г. Метод идентификации пользователей по биометрическому образу клавиатурного почерка с использованием двусвязного представления // Вопросы радиоэлектроники. 2018. № 12. С. 83–89.
DOI 10.21778/2218-5453-2018-12-83-89
УДК 004.93'12

Д.В. Пащенко[1], Е.А. Бальзанникова[1], И.Г. Сергина[2]

[1] ФГБОУ ВО «Пензенский государственный университет», [2] АО «Научно-производственное предприятие «Рубин»

МЕТОД ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЕЙ ПО БИОМЕТРИЧЕСКОМУ ОБРАЗУ КЛАВИАТУРНОГО ПОЧЕРКА С ИСПОЛЬЗОВАНИЕМ ДВУСВЯЗНОГО ПРЕДСТАВЛЕНИЯ

Задача установления авторства электронного документа или набранного на клавиатуре текста становится все более актуальной. Среди множества вариантов ее решения выделен метод, основанный на базе биометрических данных, полученных в результате анализа клавиатурного почерка. В частности, подробно рассмотрены основные параметры, получаемые на основании временных отметок нажатий клавиш, а также способ их формализованного двусвязного представления. Описан и реализован способ идентификации биометрических образов на базе статистического алгоритма, формирующего определенного вида доверительную область, внутри которой будут находиться параметры идентифицируемого объекта. Описаны методы предварительной обработки данных (фильтрации выбросов, нормализации и кластеризации), показана их эффективность на основании оценки точности распознавания методом перекрестной проверки для различных значений порога доступа. Для предложенного вероятностного алгоритма для идентификации образцов клавиатурного почерка по парольной фразе проведена оценка точности идентификации, обоснована необходимость предварительной обработки входных данных для повышения качества идентификации. На основании полученных результатов в дальнейшем предполагается формирование полного биометрического образа клавиатурного почерка с дальнейшей идентификацией по произвольному тексту.

Ключевые слова: *биометрические данные, временные характеристики, способ идентификации биометрических образов*

Введение

Установление авторства электронного документа или текста в процессе профессиональной или учебной деятельности в настоящее время является актуальной и в ряде случаев нетривиальной задачей.

Аутентификационные данные традиционно базируются на одном или нескольких компонентах:

- материальный объект: пропуск, ключ, proxy-карта, документы;
- знания: пароли, коды или ответ на вопрос;
- биометрическая информация: совокупность уникальных физиологических параметров пользователя.

Однако данные на основе атрибутов или знаний не всегда достоверны, например, в случае подмены оператора, так как системы, обеспечивающие безопасность на данном уровне, могут пострадать от подделки или воровства. В большинстве случаев решение об авторстве выносится на основании электронно-цифровой подписи электронного документа или аутентификационных данных пользователя, во время сеанса которого формировался электронный текст. Однако далеко не все электронные документы сопровождаются или могут быть сопровождены цифровой подписью.

Методы биометрической идентификации позволяют определить личность человека по его физиологическим характеристикам путем распознавания по заранее сохраненным образам. По сравнению с традиционными методами подобные способы аутентификации удобны тем, что они не могут быть случайно утеряны или забыты. Биометрический контроль доступа считается более надежным, так

как идентификаторы не могут быть переданы третьим лицам или скопированы.

Основным методом, использующим статические биометрические характеристики человека, является [1] идентификация по папиллярному рисунку на пальцах, радужной оболочке, геометрии лица, сетчатке глаза, рисунку вен руки, геометрии рук. Также существует семейство методов, использующих динамические характеристики [2]: идентификация по голосу, динамике рукописного почерка, сердечному ритму, походке. Однако для решения задачи идентификации пользователя и установления авторства документа необходим источник данных, который будет предоставлять информацию в течение всего сеанса работы пользователя.

