doi:10.11959/j.issn.1000-436x.2017272

基于顽健线性判别分析的击键特征识别方法

沈伟国[1,2]，王巍[1,2]

（1. 通信信息控制和安全技术重点实验室，浙江 嘉兴 314033;
2. 中国电子科技集团公司第三十六研究所，浙江 嘉兴 314033）

摘 要：研究了用户认证过程中的键盘击键序列特征提取和分类问题，提出一种基于顽健线性判别分析的击键特征识别方法。首先，最大化击键序列集不同类间的离散度，同时最小化序列集同类之间的离散度，保持击键序列样本的最佳判别特性。其次，最小化近邻击键序列样本间的相似性离散度，保持序列样本的区域相似性。最后，基于上述原则，对击键序列特征样本进行特征提取，并采用最近邻分类准则进行判决输出。通过与其他方法的实验对比，验证了该方法的有效性。

关键词：身份认证；击键序列识别；特征提取；顽健线性判别分析
中图分类号：TP391.4 **文献标识码**：A

Keystroke features recognition based on stable linear discriminant analysis

SHEN Wei-guo[1,2], WANG Wei[1,2]

(1. Science and Technology on Communication Information Security Control Laboratory, Jiaxing 314033, China;
2. No.36 Research Institute of CETC, Jiaxing 314033, China)

Abstract: A novel keystroke features recognition method based on stable linear discriminant Analysis (SLDA) was put forward. First of all, it maximum the dispersion between different sequences, while minimizing the dispersion between the same sequence set, maintain the best discriminant characteristics of the keystroke sequences. Secondly, the local similarity graph between keystroke sequences is constructed, minimizing the dispersion of the local similarity sequences, to keep the local similarity of keystroke sequences. Finally, based on the principles above, the feature of keystroke sequences are extracted, and the nearest neighbor classification criterion is used to judge the outputs. The effectiveness of the proposed method is certified by experiment results.

Key words: identity authentication, keystroke recognition, feature extraction, stable linear discriminant analysis

1 引言

微信、支付宝等互联网应用的蓬勃发展在改变人们生活方式的同时，随之而来的安全性问题也显得越来越严重。由于大多数网络用户安全意识薄弱以及防护机制的漏洞，导致各种身份泄露、网络资产被盗等事件频发。传统安全手段中，大多数互联网应用采用的是用户名和密码的身份认证方式，但是各种木马、病毒的出现导致密码容易被窃取或泄露。同时由于近年来人工智能和深度学习的发展，基于个人生物学特征或行为特征的用户认证方法成为研究热点[1~6]，而且很多研究成果已经被广泛应用于现实生活中，如人脸、指纹等。但是这些特征需要辅助的硬件设备进行数据采集，而击键行为特征数据通过键盘就能收集，不需要添加辅助设备，是一种省时省力的认证方法。

近年来，基于统计模式识别的击键特征识别方法引起了国内外研究者的广泛关注，取得了不错的研究成果[7~15]，但是还没有提出公认的十分有效的通用算法。Charles 等[7]最先提出了基于贝叶斯学习的击键特征识别方法，但实验结果给出的漏报率和误报率还较高。冯力等[8]对 Charles 的贝叶斯学习方

收稿日期：2017-11-09

法进行了完善，提出了一种基于贝叶斯学习的加权统计方法。Monrose 等[9]提出了基于 K 近邻聚类的识别算法，取得了较好的效果。史扬等[10]对用户的成功登入样本进行学习并提取平均特征向量。通过比较输入样本的特征向量和平均特征向量来进行认证判决。梁娟等[11]通过计算当前用户击键序列特征与共性特征的欧氏距离来对用户进行认证，给出了一种基于特征子空间的判决方法，该方法误报率较低，顽健性较好。本文对传统模式识别算法应用于击键序列识别进行了探索，在此基础上，提出了基于顽健线性判别分析的击键序列识别方法，通过采集样本进行实验测试，证明了该方法的有效性。

2　击键序列特征提取

目前，针对击键特征提取的研究主要是从击键时间特征入手，因为不同用户的知识水平和操作熟练程度的差异，在击键行为上反映出不同的节奏特性。而同一用户因为行为习惯，输入同一口令或击打同一按键时在时间特征上具有平稳随机性特征，如图 1 所示。

