Journal of Computer Applications
计算机应用,2015,35(S1):110 - 112,129
ISSN 1001-9081
CODEN JYIIDU
2015-06-20
http://www.joca.cn

文章编号:1001-9081(2015)S1-110-03

基于键盘行为数据的用户身份识别

蒋李灵*,刘家芬
(西南财经大学 经济信息工程学院,成都 611130)
(*通信作者电子邮箱 jiangllswufe@gmail.com)

摘　要:用户击键行为作为一种生物特征,具有采集成本低、安全性高的特点。然而,现有的研究方法和实验环境都是基于实验室数据,并不适用于极度不平衡的真实数据。比如,在实验室数据上效果出色的分类算法在真实数据上却无法应用。针对此问题,提出了基于真实击键行为数据的用户识别算法。该方法将聚类算法和距离算法结合起来,通过比较新来的击键行为和历史击键行为相似度以实现用户识别。实验结果表明,该算法在100名用户的3015条真实击键记录组成的数据集上准确率达到88.22%,在投入实际应用后,随着样本集的增大算法的准确率还可以进一步提升。

关键词:键盘行为;用户识别;欧氏距离;*k*-means聚类;生物认证
中图分类号:TP309.2　　**文献标志码**:A

User authentication based on keystroke dynamics

JIANG Liling*, LIU Jiafen
(*College of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu Sichuan* 611130, *China*)

Abstract: Keystroke dynamics, part of biometrics, is featured as low cost and high security. However, existing researches and experiments are mostly based on laboratory data, which are not appropriate for extremely unbalanced real data. For example, classification algorithms are not applicable to real data due to the extreme imbalance of normal and abnormal samples. In order to solve this problem, a keystroke dynamics method was proposed for real data, which combined classification algorithm with distance algorithm. It authenticated users by comparing user's new behavior with historical behavior. Experimental results show that this method has an accuracy rate of 88.21% on the real dataset of 3015 records of 100 users. It is expectable that the performance will be better with the expansion of data set.

Key words: keystroke dynamics; user authentication; Euclidean distance; *k*-means clustering; biometric authentication

0　引言

随着互联网络的发展,电子商务已经成为人们进行商务活动的新模式。在移动互联和移动智能终端的进一步推动下,互联网交易额年年攀升。如何提高互联网交易的安全性,保护用户的账号安全,成为了学术界的研究热点,也是工业界急需解决的问题。而用户身份认证则是保证账号安全的第一道防线。

传统用户身份认证方法主要是口令认证,随着用户对安全性要求的提升,出现了动态口令、智能卡认证等,但这些都属于用户外在的物理认证,存在遗失或被盗用的潜在风险。采用生物特征进行用户身份认证与传统认证方法相比,具有更强的安全性。目前被采用的生物特征主要被分为两类:一类为生物生理特征,包括指纹、虹膜等;一类为生物行为特征,包括用户击键行为等。但如果通过生物生理特征进行验证,成本要求较高,需要增加新的硬件设备,如指纹采集设备等;相对而言,生物行为特征的应用成本低且高效,通过现有设备则可完成,如击键行为数据通过键盘就能收集,无需添加新的硬件,利于系统的实施和推广[1-2]。

在学术界最早研究击键行为是贝尔实验室(Bell-Labs)的Monrose等[3-4],他们通过分析用户击键行为,发现其具有一定的规律性,且随时间的推移用户的行为习惯会有缓慢的变化。随后研究击键行为的学者越来越多,如CMU大学的Kevin等[5]提出基于距离的识别方法,以及后来学者提出的基于机器学习算法的识别方法[6]。但这些研究大都在实验室中进行,分析数据也都来自于设计的实验采样,与现实应用中的数据具有较大的差异性。真实数据往往缺乏负样本,数据极度不平衡,导致上述分类算法在真实数据上难以实施。因此本文基于某互联网公司真实的用户击键行为数据,提出了面向真实数据的键盘行为识别方法,并通过实验验证了该方法的实用性和可推广性。

1　击键行为特征

1.1　击键特征提取

击键特征提取参照Douhou[7]和Zhong等[8]提出的方法。通过收集用户击键事件的时间计算得到,如按下(press)时间、弹起(release)时间。根据以前学者研究得出的结论,用户击键特征主要体现在以下两个方面:击键的持续时间(Dwell Time)和击键的间隔时间(Flight Time)[7-9]。记每个键的按下和弹起时间为(P_i, R_i),从而可以计算出相应的持续时间和间隔时间,公式如下。

持续时间(Dwell Time):一次击键的弹起时间减去按下

收稿日期:2014-07-23;**修回日期**:2014-09-05。　**基金项目**:国家自然科学基金重大研究计划项目(91218301);国家自然科学基金青年项目(60903201);中央高校基本科研业务专项项目(JBK140129)。

