УДК 004.75 : [004.65]

Л.Э. Чалая

Харьковский национальный университет радиоэлектроники

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ АУТЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЕЙ КОМПЬЮТЕРНЫХ СИСТЕМ ПО КЛАВИАТУРНОМУ ПОЧЕРКУ

В статье рассматриваются различные подходы к аутентификации пользователей компьютерных систем по особенностям их клавиатурного почерка. Первый из подходов основан на анализе статистических характеристик клавиатурного почерка. Второй подход учитывает ритм работы пользователя на клавиатуре, а третий основывается на изменении темпа работы.
Результаты сравнительного анализа этих методов показывают, что наибольший эффект достигается при их комбинированном использовании.

биометрия, динамика клавиатурного почерка, классификация

Введение

В последнее время повышаются требования к безопасности доступа к ресурсам информационных систем. В связи с этим получили развитие методы, основанные на биометрической аутентификации пользователей. Однако эти методы достаточно чувствительны к особенностям конкретных компьютеров. Поэтому вызывает интерес применение подходов, которые были бы универсальными для всех компьютеров. Рассмотрим аутентификацию, базирующуюся на особенностях клавиатурного набора текстов пользователем, которые будем именовать клавиатурным почерком [1]. В этой статье рассматриваются три метода, позволяющих аутентифицировать пользователя на основании анализа клавиатурного набора тестовой последовательности, состоящей из пароля и ключевой фразы. Затем оценивается возможность их комбинированного использования с целью улучшения эффективности аутентификации. В области аутентификации или верификации применяется три коэффициента, широко используемых для оценки качества биометрических систем:

– коэффициент ложного отказа в доступе зарегистрированному пользователю k_1;

– коэффициент ложного предоставления доступа злоумышленнику k_2;

– коэффициент равных ошибок k_3, соответствующий ситуации, при которой k_1 и k_2 равны.

Постановка задачи

Регистрируемые данные. Регистрация данных, характеризующих клавиатурный почерк пользователя, связана с оценкой интервалов времени, разделяющих отдельные состояния клавиатуры. К таким состояниям относятся нажатие и отпускание клавиш. На рис. 1 приведен пример, иллюстрирующий последовательность формирования таких состояний для сочетания символов клавиатуры «М» и «А».

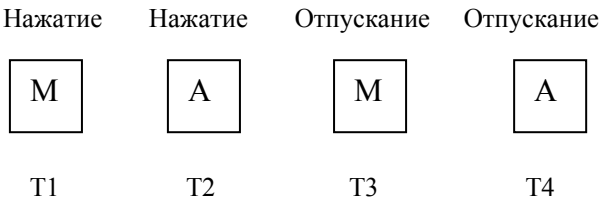

Рис. 1. Временные характеристики для последовательного набора символов М, А

Очевидно, что последовательный набор двух любых символов клавиатуры характеризуется следующим набором ситуаций:

– «Н-Н» (нажатие-нажатие): интервал времени между двумя нажатиями клавиш (Т2-Т1);

– «Н-О» (нажатие-отпускание): интервал времени между прикосновением к клавише и ее отпусканием (Т3-Т1 и Т4-Т2);

– «О-Н» (отпускание-нажатие): интервал времени между отпусканием одной клавиши и нажатием следующей (Т3-Т2);

– «О-О» (отпускание-отпускание): интервал времени между последовательными отпусканиями двух клавиш.

В дальнейшем будем рассматривать совокупность временных характеристик, соответствующих ситуациям НО, НН, ОН и ОО для ограниченной последовательности нажатий клавиш пользователем.

Анализ литературных источников. Первая работа, в которой была рассмотрена задача аутентификации пользователей компьютера по клавиатурному почерку, появилась в 1988 году [2]. Здесь исследовалась работа на клавиатуре 7 машинисток, набиравших один и тот же отрывок текста в разное время. В ней показана возможность их идентификации с помощью Т-теста. Т-тест представляет собой

статистический программный модуль, позволяющий подтвердить примерное равенство математических ожиданий и дисперсий двух статистических выборок. Эти первые результаты были обнадеживающими, но не применимыми для практики из-за малого количества участников эксперимента и недостаточного объема тестового теста. В дальнейшем были проведены многочисленные исследования по этой тематике, некоторые из которых представлены в [3].