На сегодняшний день существуют [3] методы и средства идентификации пользователей с помощью стандартных устройств ввода компьютера: клавиатуры и мыши. Подобно рукописному почерку клавиатурный почерк является уникальным для каждого человека, что позволяет с высокой долей вероятности идентифицировать пользователя, известного системе, или классифицировать пользователя как «чужого», если он не соответствует ни одному образу, имеющемуся в базе. Зачастую ввиду особенностей такой биометрической информации системы анализа клавиатурного почерка обладают более высокой точностью по сравнению с системами, основанными на физиологических параметрах [4].

С целью идентификации пользователей по клавиатурному почерку применяется множество методов [5–8]: вероятностный, основанный на предположении, что признаки не противоречат нормальному закону распределения, методы нечетких множеств, различные алгоритмы машинного обучения и нейронные сети. В данной статье подробно рассмотрен простой вероятностный алгоритм и показано, как с помощью методов предварительной обработки входных данных добиться высокой точности распознавания.

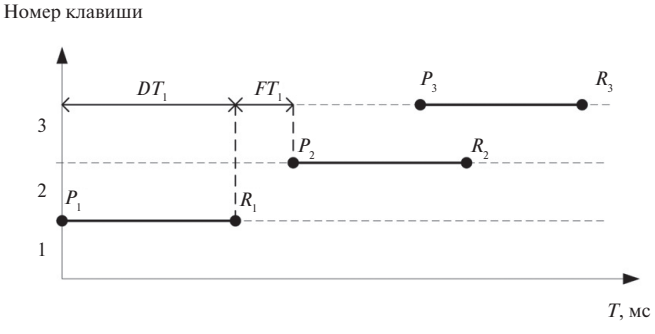

Рисунок 1. Графическое представление временных характеристик нажатия клавиш

Определение клавиатурного почерка

Для наглядности предложенных алгоритмов в данной статье будет рассмотрен метод идентификации пользователя по парольной фразе. Аналогичный алгоритм идентификации с лучшими показателями качества распознавания может быть использован для непрерывного анализа действий оператора в течение всего сеанса работы. После получения аутентификационного решения на основании данных клавиатурного почерка в дальнейшем можно сделать предположение об авторстве электронного документа или текста.

Процесс набора текста на клавиатуре можно представить в виде последовательности событий двух типов, которым сопоставляется отметка времени: P_i – нажатие клавиши и R_i – отпускание клавиши. Графически данные события представлены на рис. 1.

В связи с этим можно выделить набор признаков, отражающих динамику процесса печатания:

- время удержания клавиши DT (dwell time), где $DT_i = R_i - P_i$;
- интервал между нажатиями FT (flight time), где $FT_i = P_{i+1} - R_i$.

В связи с этим входной набор данных V, используемый для дальнейшего анализа, будет представлять собой последовательность характеристик смежных нажатий клавиш, его можно записать в виде

$$V = \{\{DT_1, FT_1\},...,\{DT_N, FT_N\}\}.$$

Совокупность представленных параметров, отражающих последовательные события, полностью описывает процесс набора текста и является достаточной для реализации идентификации пользователей по клавиатурному почерку. Несомненно, что на процесс печатания влияют как конструктивные особенности клавиатуры, так и эмоциональное состояние оператора, однако учесть подобные параметры на этапе формирования эталонного образа невозможно.

Вероятностный алгоритм

Рассмотрим основные принципы вероятностного алгоритма. Пусть имеется M образцов входных данных одного пользователя, которые соответствуют каждому из вводов парольной фразы: $\{V_1, ..., V_j, ..., V_M\}$. Если представить каждый элемент входного вектора как точку в двухмерном пространстве, то множество соответствующих элементов вектора $\{\{DT_1^j, FT_1^j\},...,\{DT_N^j, FT_N^j\}\}$ будет концентрироваться в одной области, образуя кластер. Для наглядности на рис. 2 приведен пример для первых элементов векторов признаков

из нескольких образцов ввода парольной фразы двумя пользователями.