作者简介:蒋李灵(1989-),男,四川内江人,硕士研究生,CCF会员,主要研究方向:数据挖掘、商务智能;　刘家芬(1980-),女,湖北荆州人,副教授,博士,CCF高级会员,主要研究方向:信息安全、智能信息处理。

时间。

$D_i = R_i - P_i$

间隔时间(Flight Time):一次击键的按下时间减去上次击键的弹起时间。

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

如:一个用户输入 4 个字符的击键行为特征数据为:

$T = \langle D_1, F_1, D_2, F_2, D_3, F_3, D_4, F_4, D_5 \rangle$

其中:D_1 表示第一个字符的弹起时间和按下时间的间隔;F_1 表示第二个字符的按下时间和第一个字符的弹起时间间隔;D_2, F_2, D_3, F_3, D_4, F_4, D_5 类似。

1.2　真实数据和实验室数据对比

虽然实验室数据和真实数据特征值的提取方法相同,但在数据构成上有着很大的差别。实验室数据是通过特定的击键行为实验收集而来,能够满足许多特定的条件。其采集的是多人输入相同一段字符串的击键行为数据,然后通过模型区分不同人输入相同字符串的行为习惯。对于实验室数据,可以通过加大实验人员输入密码的次数,增加数据样本量;也可以让不同实验人员输入相同的密码,从而使相同的密码字符拥有不同击键行为,使正样本和负样本维持平衡。所以在实验室数据上,可以利用分类算法,例如神经网络[10-11]、支撑向量机[6,12]等实现对同一密码不同用户击键行为的区分;同样也可以通过距离的方法计算不同样本间的相似度[13],使用 FAR(错误率,False Acceptance Rate)和 FRR(误拒绝率,False Rejection Rate)画出 ROC(Receiver Operating Characteristic)曲线,选择合适阈值,区分同一密码不同用户之间击键行为。

然而对于真实数据,每个人输入的字符串是不一样的,即每个人设置的密码或者账号是不一样的。同一密码或者账号所收集的击键行为数据十分不平衡,在真实环境中收集到的大部分数据都为正常行为,案件或者盗用行为极少,这也十分符合真实的环境。在现实生活中极其少的情况才会发生账号被盗,所以对于真实数据中一个用户的击键数据往往只包含正常的行为数据。因此实验室中可以应用的分类算法,ROC 曲线寻找阈值思想,在真实环境下是行不通的。

通过对真实数据进行统计分析,发现用户输入自己密码或者账号具有很强的习惯性和相似性,与案件或者盗用行为区分度很大。以某一用户击键行为的间隔时间(Flight Time)为例画折线图,即该用户从键盘输入 16 个字符,则间隔时间有 $\langle F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}, F_{13}, F_{14}, F_{15} \rangle$ 共 15 个数,共 4 次击键行为数据(3 次正常、1 次案件),如图 1。

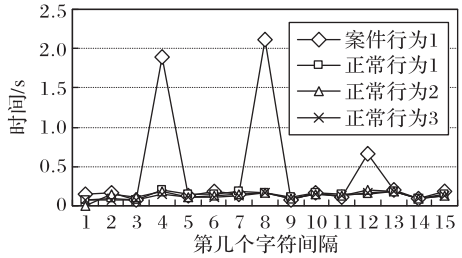

图 1　某用户账户或密码击键行为数据

由图 1 可以看出案件行为和 3 次正常行为有明显的差异性。

因此本文基于用户击键行为的真实数据的特征,针对实验室算法存在的问题,提出了基于距离算法的模型及其优化算法。

2　击键行为算法

2.1　距离算法

如何判定一个用户是否为合法用户,关键在于计算新来样本与原有样本之间的相似度,如果相似度高于某一个阈值,则有理由认为是同一个用户的击键行为。由上文可知用户一次击键行为可以用一个向量表示(即向量 $\boldsymbol{T}$),该向量由用户击键的持续时间(Dwell Time)和间隔时间(Flight Time)组成。计算不同向量之间相似度最简单、直接、高效的方法则是距离方法,通过距离的长短来度量其相似度。目前工业界和学术界使用最多的距离方法有曼哈顿距离、马氏距离、欧氏距离等。

设有如下两向量 $\boldsymbol{A}(x_1, y_1)$,$\boldsymbol{B}(x_2, y_2)$,曼哈顿距离、欧氏距离、马氏距离的计算方法可以分别表示如下。

曼哈顿距离(Manhattan Distance):

$D = |x_1 - x_2| + |y_1 - y_2|$

欧氏距离(Euclidean Distance):

$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

马氏距离(Mahalanobis Distance):