В работе [4] представлен новый подход к аутентификации по клавиатурному почерку, основанный на анализе функции плотности распределения вероятностей случайных переменных. Этот метод разделяется на четыре этапа: создание модели, группирование наблюдений, оценка критериев и принятие решений. Модель включает в себя временные характеристики НН, НО и ОН, определяемые по набору на клавиатуре текущего слова. Группирование наблюдений осуществляется с помощью классификации по возрастающей иерархии, использующей в качестве расстояния между двумя профилями евклидово расстояние. Затем выделяются значения времен, которые принимаются во внимание при формировании профиля, при этом к ним относятся только те, что находятся в группе, содержащей более 75% значений времени, исходя из иерархической классификации. Это позволяет выделить данные, которые являются репрезентативными для конкретного пользователя, и удалить данные, которые можно считать некорректными наблюдениями регистрационного эксперимента. Профиль характеризуется вектором математических ожиданий μ и отклонений σ. Вектор V_i, определяющий репрезентативность i-й характеристики, задается как

$$V_i = m_i / n_i,$$

где n_i – общее количество данных, зарегистрированных для i-й характеристики; m_i – количество данных, отобранных для последующего вычисления i-й характеристики. Затем рассматривается гипотеза, в соответствии с которой данные наблюдений распределены по закону Коши:

$$F_{\mu,\sigma}(x) = \frac{\sigma^3}{(\sigma^3 + |x-\mu|^3)}.$$

Исходя из этого, формируется функционал

$$\Psi_i = \alpha_0 \cdot F_{\mu_i,\sigma_i}(x) \cdot V_i,$$

где α_0 – коэффициент, используемый для нормализации весов.

Затем вычисляется сумма значений таких функционалов, составляющих глобальный функционал:

$$\Psi = \sum \Psi_i .$$

Значения глобальных функционалов также предполагаются распределенными по закону Коши с математическими ожиданиями и дисперсиями, вычисленными для всех пользователей, что позволяет определить следующую суммарную оценку:

$$I = \alpha V_1 + \beta V_2 + \gamma V_3 + \delta V_4,$$

где

$$V_k = F_{\mu\sigma}(S_k),\ k = \overline{1,4},$$

S_k – оценки, полученные для вектора характеристик НН, НО, ОН и ОО; $\alpha, \beta, \gamma, \delta$ – некоторые параметры.

Далее суммарная оценка сравнивается с допустимым значением для отбора или удаления соответствующего наблюдения. В соответствии с этим методом коэффициент ложного отказа в доступе зарегистрированному пользователю k_1 не превышал 1%, а коэффициент ложного предоставления доступа злоумышленнику k_2 не превышал 4,4%.

В работе [4] предпринята попытка использовать метод, обычно применяемый в задачах классификации и именуемый классификатором с векторным базисом (КВБ). Для адаптации этого метода авторы вводят гауссовское ядро, пытаясь максимизировать диапазон возможного использования КВБ. По стандартному КВБ были получены значения $k_1 = 15{,}78\%$ и $k_2 = 0\%$. Недостатком этого метода является большой объем данных, необходимых для каждого пользователя при обучении (50 выборок).

В работе [5] предлагается подход, основанный на использовании искусственных нейронных сетей (ИНС) и содержащий три этапа. Первый этап состоит в фильтрации данных, необходимых для обучения. Второй этап реализует процедуру настройки ИНС для разделения биометрических характеристик пользователей. Третий этап позволяет тестировать новые наблюдения. Для первого этапа используется сеть Кохонена, в которую вводятся данные злоумышленников. Текущие наблюдения пользователей относят к одному из компактных классов или же удаляют данные злоумышленников. На втором этапе используется ИНС типа АДАЛИНА. На третьем этапе в сеть вводится новое наблюдение, а на выходе сети формируется ответ, на основании которого принимается решение об аутентичности. По такому методу были получены значения $k_1 = 12{,}2\%$ и $k_2 = 2{,}12\%$. Основным недостатком метода является необходимость длительного обучения ИНС. Авторы рассматривают возможность сокращения количества данных, требуемых для обучения, путем их искусственного генерирования с использованием случайных чисел или модификации данных пользователя.

Задачами настоящей работы являются модификация основных подходов к созданию методов аутентификации пользователей компьютерных систем по клавиатурному почерку и анализ их комбинированного использования.

