ВЕСЦІ

НАЦЫЯНАЛЬНАЙ АКАДЭМІІ НАВУК БЕЛАРУСІ

СЕРЫЯ
ФІЗІКА-МАТЭМАТЫЧНЫХ
НАВУК

№ 3

АСОБНЫ АДБІТАК

Мінск 2002

ВЕСЦІ НАЦЫЯНАЛЬНАЙ АКАДЭМІІ НАВУК БЕЛАРУСІ №3 2002
СЕРЫЯ ФІЗІКА-МАТЭМАТЫЧНЫХ НАВУК

УДК 519.28

Н. Я. РАДЫНО

АЛГОРИТМ ИДЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЯ КОМПЬЮТЕРА ПО НАБОРУ ФИКСИРОВАННОЙ ФРАЗЫ НА КЛАВИАТУРЕ

Интенсивное развитие комьютерных технологий, средств связи и электронной обработки документов требует надежных средств защиты информации. Уже стали традиционными такие способы защиты информации, как электронно-цифровая подпись и шифрование документов. Из средств доступа к компьютеру или компьютерной сети используются магнитные карты, приборы идентификации пользователя по уникальным биометрическим признакам человека, а именно: сетчатка глаза, отпечатки пальцев, размеры ладони, голос и т.п. Упомянутые выше способы идентификации, очевидно, требуют дополнительных дорогостоящих устройств. Кроме того, для человека, проходящего процесс идентификации, такого сорта тестирование не останется незамеченным. И как правило, эта процедура является дискомфортной для человека. На практике часто используется способ идентификации пользователя по паролю. Хорошо известно, что наиболее уязвимым звеном в защите информации являются люди и их привычки. А именно, оказывается, что часто недобросовестные сотрудники организации, используя пароли своих коллег, получают доступ и к их информации, и к важной информации организации. Пароль своего коллеги нетрудно подсмотреть, а затем им воспользоваться. Чтобы исключить подобную ситуацию, нужно "научить" компьютер хотя бы с некоторой вероятностью различать манеру набора на клавиатуре сотрудников организации. Другими словами, возникает необходимость "научить" электронно-вычислительную машину различать пользователей — сотрудников организации.

В данной работе мы предлагаем один из способов идентификации пользователя компьютера. Способ идентификации основывается на статистических инвариантах манеры набора на клавиатуре заданной фразы. Преимущество такого способа идентификации заключается в том, что он может быть реализован незаметно для самого пользователя. Недостатком этого способа является то, что можно лишь с некоторой вероятностью утверждать о том, что фраза набрана тем или иным пользователем. Однако уже и такая информация во многих случаях является важной.

Ниже опишем постановку задачи и предложим алгоритм распознавания пользователя компьютера. Алгоритм проиллюстрируем на статистических данных, собранных в результате эксперимента над группой из восьми человек.

Итак, пусть имеется l пользователей компьютера, занумерованных от 1 до l. По каждому пользователю собираются статистические данные в виде матрицы $n \times m$, в которой фиксируется время перехода от клавиши к клавише при наборе определенной для всех пользователей парольной фразы (слова). Чтобы набрать фразу (слово), нужно сделать m переходов. Кроме того, каждый пользователь набирает заданную фразу n раз. После этого некий пользователь из группы набирает один раз парольную фразу (слово). Задача состоит в том, чтобы выяснить, каким пользователем вероятнее всего была набрана фраза (слово). Таким образом, нам надо выбрать среди l возможных пользователей одного или двух, трех, которые, вероятно, могли набирать заданную фразу (слово), причем следует опираться на данные, уже имеющиеся по каждому из l пользователей. Массив или матрицу, в которой будут хранится

данные о времени перехода от клавиши к клавише для каждого пользователя, обозначим через $S(i, j, k)$.

Для решения поставленной задачи мы применим модель и метод Байеса, связанные с проблемой классификации наблюдений в случае нескольких классов с известными априорными вероятностями [1, 2, 4]. Ниже рассмотрим упомянутую проблему классификации наблюдений и ее решение.