Для формирования идентифицирующих признаков необходимо ограничить область в пространстве, внутри которой будут расположены точки, соответствующие характеристикам нажатий клавиш в рамках клавиатурного почерка одного пользователя.

Предположим, что оба признака DT и FT подчиняются нормальному закону распределения. Тогда для случая с двумя признаками функция плотности распределения будет определять эллипс рассеивания (или гиперэллипсоид для большего количества параметров). Кроме того, если случайная величина имеет нормальное распределение, то 99,7% значений расположены в интервале $(\mu - 3\sigma, \mu + 3\sigma)$. Определим решающую границу в виде эллипса, оси которого соответствуют доверительному интервалу каждого из признаков DT и FT. Решающее правило о принадлежности конкретного признака обозначенной области, исходя из определения эллипса, будет следующим:

$$\frac{(x-\mu_{DT})^2}{(\mu_{DT}+3\sigma_{DT})^2}+\frac{(y-\mu_{FT})^2}{(\mu_{FT}+3\sigma_{FT})^2}<1,$$

где μ_{DT}, μ_{FT} – математические ожидания времени удержания и времени между нажатиями i-го события в обучающей выборке, σ_{DT}, σ_{FT} – среднеквадратичные отклонения данных параметров.

Для идентификации пользователя по парольной фразе будем рассматривать попадание в доверительную область каждого элемента входного вектора признаков. На основании количества попаданий каждого из элементов, в зависимости от порогового значения будет приниматься решение о принадлежности входного образца клавиатурного почерка конкретному пользователю.

Оценка качества распознавания

Для оценки точности идентификации используем два критерия: вероятность ложного срабатывания (FAR) и вероятность ложного отказа (FRR). Оценка будет производиться для различных значений порога доступа процентного соотношения попаданий (от 10 до 100%) с целью подбора оптимальной величины.

В ходе экспериментальной проверки были получены по 7–10 образцов ввода парольной фразы пяти разных пользователей. Каждый образец содержал временные характеристики последовательных нажатий соседних клавиш: DT и FT.

В процессе обучения и оценки точности результатов применялся метод перекрестной проверки [9], согласно которому набор образцов клавиатурного почерка пользователя делится на две равные части: обучающую и тестовую, после чего полученные части выборки меняются местами. Результатом является среднее значение оценки точности идентификации для двух экспериментов. Метод перекрестной проверки может применяться и для большего количества групп в обучающей выборке, но поскольку количество образцов ввода невелико, в данном случае разделение исходного набора на две части является достаточным как для объективной оценки точности, так и для формирования эталонного образца клавиатурного почерка.

Результат оценки обработки данных по указанному алгоритму показан на рис. 3. По оси абсцисс расположены значения порога доступа в виде процентного соотношения количества попаданий, а по оси ординат – доля ошибок FAR и FRR.

Из рисунка видно, что наименьшая вероятность ошибок FAR и FRR наблюдается, когда порог доступа составляет около 75%. При этом на тестовых данных имеются ошибки идентификации.

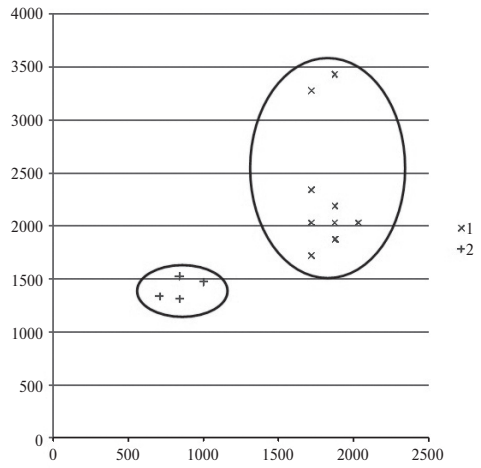

Рисунок 2. Графическое представление элементов векторов признаков для двух пользователей