Модифицированные методы аутентификации пользователей

Предлагаемые методы аутентификации пользователей компьютерных систем по клавиатурному почерку основываются на некоторых идеях рассмотренных выше подходов. Первый из этих методов базируется на анализе статистических данных (математического ожидания и среднеквадратичного отклонения значений биометрических характеристик). Второй метод основан на измерении расстояния между векторными профилями пользователей, а третий – на дискретизации времени, характеризующего темп работы пользователей

Метод, основанный на анализе статистических характеристик. Впервые подобный метод был использован для аутентификации пользователей в работе [4]. Здесь необходимые исходные данные формировались по результатам компьютерного набора некоторого текста. В процессе создания биометрического профиля определялись математические ожидания и среднеквадратичные отклонения каждой временной характеристики. Затем вычислялось значение μ_0, для которого выполняется неравенство:

$$|\mu_0 - \mu| < 0,5\sigma .$$

Аутентичность пользователя подтверждалась, если не менее 60% временных характеристик, определяемых в процессе клавиатурного набора, удовлетворяли приведенному неравенству. Были протестированы 17 пользователей. По такому методу были получены значения $k_1 = 5,5\%$ и $k_2 = 5\%$.Основным недостатком метода является большой объем текста, который необходимо набрать каждым пользователем (около 1000 нажатий клавиш) для получения приемлемого результата.

В работе [5] этот метод был исследован для фиксированной парольной фразы. В тест при этом вводился коэффициент сложности. Характеристика считалась подтверждающей аутентичность пользователя, если выполняется неравенство:

$$|\mu_0 - \mu| < \alpha_1\sigma .$$

Этот метод сводится к определению туннеля вокруг средних значений, связанных с различными характеристиками, ширина которого определяется одновременно значением среднеквадратичного отклонения и коэффициентом α_1.

По такому методу были получены значения $k_1 = 5\%$ и $k_2 = 30\%$ при варьировании коэффициента α_1.

В настоящей работе предлагаются модификации этого метода, который заключается в адаптации параметров к каждому пользователю. Предлагаемый метод вначале фиксирует значение α_1 равным 1, а затем уменьшает его, чтобы учесть все шаги обучения. Текущий профиль корректируется путем пересчета по десяти последним достоверным значениям.

Второе улучшение состоит в использовании дополнительного взвешивания, состоящего в том, что перед принятием решения об учете или удалении значения временной характеристики определяется оценка функции его расстояния до среднего значения. Эта оценка находится в интервале [0,1] и определяется по следующей зависимости:

$$I = e^{-\frac{|\mu_0 - \mu|}{\sigma}} .$$

Для определения такой глобальной оценки рассчитывается среднее значение оценок для всех временных характеристик. Затем глобальная средняя оценка сравнивается с заданным порогом. Преимущество предложенной модификации: минимизируется краевой эффект и используется лишь один регулируемый параметр (конечный порог).

Метод, основанный на анализе изменения ритма. Известно, что в музыке ритм мелодии определяется продолжительностью нот, задаваемой некоторыми символами. По аналогии можно проанализировать ритм работы на клавиатуре пользователя компьютера, т.е. рассмотреть не только отдельные значения временных характеристик клавиатурного почерка, но и интервалы времени между этими значениями. Для этого можно использовать множество очевидных методов. Рассмотрим подход, основанный на разделении значений временных характеристик на очень короткие, короткие, средние, длинные и очень длинные. Такую классификацию можно осуществить разными способами. Первый из них заключается в задании порогов, ограничивающих различные классы, например:

- $t > 200$: класс 1;
- $200 > t > 100$: класс 2;
- $100 > t > 70$: класс 3;
- $70 > t > 30$: класс 4;
- $30 > t$: класс 5.

Недостатком такого способа классификации является необходимость выбора границ интервалов для дискретизации произвольной средней скорости компьютерного набора пользователями. Второй подход состоит в классификации лишь одного наблюдения в соответствии с функцией временных интервалов между значениями временных характеристик. И, наконец, можно классы не ограничивать фиксированными порогами, а использовать только некоторую функцию разделения.

Независимо от выбранного способа классификации в процессе эксперимента для каждой временной характеристики рассчитывался профиль, и значения, полученные в процессе обучения, заносились в определенный класс. Без ограничения общности, в приводимых далее результатах рассматривается формирование профиля, в основном, по временной

характеристике ОН, так как она является наиболее информативной. Затем для сравнения текущего наблюдения с профилем необходимо сформировать функции расстояния между текущей характеристикой и классом, к которому относится профиль. Такое расстояние можно выбрать как разность между номерами классов. Например, если значение временной характеристики, соответствующей классу 4, равно 2, то расстояние, учитываемое при создании профиля, будет равно 2. Затем определяется сумма таких расстояний для получения оценки наблюдения, после чего эта оценка сравнивается с некоторым пороговым значением для принятия решения о достоверности наблюдения.