Пусть $\pi_1, \pi_2, \ldots, \pi_l$ — l классов с плотностями распределения вероятностей $p_1(x), p_2(x), \ldots, p_l(x)$ соответственно. Для того чтобы классифицировать наблюдение, необходимо разбить пространство наблюдений на l попарно непересекающихся областей $R_1, R_2, \ldots, R_l$. Если наблюдение попадает в область R_i, то говорят, что оно произведено над π_i. Пусть цена ошибочной классификации наблюдения, произведенного над π_i как наблюдения над π_j, равна $C(j|i)$. Вероятность этой ошибочной классификации равна

$$P(j|i) = \int_{R_j} p_i(x)\, dx. \tag{1}$$

Предположим, что известны априорные вероятности $q_1, \ldots, q_l$ того, что выборка произведена из соответствующего класса. Тогда математическое ожидание потерь, называемое риском классификации, будет равно

$$\sum_{i=1}^{l} q_i \left\{ \sum_{j=1, j\neq i}^{l} C(j|i) P(j|i) \right\}. \tag{2}$$

Заметим, что если $C(i|j) = 1 - \delta_{ij}$, то риск есть безусловная вероятность ошибки.

Области $R_1, \ldots, R_l$ мы желаем выбрать так, чтобы сделать риск (2) минимальным. Так как нам известны априорные вероятности, соответствующие каждому классу, то можно определить условную вероятность того, что наблюдение произведено над определенным классом при условии, что компоненты вектора x имеют данные значения. Условная вероятность того, что наблюдение произведено над классом π_i, равна

$$\frac{q_i p_i(x)}{\sum_{k=1}^{l} q_k p_k(x)}. \tag{3}$$

Если охарактеризовать наблюдение как наблюдение над π_j, то условное математическое ожидание потерь будет равно

$$\sum_{j=1, j\neq i}^{l} \frac{q_i p_i(x)}{\sum_{k=1}^{l} q_k p_k(x)} C(j|i). \tag{4}$$

Мы получим минимум математического ожидания потерь, если выберем j так, чтобы (4) было минимальным. Далее рассматривают сумму

$$\sum_{j=1, j\neq i}^{l} q_i p_i(x) C(j|i) \tag{5}$$

и выбирают j так, чтобы (4) было минимальным. Если минимум (4) достигается при двух различных значениях j, то можно выбрать любое из них. Этот метод относит точку x к одной из областей R_j. Повторяя его для каждой точки x, определяют области $R_1, \ldots, R_m$. Следовательно, метод классификации заключается в том, что наблюдение классифицируется как наблюдение над π_j, если его результаты попадают в R_j.

Теорема (см. [2, с. 199]). *Если априорная вероятность того, что наблюдение произведено над классом π_j с плотностью распределения вероятностей $p_i(x)$ $(i = 1, \ldots, l)$ равна q_r и цена ошибочной классификации этого наблюдения как наблюдения над π_j равна $C(j|i)$, то области классификации $R_1, \ldots, R_m$, при которых математическое ожидание потерь является минимальным, определяются следующим образом: область R_k состоит из тех точек x, для которых*

$$\sum_{i=1, i\neq k}^{l} q_i p_i(x) C(k|i) < \sum_{i=1, i\neq j}^{l} q_i p_i(x) C(j|i), \quad j = 1, \ldots, l, \ j \neq k. \tag{6}$$

Если (6) выполняется для всех индексов j $(j \neq k)$, за исключением h некоторых индексов, для которых неравенство заменяется равенством, то такая точка может быть отнесена к любому из $h + 1$ классов, соответствующих этим индексам. Если вероятность равенства между правой и левой частями (6) равна нулю для любых k и j при условии, что наблюдение произведено над π_i (i — любое), то метод, дающий минимум потерь, является единственным с точностью до множеств нулевой вероятности.

В предлагаемой выше теореме описан метод Байеса.

Далее перейдем непосредственно к изложению предлагаемого автором данной статьи алгоритма распознавания, приспособив для этого метод Байеса.

Шаг 1. Фиксируем выборку $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)$, произведенную пользователем, подлежащим идентификации.

Шаг 2. Априорные вероятности $q_1, q_2, \ldots, q_l$ того, что выборка произведена из соответствующего класса, полагаем равными $1/l$.