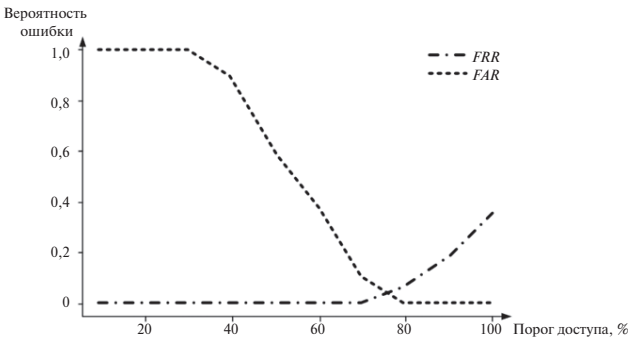

Рисунок 3. Оценка точности распознавания с применением простого вероятностного алгоритма

Фильтрация выбросов

Одной из проблем анализа клавиатурного почерка является проблема нарушения стабильности почерка, которая часто связана с ошибочным вводом. Такого рода нарушения проявляются как положительные выбросы временных интервалов событий.

Существуют два варианта решения данной проблемы: «сглаживание» выброса или исключение образца из обработки. Если предположить, что вектор, содержащий отклонение, тем не менее отражает механику набора текста пользователя, то можно скорректировать величину признака, например, уменьшив его так, чтобы сохранялась общая динамика, или привести его к среднему значению параметра. Однако такие манипуляции могут исказить реальный характер клавиатурного почерка, тем самым ухудшив качество распознавания. Авторы предлагают исключать подобные образцы из процесса обучения и получения эталонного образа, поскольку ввод, содержащий выброс, является ошибочным.

В ходе практических исследований наилучшие результаты предварительной обработки показал метод, основанный на интерквартильном размахе [10]. Данный метод использует «робастные» характеристики выборки: медиану ($Q2$), первый ($Q1$) и третий ($Q3$) квартили. Согласно ему [11], если значение превосходит третий квартиль более чем на $1{,}5$ межквартильных интервала (IQR) или меньше первого квартиля на $1{,}5\times IQR$, то оно называется «умеренным выбросом». Если значение больше третьего квартиля или меньше первого квартиля более чем на $3\times IQR$, то оно называется «экстремальным выбросом».

Исходя из этого, данный метод можно определить следующим образом:

$$v_i \begin{cases} Q1-1{,}5IQR \le v_i \le Q3+1{,}5IQR - \text{выброса нет;} \\ v_i < Q1-3IQR - \text{отрицательный экстремальный выброс;} \\ Q1-3IQR \le v_i < Q1-1{,}5IQR - \text{отрицательный умеренный выброс;} \\ Q3+1{,}5IQR < v_i \le Q3+3IQR - \text{положительный умеренный выброс;} \\ v_i > Q3+3IQR - \text{положительный экстремальный выброс.} \end{cases}$$

В результате применения данного метода на практике были отфильтрованы выбросы, которые явно соответствуют ошибкам, и оставлены образцы, содержащие небольшое отклонение характеристик от средней величины, но вводимые без ошибок.

Нормализация данных

Основной задачей нормализации является представление параметров клавиатурного почерка независимо от способа отсчета времени, что позволяет использовать любой источник в качестве таймера. В программной реализации в качестве основы формирования временных интервалов можно применять системные часы или любой доступный аппаратный счетчик. Для этого необходимо провести преобразование вектора признаков к стандартному нормальному распределению с нулевым математическим ожиданием ($\mu = 0$) и единичной дисперсией ($\sigma = 1$). Исходя из этого, нормализация параметров будет вычисляться по формуле

$$v_i' = \frac{v_i - \mu}{\sigma}.$$

В рамках решения задачи идентификации пользователей по клавиатурному почерку нормализация позволяет минимизировать зависимость вектора признаков клавиатурного почерка от некоторых внешних и внутренних факторов. В частности, нормализация величины времени удержания клавиш снижает влияние внешних конструктивных особенностей клавиатуры, таких как высота клавиш или технология их изготовления.