Метод, основанный на классификации временных характеристик. Сравнение двух клавиатурных почерков может быть основано на анализе разностей порядков временных характеристик, т.е. на определении, какая их таких характеристик является более короткой. Это можно реализовать путем измерения расстояния между некоторыми двумя векторами. Для измерения соответствующей разности порядков временных характеристик осуществляют классификацию этих характеристик по результатам каждого наблюдения от наиболее длинной к наиболее короткой. Профиль, который должен быть представлен как некоторая усредненная характеристика для каждого биометрического параметра, рассчитывается по данным обучения. Для определения оценки одного наблюдения существуют различные способы: первый (наиболее простой) основан на определении евклидова расстояния между профилем и текущим наблюдением; второй способ заключается в определении коэффициента ранговой корреляции Спирмена между двумя значениями характеристик, который широко используется на практике. Обозначим через r_i^1 ранг времени в классе 1 и через n – количество временных характеристик в классификации. Тогда коэффициент корреляции Спирмена можно представить следующим образом:

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} (r_i^1 - r_i^2)^2}{n \cdot (n^2 - 1)}.$$

Ни один из рассмотренных методов нельзя считать полностью удовлетворительным, так как динамика компьютерного набора часто характеризуется значениями учитываемых временных параметров, которые значительно превышают соответствующие эталонные значения профилей (например, из-за непредвиденных пауз в работе пользователей), что влияет не только на текущий ранг наблюдения, но и на ранги всех временных характеристик, задающих связь между текущим рангом наблюдения Ro и соответствующим рангом в профиле Rp. Предлагаемый ниже подход позволяет частично решить эту проблему. Он основан на классификации временных характеристик по повышению рангов наблюдений. В общую оценку I добавляется разность /Rp – Ro/ , что позволяет скорректировать общий результат.

Качественные характеристики методов

В тестировании качества аутентификации принимали участие 15 пользователей, набиравших текст, содержащий пароль и одну для всех тестовую фразу (всего 25 нажатий буквенных клавиш). Вначале каждый из тестируемых набрал этот текст десять раз для реализации процедуры обучения (формирования профиля), а затем (через заданные временные интервалы) для текущей аутентификации. Тестирование проводилось на протяжении шести месяцев и включало от 10 до 100 сеансов для каждого пользователя. Для определения коэффициента k_1 использовались все сеансы каждого пользователя, а для определения коэффициента k_2 в качестве злоумышленников рассматривались сеансы всех других пользователей. Средние значения коэффициентов качества аутентификации приведены в табл. 1. Пороговые значения были экспериментально подобраны таким образом, чтобы получить наименьшее значение коэффициента k_3. Тестирование было проведено для всех рассмотренных выше методов.

Таблица 1

Средние значения коэффициентов ошибок (%)

Метод	Среднее значение коэффициента k_2	Среднее значение коэффициента k_1	k_3
Математическое ожидание, дисперсия	4,39	4,81	4,6
Математическое ожидание, дисперсия + дополнительное взвешивание	3,62	3,61	3,62
Ранговая корреляция Спирмена	4,69	4,69	4,69
Коррекция рангов	3,56	3,62	3,59
Ритм клавиатурного набора с пороговым заданием классов	3,47	3,39	3,43
Ритм клавиатурного набора с пропорциональным заданием классов	3,55	4,02	3,79

Результаты тестирования показали, что наиболее предпочтительным является метод, основанный на анализе изменения ритма с фиксированными порогами, которому соответствует наименьшее значение коэффициента ошибок k_3. Следует, однако, отметить, что даже при хороших средних значениях коэффициентов ошибок система аутентификации

может оказаться малоэффективной, если хотя бы один из пользователей компьютерной системы не будет идентифицирован в процессе работы. Поэтому представляет интерес анализ максимальных значений коэффициентов качества аутентификации, полученных при тестировании, которые приведены в табл. 2. Из этой таблицы следует, что максимальные значения коэффициентов ошибок весьма значительны. Это вызвано тем, что среди тестируемых пользователей присутствовали пользователи, для которых были получены аномально большие значения анализируемых коэффициентов.