Шаг 3. Выделяем признак № 1, который характеризует манеру набора фразы. Мы считаем, что это *некоторая функция* от матрицы $S(i, j, k)$, k — фиксированное число. Напомним, что при фиксированном k $S(i, j, k)$ представляет собой матрицу $n \times m$, в которой содержатся данные измерений времени перехода от клавиши к клавише k-го пользователя. Для каждого из l пользователей по статистическим данным находим плотность распределения $p_k(x)$ этого признака. По выборке $\vec{\xi}$ находим значение указанного признака. Пусть оно равно x_1.

Шаг 4. Рассмотрим для каждого j сумму

$$\sum_{i=1, i\neq j}^{l} q_i p_i(x_1)(1 - \delta_{ij}) = p(x_1) - q_j p_j(x_1), \text{ здесь } C(i|j) = 1 - \delta_{ij} \ p(x_1) = \sum_{i=1}^{l} q_i p_i(x_1). \tag{7}$$

Выберем j так, чтобы (7) было минимальным. Если минимум (7) достигается при двух различных значениях, то выбираем оба эти значения. Метод распознавания заключается в том, что наблюдение x_1 классифицируется как наблюдение над π_j.

Шаг 5. Вычисляем условные вероятности того, что наблюдение x_1 произведено над классом π_i. Они равны

$$\frac{q_i p_i(x_1)}{\sum_{k=1}^{l} q_k p_k(x_1)}, \quad k = 1, 2, \ldots l. \tag{8}$$

На этом уже можно закончить. А можно сделать следующее.

Шаг 6. Априорные вероятности $q_1, q_2, \ldots, q_l$ того, что выборка произведена из соответствующего класса, полагаем равными условным вероятностям, вычисленным на пятом шаге.

Шаг 7. Выделяем признак № 2, который характеризует манеру набора фразы. Это опять-таки будет некоторая функция от матрицы $S(i, j, k)$, k — фиксированное число. Для каждого из l пользователей по статистическим данным находим плотность распределения $p_k(x)$ этого признака. По выборке $\vec{\xi}$ находим значение указанного признака. Пусть оно равно x_2. Затем аналогично проделываем шаг 4 и шаг 5 и так далее.

Пользователь (или пользователи), которому соответствует минимальное значение потерь после тестирования на признак (на несколько признаков), и будет искомым пользователем, которым, вероятно, и была набрана фраза (слово). А вероятность того, что фраза набрана указанным пользователем равна соответствующей вычисленной условной вероятности после тестирования на признак (на несколько признаков).

Существуют алгоритмы [3], отличные от предлагаемого здесь автором алгоритма. Однако, на наш взгляд, они имеют слабую теоретическую основу и носят полуэмпирический характер.

Рассмотрим пример. Обозначим через $\vec{\xi} = (\xi_1, \xi_2, \dots \xi_m)$ вектор значений времени m переходов при наборе неизвестным пользователем, подлежащим идентификации заданной парольной фразы (слова). Пусть $S(i,j,k)$ — массив, который при каждом фиксированном k (k изменяется от 1 до l) представляет собой матрицу $n \times m$, вектор-строка каждой такой матрицы есть значения времени перехода от клавиши к клавише при наборе слова "password" k-м пользователем. Здесь $l = 8$, $n = 10$, $m = 7$. В качестве примера рассмотрим только два признака (хотя их следует выбирать больше). Первый признак — время, затрачиваемое пользователем на набор фразы (слова "password"). Второй — среднее квадратическое отклонение времени, затрачиваемого пользователем на набор фразы (слова "password"). Это будут функции, которые каждой матрице $S(i,j,k)$ (k — фиксированное) ставят в соответствие вектор. Другими словами,

$$\begin{pmatrix} S(1,1,k) & S(1,2,k) & \dots & S(1,m,k) \\ \vdots & \vdots & \dots & \vdots \\ S(n,1,k) & S(n,2,k) & \dots & S(n,m,k) \end{pmatrix} \to \begin{pmatrix} \sum\limits_{j=1}^{m} S(1,j,k) \\ \vdots \\ \sum\limits_{j=1}^{m} S(n,j,k) \end{pmatrix}, \tag{9}$$

