УДК 004.065

Ю.А. Брюхомицкий

СТАТИСТИЧЕСКИЕ МЕТОДЫ РАСПОЗНАВАНИЯ КЛАВИАТУРНОГО ПОЧЕРКА*

Обсуждается один из возможных подходов к повышению точности клавиатурных средств аутентификации, который отличается от известных наличием двухэтапной процедуры обучения, включающей получение вначале оценок функций распределения клавиатурных параметров и затем на их основе операторных оценок соответствующих плотностей распределения. Преимущество предлагаемого подхода состоит в том, что он дает существенно более высокую асимптотическую точность оценивания, а в конечном итоге точность клавиатурной аутентификации.

Клавиатурный почерк; биометрические параметры; статистическое распознавание; эмпирические функции распределения; операторные оценки плотностей вероятностей.

Yu. A. Bryukhomitsky

STATISTICAL METHODS OF KEYSTROKE DYNAMICS ECOGNITION

We are discussing one of the possible approaches to the increase of keyboard-based authentication means, which differs from existing ones by a two-stage training procedure. It consists of estimation of keystroke feature distribution followed by operator estimation of probability densities. The advantage of this approach is based on the fact that it provides much higher asymptotical estimation precision and hence better keystroke authentication.

Keystroke dynamics; biometric features; statistical recognition; empirical distribution functions; operator estimation of probability densities.

Контроль доступа в компьютерные системы, реализуемый на основе анализа его клавиатурного почерка (КП), имеют ряд неоспоримых преимуществ. Главными из них являются: минимальная в классе биометрических средств стоимость, удобство использования, возможность эффективного сочетания с другими средствами контроля доступа, в частности, парольными. К недостаткам биометрических средств этого класса принято относить: недостаточную для самостоятельного использования точность, зависимость результатов от психофизического состояния личности, наличие определенного уровня навыков работы на клавиатуре.

В данной работе обсуждается один из возможных подходов к повышению точности клавиатурных средств аутентификации, основанный на использовании особенностей статистических методов распознавания.

Принцип аутентификации пользователя, претендующего на доступ в компьютерную систему по его КП, заключается в проведении анализа КП при вводе некоторой контрольной фразы и вынесении по результатам анализа соответствующего аутентификационного решения. Исходными данными для проведения анализа яв-

* Работа выполнена при поддержке гранта РФФИ № 08-07-00117-а.

ляются особенности динамики работы на клавиатуре данного пользователя, представленные в виде совокупности контролируемых клавиатурных параметров. Анализ состоит в формировании текущих клавиатурных параметров идентифицировавшего себя пользователя и сравнении их с эталонными параметрами пользователя с тем же именем, сформированными ранее, на этапе его регистрации.

Отправной точкой при создании методов и средств аутентификации личности по КП является принятый способ представления и использования индивидуальных клавиатурных параметров. Определим этот способ.

Пусть с клавиатуры пользователем за период времени T вводится некоторая контрольная фраза, содержащая q символов. При вводе этой фразы произойдет $r = q + p$ событий клавиатуры: q удержаний клавиш и $p = q - 1$ пауз между удержаниями. При большой скорости ввода возможны наложения времен удержания клавиш, когда нажатие очередной клавиши предшествует отпусканию предыдущей клавиши. Будем интерпретировать такой вид событий клавиатуры, как отрицательные значения длительности пауз между удержаниями.

Введем обозначения:

- τ_i – значение длительности удержания клавиши i, $\tau_i > 0$;
- τ_{ij} – алгебраическое значение длительности паузы между удержаниями клавиш i и j.

Процесс ввода некоторой контрольной фразы, в которой $r = 11$, $q = 6$, $p = 5$, иллюстрируется временной диаграммой (рис. 1).

Временная раскладка процесса клавиатурного ввода контрольной фразы в виде сочетаний длительностей удержания клавиш: $\tau_1, \tau_2, \tau_3, ..., \tau_n$ и длительностей пауз между удержаниями: $\tau_{12}, \tau_{23}, \tau_{34}, ..., \tau_{(n-1)n}$ индивидуальна для каждого пользователя и выступает в качестве эталона КП.