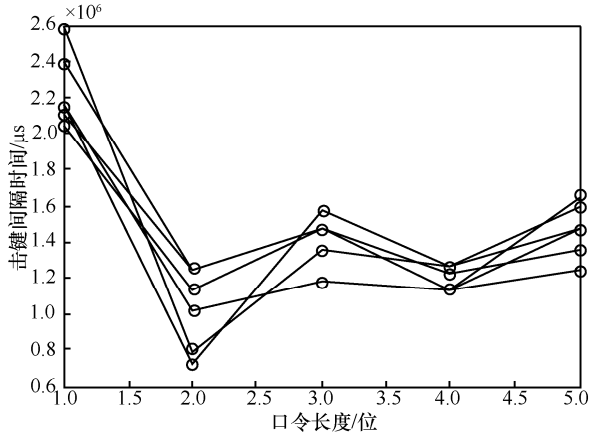


图 1 同一用户输入同一口令时的时间特征

本文击键特征的提取参照文献[3]提出的方法，通过采集用户敲击键盘的时间序列并进行预处理得到每次敲击键盘压下的时间和键盘弹起的时间。对用户击键序列的特征进行归纳提取，主要包括 2 个方面:单次击键的持续时间和相邻 2 次击键的间隔时间。如果单个按键压下和弹起的分别时间为 P_i 和 R_i，那么通过以下公式就可以计算出相应的按键持续时间和间隔时间。

持续时间：单次按键弹起的时间减去压下的时间，计算式为

$$D_i = R_i - P_i$$

间隔时间：单次按键压下的时间减去前次按键弹起的时间，计算式为

$$F_i = P_i - R_{i-1}$$

将击键持续时间和间隔时间组成的特征向量作为表征击键序列的原始特征样本，假设输入的击键长度为 n，单个击键序列样本可用 $D_1, F_1, \cdots, D_n, F_{n-1}$ 来表示。单个击键序列样本总共由 $2n-1$ 维击键特征向量组成，其中 n 个用于表示单次按键的持续时间，$n-1$ 个用于表示相邻 2 次击键的时间间隔。

3　顽健线性判别分析与击键特征识别

为了去除击键特征之间相关性、降低特征维数的同时，保持分类样本的判别信息，需要寻找一个最佳特征子空间。顽健线性判别分析（SLDA, stable linear discriminant analysis）的基本原则是寻找最优投影向量，通过最大化击键序列样本集不同类间的离散度，同时最小化序列集同类之间的离散度，保持击键序列样本的最佳判别特性。同时，相似邻接图的构造使区域最近的击键序列样本映射到最佳特征空间后仍然距离相近，从而保持击键序列样本空间的区域相似性信息。

假设击键序列训练样本集为 $\{(x_i, \tau_i)\}_{i=1}^{N}$，$N$ 为样本个数，$x_i \in R^{d\times 1}(i=1,2,\cdots,N)$ 表示第 i 个样本，每个样本由 d 个原始特征参量表示，τ_i 为训练样本中 x_i 的类别标签。假设击键序列样本集中第 k 类的个数为 N_k，则 $\sum_{k=1}^{c} N_k = N$。根据 SLDA 的算法思想，建立如下目标函数

$$\max_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{\alpha} + \beta \sum_{i,j=1}^{n} \left(\boldsymbol{a}^{\mathrm{T}} \boldsymbol{x}_i - \boldsymbol{a}^{\mathrm{T}} \boldsymbol{x}_j\right)^2 \boldsymbol{S}_{ij}} \tag{1}$$

$$\boldsymbol{S}_b = \frac{1}{N} \sum_{k=1}^{c} N_k \left(\boldsymbol{m}^{(k)} - \boldsymbol{m}\right)\left(\boldsymbol{m}^{(k)} - \boldsymbol{m}\right)^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{S}_w = \frac{1}{N} \sum_{k=1}^{c} \left(\sum_{i=1}^{N_k} \left(\boldsymbol{x}_i^{(k)} - \boldsymbol{m}^{(k)}\right)\left(\boldsymbol{x}_i^{(k)} - \boldsymbol{m}^{(k)}\right)^{\mathrm{T}} \right) \tag{3}$$