$$\begin{pmatrix} S(1,1,k) & \dots & S(1,m,k) \\ \vdots & \dots & \vdots \\ S(n,1,k) & \dots & S(n,m,k) \end{pmatrix} \to \begin{pmatrix} \sqrt{\dfrac{1}{m-1}\sum\limits_{s=1}^{m}\left(S(1,s,k) - \sum\limits_{j=1}^{m} S(1,j,k)\right)^2} \\ \vdots \\ \sqrt{\dfrac{1}{m-1}\sum\limits_{s=1}^{m}\left(S(n,s,k) - \sum\limits_{j=1}^{m} S(n,j,k)\right)^2} \end{pmatrix}. \tag{10}$$

Найдем плотности распределений интересующих нас признаков по вектор-столбцам (9), (10) в предположении, что рассматриваемые распределения нормальны. Последнее, вообще говоря, неверно в общем случае. Параметры распределений для каждого пользователя указаны в табл. 1 и 2, (a_i — выборочное среднее значений признака, соответствующее i-му пользователю, σ_i — выборочное среднее квадратичное отклонение значений признака, соответствующее i-му пользователю), а графики плотностей распределений — на рис. 1 и 2.

Таблица 1. **Параметры распределений первого признака для каждого пользователя**

№	1	2	3	4	5	6	7	8
a_i	140,60	442,30	267,50	218,50	359,80	256,60	295,90	312,20
σ_i	27,29	89,85	61,32	23,16	94,35	24,45	67,64	122,08

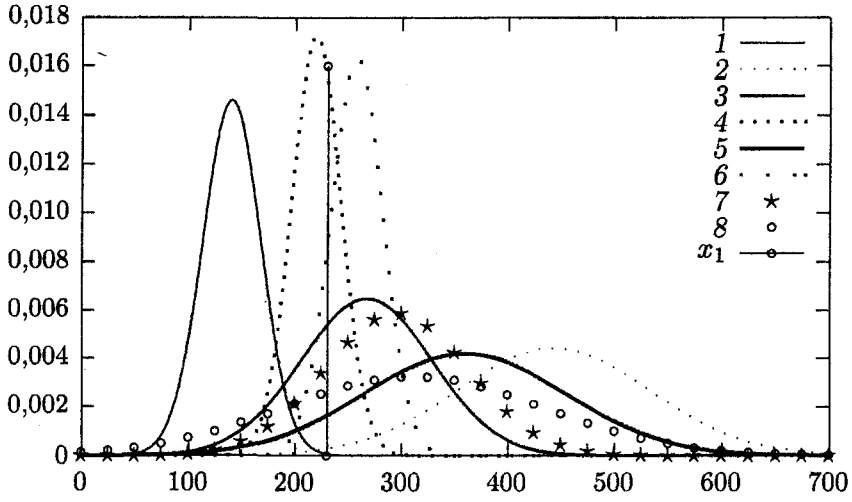

Рис. 1. Графики плотностей распределений первого признака

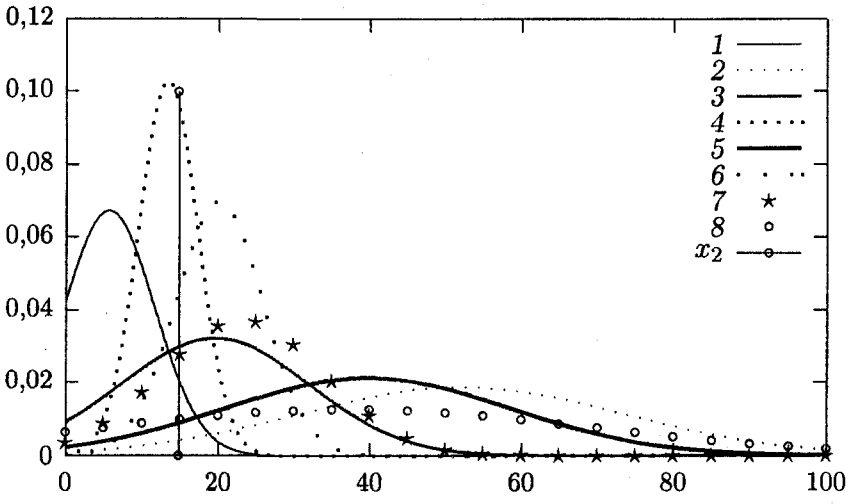

Рис. 2. Графики плотностей распределений второго признака

Таблица 2. **Параметры распределений второго признака для каждого пользователя**

№	1	2	3	4	5	6	7	8
a_i	5,72	52,08	19,65	13,26	39,77	20,01	23,23	37,44
σ_i	5,94	21,29	12,39	3,88	18,77	5,73	10,70	31,70