Таблица 2

Максимальные значения коэффициентов ошибок, (%)

Метод	Максимальное значение коэффициента k_2	Максимальное значение коэффициента k_1
Математическое ожидание, дисперсия	27,14	13
Математическое ожидание, дисперсия + дополнительное взвешивание	21,59	10
Ранговая корреляция Спирмена	19,02	24
Коррекция рангов	11,51	18
Ритм клавиатурного набора с пороговым заданием классов	10,89	18
Ритм клавиатурного набора с пропорциональным заданием классов	9,76	18

Для выявления таких аномалий целесообразно провести анализ результатов тестирования по каждому из пользователей для всех рассмотренных методов. Этот анализ дает примерно сходные результаты по каждому из методов.

Примерно половине пользователей соответствуют хорошие результаты аутентификации (k_3 менее 2,5%). Большинству пользователей из второй половины соответствуют приемлемые результаты аутентификации (k_3 менее 5%). И лишь один или два пользователя (в зависимости от метода) характеризуются значением k_3, превышающим 10%. В частности, для метода, основанного на анализе статистических характеристик (табл.а 3), четыре пользователя имеют коэффициенты k_1 и k_2, превышающие 5%. Один из этих пользователей имеет особенно высокий уровень коэффициента k_2 (21,59 %). Дополнительный анализ его клавиатурного почерка показал высокую скорость набора в среднем, но с большим разбросом, что является весьма благоприятным для атак злоумышленников.

Таблица 3

Показатели аутентификации пользователей для метода, основанного на анализе статистических характеристик, (%)

№ пользователя	k_2	k_1	k_3
1	0,00	6,74	3,37
2	5,38	2,00	3,69
3	8,14	0,00	4,07
4	0,12	0,00	0,06
5	0,84	0,00	0,42
6	21,59	0,00	10,80
7	8,39	4,00	6,20
8	0,00	1,25	0,63
9	0,13	3,00	1,57
10	0,00	2,00	1,00
11	1,67	10,00	5,84
12	0,00	9,00	4,50
13	0,80	9,00	4,90
Среднее значение	3,62	3,61	3,62
Максимальное значение	21,59	10,00	15,80
Минимальное значение	0,00	0,00	0,00

Для метода ранговой корреляции были получены глобальные оценки, близкие к глобальным оценкам метода, основанного на анализе статистических характеристик. В то же время распределения таких оценок по отдельным пользователям существенно разнятся для этих методов, что можно видеть из табл. 3 и 4.

Таблица 4

Показатели аутентификации пользователей для метода ранговой корреляции (%)

№ пользователя	k_2	k_1	k_3
1	0,26	0,00	0,13
2	0,75	2,00	1,38
3	11,51	18,00	14,76
4	1,07	0,00	0,54
5	3,98	10,00	6,99
6	8,32	0,00	4,16
7	9,51	10,00	9,76
8	1,04	0,00	0,52
9	0,40	2,00	1,20
10	0,88	0,00	0,44
11	2,98	0,00	1,49
12	2,54	1,00	1,77
13	3,07	4,00	3,54
Среднее значение	3,56	3,62	3,59
Максимальное значение	11,51	18,00	14,76
Минимальное значение	0,26	0,00	0,13

Для метода, основанного на анализе изменения ритма (табл. 5), аномально большие значения коэффициентов ошибок также присущи четырем пользователям.

Таблица 5

Показатели аутентификации пользователей для метода, основанного на анализе изменения ритма, (%)

№ пользователя	k_2	k_1	k_3
1	0,00	1,12	0,56
2	0,75	2,00	1,38
3	10,64	2,00	6,32
4	1,07	0,00	0,54
5	3,74	5,00	4,37
6	8,44	5,00	6,72
7	10,89	18,00	14,45
8	1,17	0,00	0,59
9	0,27	5,00	2,64
10	0,13	2,00	1,07
11	0,48	0,00	0,24
12	5,07	0,00	2,54
13	2,40	4,00	3,20
Среднее значение	3,47	3,39	3,43
Максимальное значение	10,89	18,00	14,45
Минимальное значение	0,00	0,00	0,00

Таким образом, результаты тестирования различных методов показывают близость средних значений коэффициентов ошибок, но существенно разнятся для конкретных пользователей. Представляется целесообразным рассмотреть возможность комбинирования методов для повышения качества работы биометрической системы аутентификации по клавиатурному почерку.