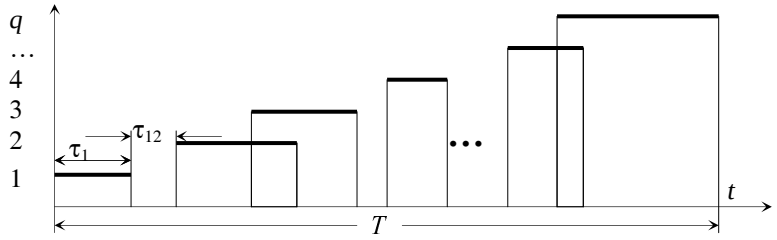

Рис. 1. Временная диаграмма процесса ввода контрольной фразы

Поставим в соответствие результату клавиатурного ввода контрольной фразы, временная диаграмма которой показана на рис.1, r-мерный вектор биометрических (клавиатурных) параметров

$$V = \{v_j\}, \quad j = \overline{1,r},$$

каждый компонент v_j которого соответствует длительности любого очередного события клавиатуры (будь то удержание клавиши или пауза между удержаниями), произошедшего за период T. События клавиатуры, состоящие в наложении времен удержания клавиш, будем интерпретировать отрицательными значениями соот-

ветствующих компонент вектора V. При таком представлении вектор биометрических параметров V можно рассматривать как образец КП данного пользователя.

Так, для временной диаграммы, показанной на рис. 1, вектор биометрических параметров будет иметь вид:

$$V = \{v_1, v_2, ...v_{11}\},$$

где

$v_1 = \tau_1$; $v_2 = \tau_{12}$; $v_3 = \tau_2$; $v_4 = -\tau_{23}$; $v_5 = \tau_3$; $v_6 = \tau_{34}$; $v_7 = \tau_4$; $v_8 = \tau_{45}$; $v_9 = \tau_5$; $v_{10} = -\tau_{56}$; $v_{11} = \tau_6$.

Для получения клавиатурного эталона пользователя необходимо иметь серию, состоящую из L образцов КП, которая составит обучающую выборку для некоторого s-класса

$$\Psi^{(s)} = \{V_i\}, \quad i = \overline{1, L}.$$

В общем случае в системе может быть зарегистрировано множество $K = \{k_1, k_2, \ldots, k_M\}$ пользователей, каждый из которых будет представлен своим эталоном и будет соотнесен с определенным классом из множества классов $s = \{s_1, s_2, \ldots, s_M\}$. Таким образом, образуется однозначное отображение совокупности пользователей $\{K\}$ на множество классов $\{s\}$.

Для формирования эталонов всех M легитимных пользователей потребуется соответственно M обучающих выборок

$$\Psi^{(s_1)}, \Psi^{(s_2)}, ..., \Psi^{(s_M)}.$$

В режиме аутентификации неизвестный x-пользователь предъявляет обученной клавиатурной системе контроля доступа (КСКД) образец своего КП в виде вектора биометрических параметров $V^{(s_x)} = \{v_j\}, \quad j = \overline{1, r}$. Система должна на основе вектора $V^{(s_x)}$ сформировать эталонное описание неизвестного x-класса, сравнить его с эталонами всех зарегистрированных в системе $\{k_1, k_2, \ldots, k_M\}$ пользователей и по результатам сравнения вынести соответствующее аутентификационное решение. В такой постановке фактически решается задача классификации вектора $V^{(s_x)}$ на M+1 взаимоисключающих классов: M классов из множества $s = \{s_1, s_2, \ldots, s_M\}$, соответствующих зарегистрированным в системе пользователям и (M+1)-й класс, отведенный всем остальным пользователям, объединяемым понятием «чужие». При наличии процедуры предварительной авторизации пользователей задача упрощается и сводится к классификации вектора $V^{(s_x)}$ на два класса: s_c – «свой», то есть принадлежащий к какому-либо классу из множества $\{s\}$, и $s_ч$ –«чужой», то есть не принадлежащий ни к одному классу из множества $\{s\}$.