Кластеризация признаков

Для подавляющего большинства алгоритмов анализа клавиатурного почерка основной проблемой является нестабильность основных характеристик, обусловленная вариациями способов набора одной и той же парольной фразы в рамках клавиатурного почерка одного пользователя.

Помимо рассмотренных способов предварительной обработки данных, предназначенных для повышения качества распознавания биометрического образа, авторами предложен способ повышения качества самого алгоритма распознавания в условиях переменной динамики набора текста.

В большинстве случаев набор значений одного признака рассматривается как единая выборка, подчиняющаяся некоторому закону распределения. Исходя из предположения о возможной нестабильности клавиатурного почерка, рассматриваются значения признаков в виде отдельных групп, соответствующих альтернативным вариантам способа набора парольной фразы или текста. Предполагается, что каждая из групп также подчиняется нормальному закону распределения. Таким образом, данный способ обработки осуществляет предварительную кластеризацию каждого набора признаков с последующим применением к каждому кластеру одного из рассмотренных ранее алгоритмов идентификации.

Рисунок 4. Процесс обработки данных клавиатурного почерка

Для достижения поставленной задачи могут быть использованы различные алгоритмы кластеризации [12]. В рамках проведенного исследования применялся метод агломеративной кластеризации, так как он хорошо подходит для задач, где количество кластеров заранее неизвестно, при возможности экспериментального подбора порогового расстояния между кластерами, на основании которого осуществляется разделение данных. В результате для каждого кластера вычисляются математическое ожидание и дисперсия по каждому признаку и определяется решающая граница в виде эллипса. Если соответствующий признак попадает хотя бы в одну доверительную область, засчитывается совпадение.

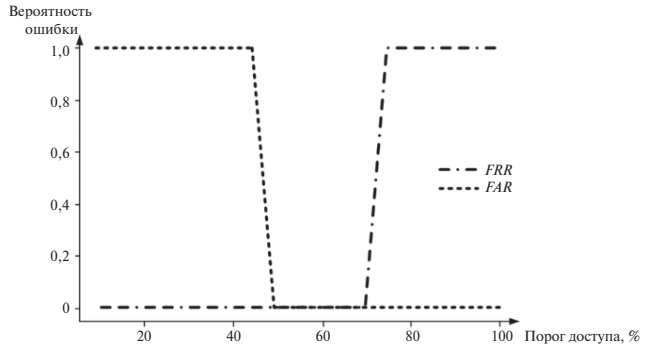

Рисунок 5. Результаты оценки точности идентификации с применением алгоритмов предварительной обработки

Таким образом, процесс обработки данных клавиатурного почерка можно представить в виде алгоритма (рис. 4). Результаты оценки точности идентификации на основании данного алгоритма показаны на рис. 5.

Как показывает график, присутствует широкая область порога доступа, в которой на тестовых данных наблюдается нулевая вероятность ошибок *FAR* и *FRR*. Хотя результаты получены на основании небольшого набора экспериментальных данных, очевидно преимущество применения процедуры кластеризации исходных данных, позволяющей значительно повысить качество идентификации с применением самого простого вероятностного алгоритма [13].

Выводы

Вероятностный алгоритм распознавания биометрического образа клавиатурного почерка с использованием предварительной обработки данных – фильтрации выбросов и нормализации – позволил повысить точность идентификации пользователей по клавиатурному почерку. Экспериментально показана эффективность применения алгоритмов кластеризации исходных данных для повышения точности идентификации биометрического образа.