Комбинированный метод

Рассмотрим возможность создания комбинированного алгоритма аутентификации. Существуют различные подходы для реализации процедуры объединения классификаторов, лежащих в основе методов идентификации ситуаций, например, использование теории очевидности [9, 11]. В системах аутентификации принятие решений сводится к проблеме формирования на выходе только двух состояний: «да» или «нет». Существует три группы методов такой классификации, позволяющих получить численную оценку в интервале между 0 и 1. К первой группе методов относятся методы, основанные на фиксированной схеме выбора. При этом каждый из используемых классификаторов принимает свое локальное решение («да» или «нет»). Например, для трех классификаторов существуют три схемы выбора: полное совпадение результатов (все три классификатора формируют положительный результат); большинство совпавших результатов (два из трех классификаторов формируют положительный результат); минимально достаточный результат (хотя бы один из трех классификаторов формирует положительный результат). Вторая группа методов основана на использовании взвешенной комбинации оценок. Трудности их применения состоят в том, что распределение оценок здесь весьма чувствительны к статистическим оценкам и к особенностям классификаторов. Чтобы иметь возможность комбинирования оценок, необходимо осуществить их предварительную нормализацию. Наиболее распространенными разновидностями нормализации являются:

– нормализация по известному максимуму

$$I' = \frac{I}{I_{max}};$$

– нормализация по максимальному диапазону изменения оценок

$$I' = \frac{I - I_{min}}{I_{max} - I_{min}}.$$

Отметим, что для такой нормализации используется только информация об известных данных, поэтому после нормализации могут получиться оценки, превышающие 1, или же отрицательные. Это происходит, если наблюдение является очень близким к эталонному профилю или же очень удаленным от него. В этом случае такие наблюдения попросту отбрасываются.

– нормализация по z-оценке

$$I' = \frac{\left|I - \bar{I}\right|}{\sigma_I},$$

где $\bar{I}$ – среднее значение оценок для всех методов.

Нормализация по z-оценке не всегда приводит к оценкам, находящимся в интервале между 0 и 1. Недостатком такого оператора нормализации является необходимость априорного знания распределения оценок, а, следовательно, их максимума, минимума, математического ожидания и дисперсии. Это не имеет значения, если при аутентификации используется одна и та же парольная фраза для всех пользователей. В противном же случае оценивание математического ожидания и дисперсии по текущим данным и последующая нормализация являются проблематичными при малом количестве сеансов обучения.

После проведения нормализации можно перейти к выбору процедуры комбинированного использования методов аутентификации. Обозначим через I' оценку, полученную по i-му классификатору. Предлагаемый подход предполагает разделение решений на два класса, поэтому

$$P_1 = 1 - P_2,$$

где P_1 – вероятность подтверждения личности пользователя, P_2 – вероятность отказа в доступе пользователю.

Чтобы рассматривать не вероятности, а нормализованные оценки, положим:

$$P_{1i} = I^i,$$

где i – индекс пользователя.

Поскольку априори вероятности исходов «да» или «нет» неизвестны, то их нельзя непосредствен-

но включать в процедуру, но можно использовать некоторые пороговые значения, с которыми будет сравниваться выходная оценка.

Основными операторами, используемыми в предлагаемом комбинированном методе, являются:

– операторы максимума и минимума

$$I = \max_i (I^i);$$

$$I = \min_i (I^i);$$

– медианная оценка

$$I = \operatorname{median}_i (I^i);$$

– произведение оценок

$$I = \prod_i (I^i);$$

– сумма оценок

$$I = \sum_i (I^i).$$

Результаты анализа комбинированного метода

Первый тест такого анализа состоял в исследовании по оператору суммы различных процедур нормализации оценок (табл. 6). Наилучшие результаты получены при нормализации по z-оценке, поэтому в дальнейшем будем рассматривать только такой вариант нормализации.

Таблица 6

Показатели аутентификации пользователей для различных процедур нормализации оценок, (%)

Нормализация	k_2	k_1	k_3
По максимуму	1,75	2,46	2,11
По максимальному диапазону	2,00	2,00	2,00
По z-оценке	1,81	1,69	1,75

Результаты применения различных операторов комбинирования приведены в табл. 7.

Таблица 7

Показатели аутентификации пользователей для различных операторов комбинирования, (%)

Метод	k_2	k_1	k_3
полное совпадение результатов	1,17	7,92	4,55
большинство совпавших результатов	2,39	2,15	2,27
минимально достаточный результат	7,60	0,54	4,07
произведение оценок	2,00	2,00	2,00
медианная оценка	3,34	3,39	3,37
максимальная оценка	3,62	3,61	3,62
минимальная оценка	3,62	3,62	3,62
сумма оценок	1,81	1,69	1,75