Вычисляем $x_1 = \sum_{j=1}^{m} \xi_j$ и $x_2 = \sqrt{\frac{1}{m-1} \sum_{s=1}^{m} \left(\xi_s - \sum_{j=1}^{m} \xi_j \right)^2}$. Оказывается, что $x_1 = 230$, $x_2 = 14.84$. Полагаем, что $q_1 = q_2 = \ldots = q_8 = \frac{1}{8}$.

Сначала рассмотрим классификацию по первому признаку, для этого выбираем j так, чтобы $\sum_{i=1,\ i \neq j}^{l} q_i p_i(x_1)$ было минимальным (см. выражение (4)). В рассматриваемом случае

$j = 4$. Тогда наблюдение $\vec{\xi}$ классифицируем как наблюдение над π_4 по первому признаку. Математическое ожидание потерь равно: $M_{\pi_1} = 0,9982$, $M_{\pi_2} = 0,9928$, $M_{\pi_3} = 0,8577$, $M_{\pi_4} = 0,5983$, $M_{\pi_5} = 0,9567$, $M_{\pi_6} = 0,7618$, $M_{\pi_7} = 0,9032$, $M_{\pi_8} = 0,9313$.

Условные вероятности Q_i того, что наблюдение произведено над классом π_i, равны:

$Q_1 = 0,0018$, $Q_2 = 0,0072$, $Q_3 = 0,1423$, $Q_4 = 0,4017$, $Q_5 = 0,0433$, $Q_6 = 0,2382$, $Q_7 = 0,0968$, $Q_8 = 0,0687$.

Для сравнения рассмотрим классификацию по второму признаку. Для этого выбираем j так, чтобы $\sum_{i=1, i\neq j}^{l} q_i p_i(x_2)$ было минимальным (см. выражение (4)). В этом случае опять получаем $j = 4$. Наблюдение $\vec{\xi}$ классифицируем как наблюдение над π_4 по второму признаку. Математическое ожидание потерь равно: $M_{\pi_1} = 0,9145$, $M_{\pi_2} = 0,9832$, $M_{\pi_3} = 0,8764$, $M_{\pi_4} = 0,6081$, $M_{\pi_5} = 0,9636$, $M_{\pi_6} = 0,8082$, $M_{\pi_7} = 0,8865$, $M_{\pi_8} = 0,9596$. Условные вероятности Q_i того, что наблюдение произведено над классом π_i, равны: $Q_1 = 0,0855$, $Q_2 = 0,0168$, $Q_3 = 0,1236$, $Q_4 = 0,3919$, $Q_5 = 0,0364$, $Q_6 = 0,1918$, $Q_7 = 0,1135$, $Q_8 = 0,0404$.

Заметим, что качество работы алгоритма в большóй степени зависит от того, какие задать признаки и в какой степени близка друг к другу динамика набора слов пользователями.

Автор признателен профессору Ю.С. Харину за внимание к работе.

Summary

The algorithm of a PC user identification by his keystroke dynamics is proposed. Motivation of proposed algorithm is discussed. An example of manipulation of statistical data taken from the eight PC users is demonstrated.

Литература

1. Айвазян С. А, Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Основы моделирования и первичная обработка данных. М., 1983.

2. Андерсон Т. Введение в многомерный статистический анализ. М., 1963.

3. Расторгуев С. П. Программные методы защиты информации в компьютерах и сетях. М., 1993.

4. Харин Ю. С. Робастность в статистическом распознавании образов. Мн., 1992.

Институт математики
НАН Беларуси

Поступила в редакцию
12.03.2001