СПИСОК ЛИТЕРАТУРЫ

1. Руководство по биометрии / Р.М. Болл, Д.Х. Коннел, Ш. Панканти, Н.К. Ратха, Э.У. Сеньор. М.: Техносфера, 2007. 368 с.
2. Кухарев Г.А. Биометрические системы. Методы и средства идентификации личности человека. М.: Политехника, 2001. 240 с.
3. Борисов Р.В., Зверев Д.Н., Сулавко А.Е., Писаренко В.Ю. Оценка идентификационных возможностей особенностей работы пользователя с компьютерной мышью // Вестник СибАДИ. 2015. № 5. С. 45–50.
4. Иванов А.И. Биометрическая идентификация личности по динамике подсознательных движений. Пенза: Изд-во ПГУ, 2000. 188 с.
5. Teh Sh. P., Teoh B.J. A., Yue S. A survey of keystroke dynamics biometrics // The Scientific World Journal. 2013. P. 1–24.
6. Брюхомицкий Ю.А. Статистические методы распознавания клавиатурного почерка // Известия Южного федерального университета. Технические науки. 2009. № 11 (100). С. 139–147.
7. Брюхомицкий Ю.А. Гистограммный метод распознавания клавиатурного почерка // Известия Южного федерального университета. Технические науки. 2010. № 11 (112). С. 8–12.

8. Ходашинский И. А., Савчук М. В., Горбунов И. В., Мещеряков Р. В. Технология усиленной аутентификации пользователей информационных процессов // Доклады Томского государственного университета систем управления и радиоэлектроники. 2011. № 2 (24). Ч. 3. С. 236–248.
9. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015. 400 с.
10. Крамер Г. Математические методы статистики. М.: МИР, 1975. 721 с.
11. Зайдель А. Н. Элементарные оценки ошибок измерений. М.: Наука, 1968. 98 с.
12. Жамбю М. Иерархический кластер-анализ и соответствия. М.: Финансы и статистика, 1988. 342 с.
13. Пащенко Д. В., Бальзанникова Е. А. Повышение точности идентификации пользователя по биометрическим данным клавиатурного почерка // Новые информационные технологии и системы (НИТИС-2017): тр. XIV Междунар. науч.-техн. конф. Пенза: Изд-во ПГУ, 2017. С. 166–169.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Пащенко Дмитрий Владимирович, д.т.н., профессор, заведующий кафедрой «Вычислительная техника», ФГБОУ ВО «Пензенский государственный университет», Российская Федерация, 440026, Пенза, ул. Красная, д. 40, тел.: 8 (927) 287-33-32, e-mail: dmitry.pashchenko@gmail.com.
Бальзанникова Елена Алексеевна, аспирант, ФГБОУ ВО «Пензенский государственный университет», Российская Федерация, 440026, Пенза, ул. Красная, д. 40, тел.: 8 (927) 378-22-21, e-mail: elenabalzannikova@gmail.com.
Сергина Ирина Геннадьевна, инженер-программист, АО «Научно-производственное предприятие «Рубин», Российская Федерация, 440000, Пенза, ул. Байдукова, д. 2, тел.: 8 (8412) 20-89-88, e-mail: tig9477@ya.ru.

For citation: Pashchenko D. V., Balzannikova E. A., Sergina I. G. User identification method by means of biometric image of keystroke dynamics with double-chained representation. Voprosy radioelektroniki, 2018, no. 12, pp. 83–89.
DOI 10.21778/2218-5453-2018-12-83-89

D. V. Pashchenko, E. A. Balzannikova, I. G. Sergina

USER IDENTIFICATION METHOD BY MEANS OF BIOMETRIC IMAGE OF KEYSTROKE DYNAMICS WITH DOUBLE-CHAINED REPRESENTATION

To date, more and more relevant is the task of establishing the authorship of an electronic document or typed on the keyboard text. Among the many options for solving this problem, a method based on biometric data obtained as a result of the analysis of the keystroke dynamics is highlighted. In particular, the paper considers the main parameters obtained on the basis of time stamps of keystrokes, as well as the method from the formalized doubly-chained representation. A method for identifying biometric images based on a statistical algorithm that forms a certain type of trust area within which the parameters of the identified object will be located has been described and implemented. In addition, the paper describes methods for preliminary processing of data (emission filtering, normalization and clustering), as well as their effectiveness based on the evaluation of the accuracy of cross-validation for different values of the access threshold. Thus, in the course of practical research, a simple probabilistic algorithm was implemented to identify samples of the keyboard handwriting with a pass phrase, an estimation of the identification accuracy was made, the necessity of preliminary processing of input data for improving the quality of identification was justified. Based on the results obtained, it is further assumed that a complete biometric image of the keyboard handwriting will be formed, with further identification by arbitrary text.