其中，m 表示全体击键序列训练样本的均值，m^k 表示第 k 个类的击键序列样本均值。$\boldsymbol{\alpha}$ 为投影向量，$\boldsymbol{S}_w$ 和 $\boldsymbol{S}_b$ 分布表示击键序列样本的同类之间的离散度矩阵和不同类间的离散度矩阵。$\boldsymbol{x}_i^{(k)}$ 为第 k 类中的第 i 个击键样本。β 表示调节系数，对分母前后

2 个部分进行权衡。S_{ij} 为邻接图的权重矩阵中的元素，定义如下

$$S_{ij}=\begin{cases}1,\ x_i\in N_k(x_j)\text{ 或 }x_j\in N_k(x_i)\\0,\text{其他}\end{cases}\tag{4}$$

目标函数可分解成如下 3 部分。

$$\max\sum_{k=1}^{C}N_k\boldsymbol{\alpha}^{\mathrm{T}}\left(\boldsymbol{m}^{(k)}-\boldsymbol{m}\right)\left(\boldsymbol{m}^{(k)}-\boldsymbol{m}\right)^{\mathrm{T}}\boldsymbol{\alpha}\tag{5}$$

$$\min\frac{1}{n}\sum_{i=1}^{C}\left(\sum_{j=1}^{N_i}\boldsymbol{\alpha}^{\mathrm{T}}\left(\boldsymbol{x}_j^{(i)}-\boldsymbol{m}^{(i)}\right)\left(\boldsymbol{x}_j^{(i)}-\boldsymbol{m}^{(i)}\right)^{\mathrm{T}}\boldsymbol{\alpha}\right)\tag{6}$$

$$\min\sum_{i,j=1}^{N}\left(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_j\right)^2S_{ij}\tag{7}$$

式(5)的作用是使不同类击键序列样本的离散度最大化，即让不同类的击键序列样本映射之后距离越远越好，式(6)的主要作用是使同类击键序列样本之间的离散度最小化，即让同类击键序列样本经过映射之后保持距离相近，从而对击键序列样本空间的几何结构信息进行刻画。式(7)的主要作用是使区域最近的击键序列样本映射到最佳特征空间后仍然距离相近，从而保持击键序列样本空间的区域相似性信息。

将式(7)化简可得

$$\begin{aligned}&\sum_{ij}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_j)^2S_{ij}\\&=2\sum_i\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_iD_{ii}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha}-2\sum_{ij}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_iS_{ij}\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\alpha}\\&=\ 2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{S}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}\\&=\ 2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{D}-\boldsymbol{S})\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}\\&=\ 2\,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}\end{aligned}\tag{8}$$

其中，$\boldsymbol{X}=[x_1,x_2,\cdots x_N]$，矩阵 $\boldsymbol{S}$ 中的元素表示 2 个击键序列样本间距离权重值，对角矩阵 $\boldsymbol{D}$ 的元素值等于 $\boldsymbol{S}$ 中各列或者各行元素值的和，即 $D_{ii}=\sum_j S_{ij}$。

$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S}$ 称为拉普拉斯矩阵。

将式(8)代入式(1)后，目标函数变为

$$\max_{\boldsymbol{\alpha}}\frac{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{S}_t+\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}})\boldsymbol{\alpha}}\tag{9}$$

其中，$\boldsymbol{S}_t=\boldsymbol{S}_b+\boldsymbol{S}_w$，通过化简可得，以上目标的函数的解即式(10)特征方程所求得的特征向量。

$$\boldsymbol{S}_b\boldsymbol{\alpha}=\lambda(\boldsymbol{S}_t+\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}})\boldsymbol{\alpha}\tag{10}$$

为了能够充分地保留击键序列样本特征，通常选取的特征向量个数 $l(l\geqslant 2)$，假设 $\boldsymbol{\alpha}=[\alpha_1\,\alpha_2\cdots\alpha_l]$ 为最佳映射矩阵，那么 $\alpha_1,\alpha_2,\cdots,\alpha_l$ 即是的前 l 个最大特征值所对应的特征向量。

任意击键序列样本 $\boldsymbol{x}^*$ 在映射空间中的向量表示 $\boldsymbol{y}^*$ 可以由公式 $\boldsymbol{y}=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}$ 计算得到。经过特征提取和去相关性后，采用最近邻分类准则进行判决输出。即如果 $d\left(\boldsymbol{y}_k,\boldsymbol{y}^*\right)=\min_i\left\{d\left(\boldsymbol{y}_i,\boldsymbol{y}^*\right)\right\}$，则测试击键序列样本 $\boldsymbol{x}^*$ 与已知击键序列 $\boldsymbol{x}_k$ 属于同一类。其中，$d\left(\boldsymbol{y}^*,\boldsymbol{y}_i\right)=\left\|\boldsymbol{y}^*-\boldsymbol{y}_i\right\|^2$。