Конечной целью обучения является формирование эталонных описаний классов. Форма этих описаний определяется способом их использования в решающих правилах. Природа данных при анализе КП носит случайный характер, поэтому вид решающего правила может быть заимствован из теории статистических решений, сведен к формированию отношения правдоподобия условных плотностей распределения и сравнению его с некоторым порогом $C_п$:

$$\frac{w_r(V \mid s_1)}{w_r(V \mid s_2)} \geq C_{\ddot{I}} \ , \qquad (1)$$

где $w_r(V \mid s_i)$ – условная совместная *r*-мерная плотность вероятности выборочных значений $\{v_j\}, \quad j = \overline{1,r}$ r при условии их принадлежности к классу s_i.

В том случае, если хотя бы с некоторым приближением вид закона распределения известен, а априорная неопределенность относится лишь к параметрам этого распределения, то применяются параметрические методы распознавания. Целью обучения при этом является получение оценок параметров известного распределения, по которым затем вычисляются плотности вероятностей. Например, в ряде приложений, связанных с использованием биометрических систем контроля доступа, вводится допущение о нормальном законе распределения биометрических параметров, что позволяет с помощью параметрических методов распознавания получить хорошие результаты по точности аутентификации [1].

Более общим и сложным является случай, когда нет априорных сведений не только о параметрах, но и о законе распределения. Тогда применяются непараметрические методы распознавания. Целью обучения в такой ситуации является получение оценок условных плотностей вероятностей.

В задаче классификации клавиатурных биометрических параметров в силу ряда специфических причин, связанных с нестабильностью КП, допущение о «нормальности» закона распределения может привести к ошибкам аутентификации. Поэтому в тех случаях, когда указанные причины невозможно игнорировать, приходится обращаться к непараметрическим методам [2].

Особенность реализации отношения правдоподобия (1) при непараметрической классификации параметров КП состоит в том, что плотности $w_r(V \mid s_i)$ априорно не известны и должны быть представлены своими оценками $\widehat{w}_r(V \mid s_i)$, полученными при обучении на основе образцов векторов $\{V_i\}, \quad i = \overline{1,L}$.

Анализ наиболее распространенных непараметрических методов восстановления плотности вероятности с помощью гистограммных, парценовских, *k* ближайших соседей, полигональных и других известных оценок показывает, что они не обеспечивают точности оценивания при реальных конечных объемах обучающих выборок. Кроме того, ряд методов (разложений по базисным функциям, полигональный) могут быть использованы только для оценивания одномерных плотностей вероятностей.

В такой ситуации целесообразно представлять исходные данные для принятия решений не оценками плотностей распределения $\widehat{w}_r(V \mid s_i)$, а оценками функций распределения $\widehat{F}_r(V \mid s_i)$ [3]. Основное преимущество такого подхода состоит в том, что появляется принципиальная возможность использования значений эмпирической функции распределения $\widehat{F}_r(V \mid s_i)$ во всех точках V области ее определения. Как известно, оценивание плотности распределения возможно только на основе конечного множества наблюдений, при этом недостаток важной информации восполняется всевозможными допущениями (введение весовых функ-

ций, функций потенциала и т.п.), которые собственно и породили множество различных непараметрических методов. Между тем, функции распределения $F_r(V \mid s_i)$ содержат всю доступную информацию о классах образов, а их оценки $\hat{F}_r(V \mid s_i)$ позволяют контролировать точность аппроксимации функций $F_r(V \mid s_i)$ при любых объемах L обучающей выборки. Однако непосредственно использовать оценки $\hat{F}_r(V \mid s_i)$ вместо оценок $\hat{w}_r(V \mid s_i)$ при построении решающего правила не представляется возможным.

Для приведения исходных данных, представленных оценками функций распределения $\hat{F}_r(V \mid s_i)$, к традиционной структуре решающего правила (1) можно перейти от оценок $\hat{F}_r(V \mid s_i)$ к оценкам $\hat{w}_r(V \mid s_i)$, исходя из определения плотности распределения $\hat{w}_r(V \mid s_i)$ как производной от функции $F_r(V \mid s_i)$ [3].