Анализ табл. 7 свидетельствует о том, что применение комбинированных методов существенно улучшает результаты аутентификации по сравнению с локальным применением рассмотренных ранее методов. Наилучшие результаты получены для классификатора, работающего по схеме выбора «большинство совпавших результатов», а также для операторов произведения и суммы оценок. Однако нельзя сравнивать различные варианты комбинированного метода лишь по значениям коэффициентов ошибок. Следует также принимать во внимание необходимость использования в каждом из вариантов дополнительной информации. Например, метод принятия решений по большинству совпавших результатов требует задания трех пороговых значений. Метод суммирования требует определения математического ожидания и дисперсии оценок для нормализации, а также фиксации порового значения для процедуры принятия решений. Метод произведения требует использования лишь одного порового значения. Он является более перспективным для практического применения, чем метод суммирования, несмотря на более низкие качественные характеристики. Это связано с тем, что при переменной длине парольных фраз трудно реализовать нормализацию оценок перед применением оператора суммирования.

Результаты применения оператора суммирования для тестируемых пользователей приведены табл. 8.

Таблица 8

Показатели аутентификации пользователей для метода, использующего оператор суммирования оценок, (%)

№ пользователя	k_2	k_1	k_3
1	0,00	0,00	0,00
2	0,25	2,00	1,13
3	4,38	0,00	2,19
4	0,00	0,00	0,00
5	0,60	5,00	2,80
6	7,48	0,00	3,74
7	8,39	8,00	8,20
8	0,00	0,00	0,00
9	0,00	2,00	1,00
10	0,00	0,00	0,00
11	0,24	0,00	0,12
12	0,93	1,00	0,97
13	1,20	4,00	2,60
Среднее значение	1,81	1,69	1,75
Максимальное значение	8,39	8,00	8,20
Минимальное значение	0,00	0,00	0,00

Табл. 8 свидетельствует об общем снижении коэффициентов ошибок для пользователей при использовании комбинированного метода по сравнению с применением рассмотренных ранее локальных методов аутентификации, хотя для отдельных пользователей комбинирование привело к некоторому ухудшению результатов (пользователи 5 и 7).

Следует отметить существенное снижение числа коэффициентов ошибок, значение которых превышает 8%. Лишь для двух пользователей значение k_1 превысило 5% и ни для одного из пользователей это значение не превысило 10%. Эти результаты являются достаточными, чтобы случаи некорректного отказа в доступе санкционированным пользователям были редкими. В то же время для четырех пользователей значение k_2 превысило 1%, а для двух 5%, что является слишком высоким показателем, так как вместо этих пользователей злоумышленник сможет с большой вероятностью получить несанкционированный доступ к ресурсам компьютера, зная чужой пароль, являющийся, как правило, коротким. Поэтому в ходе тестирования было проведено варьирование порогового значения, используемого для принятия решений по аутентификации, чтобы повысить безопасность компьютерной системы.

Результаты соответствующего исследования представлены на рис. 2.

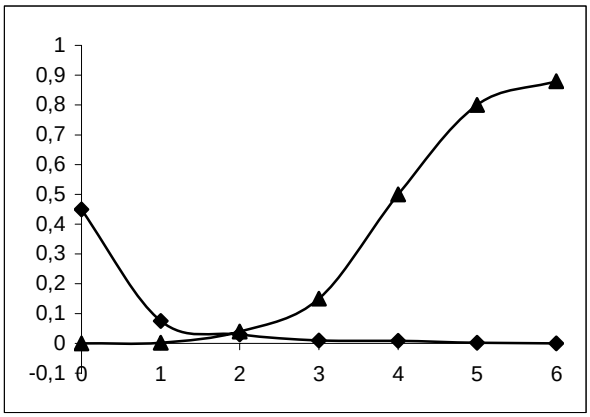

Рис. 2. Зависимость коэффициентов ошибок k_1 и k_2 от пороговых значений

Полученные кривые являются характерными для биометрической идентификации, когда низкому значению порога соответствуют высокое значение k_2 и низкое значение k_1. С возрастанием величины порога k_2 уменьшается, а k_1 увеличивается. Пересечение двух кривых происходит для порога, равного 1,7, при величине коэффициентов ошибок 1,8%, соответствующей значению k_3. Для порога ниже, чем 0,15, значения k_1 и k_2 соответственно равны 0% и 37%, что является неприемлемым для реального использования. Можно ввести ограничение, в соответствии с которым значение k_2 не может быть нулевым, и определить точку, после которой выполняется условие $k_2 < 0{,}5$. Для такой точки получено значение $k_1 = 6\%$, являющееся достаточно высоким, но возможным для практического применения. Однако при этом зафиксировано максимальное значение $k_1 = 20\%$.