4 实验和性能分析

实验通过 VC++生成击键序列采集客户端，设定长度 n 为 5~10 的 6 组不同口令，按第 2 节所述方法，每组口令分别采集 10 个人的击键序列样本。实验中，每人每组口令采集 40 次，随机选取其中的 25 次作为训练样本，其余 15 次的作为测试样本。所以，每组口令总共产生训练样本 250 个，测试样本 150 个。

为了验证本文算法的有效性，实验在口令长度为 5 的情况下，选取了其中 4 个人的 100 个训练样本，经过训练产生最佳投影矩阵。为了对击键序列训练样本映射到特征子空间后的分布情况进行分析，本文选取了 2 组特征向量来绘制二维分布图。从图 2 中可以发现，经过投影之后，除了个别样本存在交错重叠，大部分都能得到很好的区分。

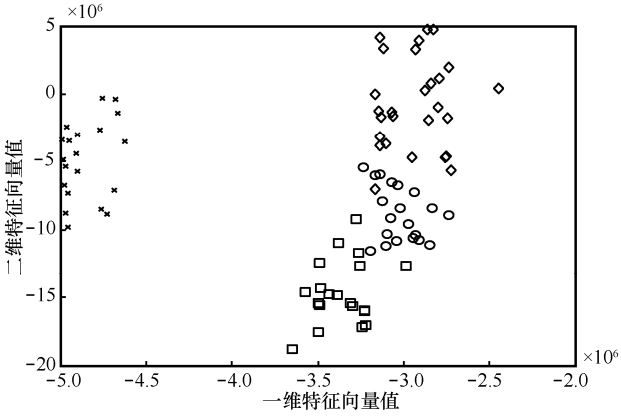


图 2 投影后样本分布情况

同时，实验在不同口令长度下测试了本文算法的性能，并与基于欧氏距离、PCA 和 LDA 的识别方法进行了对比。PCA 和 LDA 都是基于特征子空间的识别方法，其主要思想都是寻找一个最优投影方向，使投影到子空间之后数据样本的分类特性更好。图 3 表示在口令长度从 5~10 变化时，4 种方法的识别率对比

情况。从图 3 中可以发现，本文算法的稳定性更好、识别率更高。图 4 表示当口令长度为 10 时，在选取不同特征向量个数情况下，4 种算法的识别率对比情况。从中可以发现，随着特征向量个数的增加，识别率有了较大的提高，本文算法整体性能更好。