В результате получаем двухэтапную процедуру обучения. На первом этапе по обучающим выборкам $\Psi^{(s_1)}, \Psi^{(s_2)}, ..., \Psi^{(s_M)}$ строятся эмпирические функции распределения $\hat{F}_r(V \mid s_i)$ для всех классов образов $s = \{s_1, s_2, \ldots, s_M\}$. На втором этапе по эмпирическим функциям $\hat{F}_r(V \mid s_i)$ формируются оценки плотностей вероятностей $\hat{w}_r(V \mid s_i)$, которые и становятся эталонными описаниями классов.

Рассмотрим принципы реализации первого этапа обучения КСКД. Процесс формирования клавиатурного эталона образцов некоторого s_i-класса будем трактовать как многомерный случайный процесс $\xi_r^{(s)}(t)$ (мерности r), представляющий собой случайные изменения во времени признака V. При этом будем полагать (возможно, с некоторым приближением), что случайный процесс $\xi_r^{(s)}(t)$ удовлетворяет условию эргодичности.

Для одномерного случайного процесса $\xi_j^{(s)}(t)$ можно получить соответствующую ему оценку одномерной функции распределения клавиатурных параметров пользователя s_i-класса $\hat{F}_r(v_j \mid s_i)$ – как отношение суммарного времени пребывания реализации случайного процесса $\xi_j^{(s)}(t)$ под некоторым уровнем v к длительности реализации T [3]:

$$\hat{F}_r(v_j \mid s_i) = \frac{1}{T}\sum\nolimits_k t_k ,$$

где t_k – длительность k-го выброса процесса $\xi_j^{(s)}(t)$ под уровнем v.

Например, временная диаграмма одномерного случайного процесса $\xi_j^{(s)}(t)$, наблюдаемого по координате v_j вектора V**,** может быть такой, как показана на рис. 2.

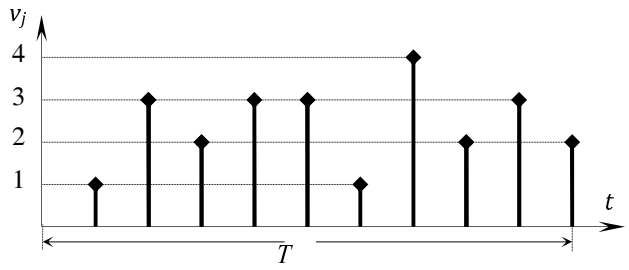

Рис. 2. Временная диаграмма одномерного случайного процесса

На рис. 3 приведен график оценки функции распределения $\widehat{F}_r(v_j \mid s_i)$ случайного процесса $\xi_j^{(s)}(t)$, показанного на рис. 2.

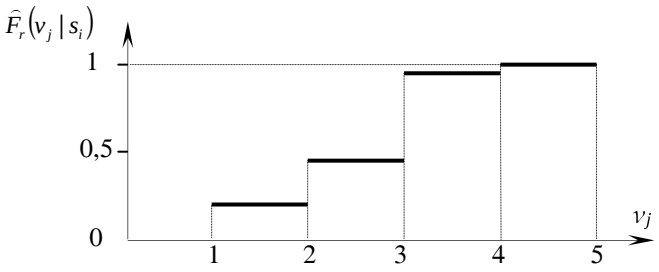

Рис. 3. График функции распределения $\widehat{F}_r(v_j \mid s_i)$

Для многомерного случайного процесса $\xi_r^{(s)}(t)$ соответствующую ему оценку многомерной функции распределения $\widehat{F}_r(V \mid s_i)$ биометрических параметров пользователя *s*-класса можно получить как отношение суммарного времени пребывания реализации случайного процесса $\xi_r^{(s)}(t)$ внутри области, ограниченной некоторой гиперплоскостью Q(V), к длительности реализации *T*:

$$\widehat{F}_r(V \mid s_i) = \frac{1}{T}\sum\nolimits_k t_k \,, \tag{2}$$

где t_k – длительность *k*-го выброса случайного процесса $\xi_r^{(s)}(t)$ внутри области, ограниченной гиперплоскостью Q(V).