Для точки, после которой $k_1 < 0{,}1$, получено значение $k_2 = 7\%$. С уменьшением k_1 (при увеличении порога) до 0% значение k_2 сильно возрастает (для $k_1 = 0{,}05\%$ значение $k_2 = 80\%$). Очевидно, нецелесообразно уменьшать значение k_1 ниже 1%, что является вполне приемлемым с учетом возможности влияния естественных помех на результат аутентификации, вызванных, например, возникновением больших пауз при наборе, если пользователь находится в нехарактерном для него эмоциональном состоянии.

Последний проведенный в процессе эксперимента тест был посвящен проверке работоспособности исследуемых методов при изменении условий аутентификации.

В частности, были рассмотрены различные для всех 15 пользователей тестовые последовательности «имя пользователя – парольная фраза» и различное число сеансов обучения. При этом получено значение коэффициента $k_3 = 3{,}2$ с максимальными значениями k_2 и k_1 37% и 18% соответственно. Для этого теста использовался комбинированный метод с тремя классификаторами и оператором произведения оценок. Результаты показали увеличение коэффициента k_3 лишь на 1,2% по сравнению с результатами тестирования для фиксированных тестовых последовательностей. Это вполне объяснимо, так как отдельные пользователи участвовали в малом числе сеансов на этапе обучения и не успели сформировать устойчивый профиль.

Выводы

В работе была рассмотрена система аутентификации, основанная на анализе клавиатурного почерка. Приведены результаты тестирования 15 пользователей в течение 6 месяцев. Система предусматривает возможность реализации трех базовых методов: метод, основанный на анализе математического ожидания и дисперсии интервалов времени между нажатиями клавиш; метод, основанный на классификации временных характеристик, и метод, основанный на анализе изменения ритма клавиатурного набора. Результаты тестирования этих методов при их изолированном использовании соответствуют среднему значению коэффициента ошибок k_3, равному 3,7%. Затем был рассмотрен комбинированный метод, основанный на совместном использовании трех базовых методов с помощью различных операторов объединения. Анализ показывает, что наилучшие качественные характеристики дает оператор суммирования оценок, но он требует наличия информации, которая не всегда может быть получена априори. В соответствии с проведенными тестами применение этого оператора в комбинированном методе позволило получить среднее значение коэффициента ошибок k_3, равное 1,8% при одновременном снижении максимальных значений k_3 для тестируемых пользователей. Это позволяет рассчитывать на перспективы широкого практического

применения динамической аутентификации пользователей по компьютерному почерку. Многочисленные улучшения качества такой аутентификации связаны, прежде всего, с комбинированным использованием традиционных методов. Отметим, что в процессе эксперимента для некоторых пользователей применение комбинированного подхода не вносило положительных изменений по сравнению с применением локального метода. В связи с этим представляется целесообразным разработать гибкую систему назначения весовых коэффициентов для отдельных составляющих комбинированного метода с учетом особенностей пользователей.

В целом, проведенные исследования свидетельствуют о возможности использования анализа динамики клавиатурного почерка для создания простых и недорогих схем аутентификации пользователей компьютеров и компьютерных систем, в частности, и по традиционной исходной фразе «имя пользователя - пароль».

Список литературы

1. Чалая Л.Э. Модель идентификации пользователей по клавиатурному почерку // Искусственный интеллект. – 2004. – № 4. – С. 811-817.

2. Leggett J., Williams G. Verifying identity via keystroke characteristics // International Journal of Man-Machine Studies. – 1988. – V. 28, n. 1. – P. 67-76.

3. Leggett J., Williams G., Usnick M., Longnecker M. Dynamic identity verification via keystroke characteristics // International Journal of Man-Machine Studies. – 1991. – Vol. 35, n. 6. – P. 859-870.

4. Coltell O., Badia J. M., Torres G. Biometric Identification System Based in Keyboard Filtering // Proc. of XXXIII Annual IEEE International Carnahan Conference on Security Technology. – 1999. – P. 203-209.

5. Kittler J., Hatef M., Duin R.P.W., Matas J. On Combining Classifiers // IEEE Trans. Pattern Anal. Mach.Intell., IEEE Computer Society. – 1998. – Vol. 20. – P. 226-239.

Поступила в редколлегию 10.12.2007

Рецензент: д-р техн. наук, проф. С.Г. Удовенко, Харьковский национальный университет радиоэлектроники, Харьков.