Keywords: biometric data, time characteristics, method of identification of biometric images

REFERENCES

1. Bolle R.., Connell J., Pankanti S., Ratha N. K., Senior A. W. *Guide to biometrics.* Springer, 2004, 334 p.
2. Kuharev G. A. *Biometricheskie sistemy. Metody i sredstva identifikacii lichnosti cheloveka* [Biometric systems. Methods and means of identification of a person]. Moscow, Politehnika Publ., 2001, 240 p. (In Russian).
3. Borisov R. V., Zverev D. N., Sulavko A. E., Pisarenko V. Ju. Evaluation of the identification capabilities of the features of the user with a computer mouse. *Vestnik SibADI,* 2015, no. 5, pp. 45–50. (In Russian).
4. Ivanov A. I. *Biometricheskaja identifikacija lichnosti po dinamike podsoznatelnyh dvizhenij* [Biometric identification based on the dynamics of subconscious movements]. Penza, PGU Publ., 2000, 188 p. (In Russian).
5. Teh Sh. P., Teoh B. J. A., Yue S. A survey of keystroke dynamics biometrics. *The Scientific World Journal,* 2013, pp. 1–24.
6. Bryukhomitsky Yu. A. Statistical methods of keystroke dynamics recognition. *Izvestiya SFedU. Engineering sciences,* 2009, no. 11 (100), pp. 139–147. (In Russian).
7. Bryukhomitsky Yu. A. Histogram recognition keyboard writing style. *Izvestiya SFedU. Engineering sciences,* 2010, no. 11 (112), pp. 8–12. (In Russian).
8. Hodashinsky I. A., Savchuk M. V., Gorbunov I. V., Meshcheryakov R. V. Strong authentication technology of the users of information processes. *Proceedings of TUSUR,* 2011, no. 2–3 (24), pp. 236–248. (In Russian).
9. Flach P. *Machine learning: the art and science of algorithms that make sense of data.* Cambridge University Press, 2012, 409 p.
10. Cramér H. *Mathematical methods of statistics.* Princeton University Press, 1947, 575 p.
11. Zajdel A. N. *Jelementarnye ocenki oshibok izmerenij* [Elementary estimates of measurement errors]. Moscow, Nauka Publ., 1968, 98 p. (In Russian).

12. Jambu M. *Classification automatique pour l'analyse des données* [Hierarchical cluster analysis and matching]. Dunod, 1978, 342 p. (In French).
13. Pashchenko D. V., Balzannikova E. A. Improving the accuracy of user identification based on the biometric data of keyboard handwriting. (Conference proceedigs) Novye informacionnye tehnologii i sistemy (NITIS-2017). Penza, PGU Publ., 2017, pp. 166–169. (In Russian).

AUTHORS

Pashchenko Dmitry, D. Sc., professor, head of computer science department, Penza State University, 40, Krasnaya St, Penza, 440026, Russian Federation, tel.: +7 (927) 287-33-32, e-mail: dmitry.pashchenko@gmail.com.
Balzannikova Elena, post-graduate student, Penza State University, 40, Krasnaya St, Penza, 440026, Russian Federation, tel.: +7 (927) 378-22-21, e-mail: elenabalzannikova@gmail.com.
Sergina Irina, software engineer, JSC Research and Production Enterprise «Rubin», 2, Baydukova St, Penza, 440000, Russian Federation, tel.: +7 (8412) 20-89-88, e-mail: tig9477@ya.ru.