Доказано [3], что оценка (2) является несмещенной и состоятельной.

На этом заканчивается первый этап обучения КСКД.

На втором этапе обучения по эмпирическим функциям $\widehat{F}_r(V \mid s_i)$ формируются оценки плотностей вероятностей $\widehat{w}_r(V \mid s_i)$, которые и становятся эталон-

ными описаниями классов. Непосредственный переход от $\hat{F}_r(V \mid s_i)$ к $\hat{w}_r(V \mid s_i)$ по правилам численного дифференцирования неприемлем. Известные методы численного дифференцирования не обеспечивают требуемой сходимости, поскольку исходно являются некорректными (малые вариации дифференцируемой функции приводят к значительным изменениям результата). Поэтому предлагается использование операторных оценок плотности вероятности $\hat{w}_r(V \mid s_i)$, основанных на аппроксимации оператора дифференцирования [3].

Суть метода, позволяющего получить операторные оценки плотности вероятности $\hat{w}_r(V \mid s_i)$, сводятся к следующему. В *r*-мерном пространстве V^r биометрических параметров вводятся в рассмотрение точки

$$V_m = V + m\mathbf{h},$$

где $V \in V^r$ – произвольная точка пространства V^r, в которой делается оценка плоности $\hat{w}(V)$;

$\mathbf{h} = h\mathbf{e}$, $\mathbf{e} = (1, 1, \ldots, 1)^T$ – единичный *r*-мерный вектор;

$m = -N, -N+1, \ldots, N-1, N$.

Величины h (шаг квантования) и N (количество шагов) являются дополнительными параметрами, которые могут выбираться независимо друг от друга и позволяют повысить точность приближения оценки $\hat{w}(V)$ к истинной плотности $w(V)$.

Множество точек $\{V_m\}$ образует в пространстве V^r гиперкубическую решетку, равномерно заполняющую гиперкуб со сторонами длиной $2Nh$, точка V находится в центре гиперкуба. Точки V_m не обязательно совпадают с векторами обучающей выборки $\{V_1, V_2, \ldots, V_L\}$.

Значения эмпирической функции распределения $\hat{F}_r(V_m)$ в точках V_m, $m = 0, \pm 1, \ldots, \pm N$ образуют поверхность в $(r+1)$-мерном пространстве, которая аппроксимируется затем гиперплоскостью D(V). Параметры гиперплоскости D(V) рассчитываются из условия минимума суммы квадратов отклонений D(V) от $\hat{F}_r(V_m)$:

$$\sum_{m=-N}^{N} \left[\hat{F}(V_m) - D(V)\right]^2 \to \min .$$

Для одномерной функции распределения ($r = 1$) при условии некоррелированности компонент вектора V полное выражение оценки плотности вероятности $\hat{w}(v)$ в точке v имеет вид

$$\hat{w}(v) = 3\sum_{m=-N}^{N} \frac{m \cdot \hat{F}(v + mh)}{N(N+1)(2N+1)h}.$$

Оценка *r*-мерной плотности вероятности получается как произведение одномерных плотностей:

$$\widehat{w}(V) = \prod_{j=1}^{r} \widehat{w}(v_j).$$

Аналогичную оценку можно получить и для общего случая, при наличии корреляции между компонентами вектора V. При этом по каждой мерности $j = \overline{1,r}$ выбирается свой размер шага h и количество шагов N. Эти выражения получены в [3] и из-за громоздкости здесь не приводятся. Там же показано, что при одновременном выполнении условий $h \to 0$ $N \to \infty$ среднеквадратичная ошибка аппроксимации оператора дифференцирования сходится к нулю.