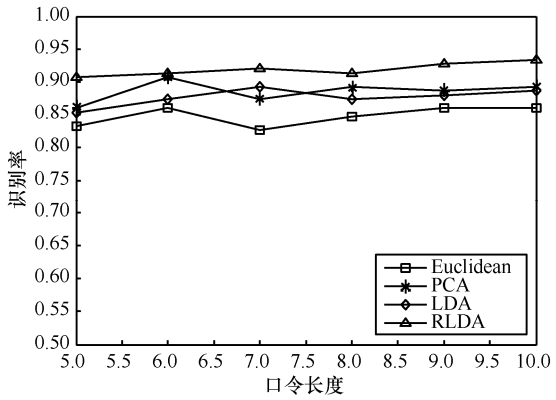


图 3　不同口令长度时的识别率对比

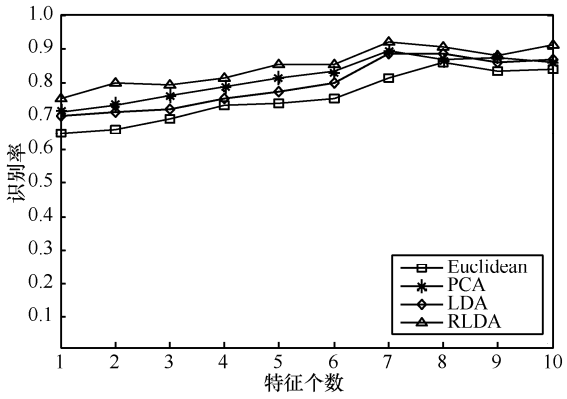


图 4　特征向量个数变化时识别率对比

5　结束语

传统的用户安全机制亟待改进，随着基于个人生物学特征或行为特征的用户认证方法的兴起，相关的研究成果已经被广泛应用。基于击键特征的用户认证由于其成本低、效率高可以很好地弥补传统认证方式中存在的缺陷。本文从寻找最佳特征子空间入手，提出了一种基于顽健线性判别分析的击键特征识别方法。通过与其他识别方法的对比验证了该方法的有效性。在今后的工作中，将对算法进行改进，同时研究在大数据情况下，如何保持算法的稳定性和顽健性。

参考文献：

[1] GAINES R, LISOWSKI W, PRESS S. Authentication by keystroke timing: some preliminary results[R]. Rand Corporation: Rand Report R-2560-NSF, 1980.

[2] JOYCE R, GUPTA G. Identity authentication based on keystroke latencies[J]. Commun ACM, 1990, 33(2): 168-176.

[3] ROBINSON J, LIANG V. Computer user verification login string keystroke dynamics[J]. IEEE Trans Syst Man Cybern, 1998, 28(2): 236-241.

[4] LEGGETT J, WILLIAMS G, USNICK J. Dynamic identity verification via keystroke characteristics[J]. International Journal of Man-machine Studies, 1991, (35): 859-870.

[5] NAPIER R, LABERTY W, MAHAR W. Keyboard user verification: toward an accurate, efficient, andecologically valid algorithm[J]. International Journal of Human Computer Studies, 1995(43): 213-222.

[6] DUNN A. biometric authentication-real identities for a virtual world, MIDAS[R]. Project Development Final Report, 2002.

[7] SALEH B, CHARLES S, BASSAM E. Computer-access security systems using keystroke dynamics[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1990, 12(12): 1217-1222.

[8] 高艳, 管晓宏, 冯力. 基于实时击键序列的主机入侵检测[J]. 计算机学报, 2004, 27(3): 396-401.
GAO Y, GUAN X H, FENG L. The host-based intrusion detection based on real time keystroke sequences[J]. Chinese Journal of Computers, 2004, 27(3): 396-401.

[9] MONROSE F, RUBIN A. Authentication via keystroke dynamics[C]//Fourth ACM Conference on Computer and Communications Security, 19973.

[10] SHI Y, CAO L M. User identity verification based on recognition of typing style[J]. Computer Engineering, 2005, 31(6).

[11] LIANG J, WANG X, CHEN W W. Recognition of user's keystroke features based on difference subspace[J]. Computer Engineering. 2007, 33(11).

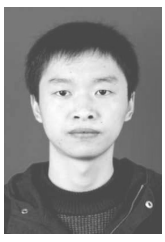
[12] WANG Y J, ZHAO P H, WANG M M. Method of user identification based on keystroke behavior and its application[J]. Computer Science. 2015, 42(11).

[13] 王振辉, 王振铎, 支佩买. 基于鼠标和键盘行为特征组合的用户身份认证[J]. 计算机应用与软件, 2016, 33(7): 308-312.
WANG Z H, WANG Z D, ZHI K M. User identity authentication based on mouse and keyboard behavioural biometrics combination[J]. Computer Applications and Software, 2016, 33(7):308-312.

[14] 王晅, 陈伟伟, 马建峰. 基于遗传算法和灰色关联分析的击键特征识别算法[J]. 计算机应用, 2007, 27(5): 1054-1057.
WANG X, CHEN W W, MA J F. User authentication algorithm with keystroke features based on genetic algorithms and grey relational analysis[J].Journal of Computer Applications,2007,27(5):1054-1057.

[15] 蒋李芬, 刘家芬. 基于键盘行为数据的用户身份识别[J]. 计算机应用, 2015, 35(SI): 110-112.
JIANG L F, LIU J F. User authentication based on keystroke dynamics[J]. Journal of Computer Applications, 2015, 35(SI): 110-112.

作者简介：



沈伟国（1987-），男，浙江湖州人，中国电子科技集团公司第三十六研究所工程师，主要研究方向为网络安全。

王巍（1980-），男，河北张家口人，博士，中国电子科技集团公司第三十六研究所高级工程师，主要研究方向为网络安全。