Таким образом, использование операторных оценок плотности вероятности $\widehat{w}_r(V \mid s_i)$ в сравнении с другими известными непараметрическими методами оценивания плотности вероятности, а также методами прямого численного дифференцирования функций распределения $\widehat{F}_r(V \mid s_i)$ дает более высокую асимптотическую точность оценивания плотности $\widehat{w}_r(V \mid s_i)$.

Имея оценки плотности $\widehat{w}_r(V \mid s_i)$, $s_i = \overline{s_1, s_M}$, можно строить решающее правило для принятия аутентификационного решения.

В соответствии с выражением (1) отношение правдоподобия будет иметь вид

$$C = \frac{\widehat{w}_r(V \mid s_x)}{\widehat{w}_r(V \mid s_c)},$$

где $\widehat{w}_r(V \mid s_x)$ – оценка плотности распределения биометрических параметров $V^{(s_x)}$ неизвестного пользователя;

$\widehat{w}_r(V \mid s_c)$ – оценка плотности распределения биометрических параметров «своего» пользователя.

В итоге, решающее правило приобретает вид

$$s = \begin{cases} s_c, & \text{åñëè } \tilde{N} \geq \tilde{N}_\text{Ï}; \\ s_\div, & \text{åñëè } \tilde{N} < \tilde{N}_\text{Ï}, \end{cases}$$

где $C_п$ – значение порога, выбираемое с учетом ошибок первого рода.

Предложенный подход к аутентификации пользователей по КП отличается от известных наличием двухэтапной процедуры обучения. На первом этапе по обучающим выборкам строятся эмпирические функции распределения клавиатурных параметров, а на втором этапе по этим функциям формируются операторные оценки плотностей вероятностей, которые и становятся эталонными описаниями классов. Основное преимущество такого подхода состоит в том, что значения эмпирической функции распределения известны во всех точках области ее определения, что невозможно при прямом оценивании плотности распределения. Использование операторных оценок плотности распределения клавиатурных параметров в сравнении с другими непараметрическими методами дает существенно более

высокую асимптотическую точность оценивания плотности, а в конечном итоге и точность клавиатурной аутентификации.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

3. *Брюхомицкий Ю.А., Казарин М.Н.* Параметрическое обучение биометрических систем контроля доступа / Вестник компьютерных и информационных технологий. – М.: Изд-во Машиностроение, 2006. – № 2 (20). – С. 6–13.

4. *Брюхомицкий Ю.А.* Классификация нестационарных вероятностных биометрических параметров личности // Известия ЮФУ. Технические науки. – 2008. – №8 (85) – С. 147 – 154.

5. *Фомин Я.А., Тарловский Г.Р.* Статистическая теория распознавания образов. – М.: Радио и связь, 1986. – 264 с.

Брюхомицкий Юрий Анатольевич
Технологический институт Федерального государственного образовательного учреждения высшего профессионального образования «Южный федеральный университет» в г. Таганроге.
E-mail: bya@tsure.ru.
347928, г. Таганрог, ул. Чехова, 2.
Тел.: 8 (8634) 371-905.
Кафедра безопасности информационных технологий; доцент.

Bryukhomitsky Yuri Anatolyevich
Taganrog Institute of Technology – Federal State-Owned Educational Establishment of Higher Vocational Education “Southern Federal University”.
E-mail: bya@tsure.ru.
2, Chekhova st., Taganrog, 347928, Russia.
Phone: +7 (8634) 371-905.
Department of IT-Security; associate professor.

УДК 681.324

Г.Э. Абрамов

МОДЕЛЬ АНОМАЛЬНОГО ПОВЕДЕНИЯ СИСТЕМЫ НА ОСНОВЕ ВЕРОЯТНОСТНЫХ СУФФИКСНЫХ ДЕРЕВЬЕВ

Описывается метод применения вероятностных суффиксных деревьев для обнаружения аномального поведения программ. Используется «отпечаток» нормального поведения приложений с целью в дальнейшем обнаружить аномальное поведение как нечто, отклоняющееся от модели. В качестве основной модели используется вероятностные суффиксные деревья.

Вероятностное суффиксное дерево; PST, обнаружение аномального поведения.