$D = \sqrt{(\boldsymbol{A} - \boldsymbol{B})^{\mathrm{T}} \boldsymbol{S}^{-1} (\boldsymbol{A} - \boldsymbol{B})}$

其中 $\boldsymbol{S}$ 为协方差矩阵。

为了对比不同距离表示方法在真实数据上的效果,分别用曼哈顿距离、马氏距离、欧氏距离进行实验。实验思路是采用 3 种距离分别对每个用户的账号或密码进行建模,判断模型的准确率,具体算法如下。

算法一

输入:一个账号或密码的击键行为($T_1, T_2, T_3, \cdots, T_n$),前 k 个作为训练数据,后 $n - k$ 个作为测试数据。

步骤:

1) 对 $T_1, T_2, \cdots, T_k$,共 k 条行为记录两两之间计算距离,得到距离数组 $d_1, d_2, \cdots, d_m$,其中 $m = C_K^2$。

2) 对距离数组排序,选择 80% 分位数作为阈值。

3) 测试数据 $T_{k+1}, T_{k+2}, \cdots, T_n$,分别与历史数据计算距离,如果计算出的距离有 80% 都小于阈值,则认为该条记录为该用户的正常击键行为。

输出:阈值,预测为同一用户击键行为个数,预测为不同用户击键行为个数。

实验结果如表 1 所示。

表 1　不同距离的准确率　%

样本数	曼哈顿距离	马氏距离	欧氏距离
20	80.22	82.23	81.08
50	79.91	82.79	82.54
100	80.53	82.70	82.02

由表 1 可以看出,曼哈顿距离的效果比另外两个算法稍差,马氏距离和欧氏距离效果较好。由马氏距离的公式可知,计算马氏聚类需要计算协方差矩阵的逆矩阵。但在真实数据中,很多密码或者账号所收集到的击键行为数据很有限,导致大部分模型无法计算协方差矩阵的逆矩阵,使马氏距离的应用范围受到了很大的限制。所以在本文将采用效果差异不大的欧氏距离代替马氏距离,并采用聚类方法对算法一进行改进,从而提高模型的可用性和鲁棒性。

2.2　算法优化

在算法一判定一次新的击键行为是否由该用户产生时,

需要将这次击键数据与历史中所有数据进行比对并计算距离,然后才能得出结果,这就使得算法的执行效率非常低。

因此本文考虑对算法一进行改进,利用 k- means 聚类算法对历史数据进行聚类,用一个或者多个新的中心点代替历史数据,当一次新的击键行为数据进入模型,就只需与极少量的中心点计算距离,改进算法如下所示。

算法二

输入:一个账号或者密码的击键行为$(T_1,T_2,\cdots,T_n)$,前 k 个作为训练数据,后 $n-k$ 个作为测试数据,历史数据聚为 i 类。

步骤:

1)对 $T_1,T_2,\cdots,T_k$,共 k 条行为记录两两之间计算欧氏距离,得距离数组 $d_1,d_2,\cdots,d_m$,其中 $m = C_K^2$。

2)对距离数组排序,选择 80% 分位数作为阈值。

3)对训练数据 $T_1,T_2,\cdots,T_k,k$ 条行为记录进行聚类,一共聚为 i 类,得聚类中心 $C_1,C_2,\cdots,C_i$ 及每一类的样本个数 $N_1,N_2,\cdots,N_i$。

4)测试数据 $T_{k+1},T_{k+2},\cdots,T_n$,计算每一个测试数据与聚类中心的距离记为 $d_1,d_2,\cdots,d_i$,最终测试数据的距离为 d,公式如下:

$$d = \frac{N_1}{k} * d_1 + \frac{N_2}{k} * d_2 + \cdots + \frac{N_i}{k} * d_i$$

5)阈值比较,如果 d 小于阈值,则分为同一用户;反之亦然。

输出:阈值,预测为同一用户的行为数,预测为不同用户的行为数。

3 实验结果及分析

本文从模型的可用性和鲁棒性出发,选择了欧氏距离作为模型的距离公式。为了提高模型的效率,预先用 k- means 聚类算法对历史数据进行聚类,对算法一改进优化,得到了算法二。根据统计抽样的方法,从真实数据中选出 3 个不同大小的样本数据集:20 个用户账号的 727 条击键行为记录组成的 DataSet1、50 个用户账号的 1 819 条击键行为记录组成的 DataSet2 和 100 个用户账号的 3 015 条击键行为记录组成的 DataSet3。本文对这三个样本集进行了训练和测试,经过多次实验发现,选择 k 值为 10 的时候能达到比较好的聚类效果。三个样本集在算法优化前后的运行时间对比见图 2,其中纵轴为在样本集上进行一次用户身份识别时所花的时间,单位为秒。

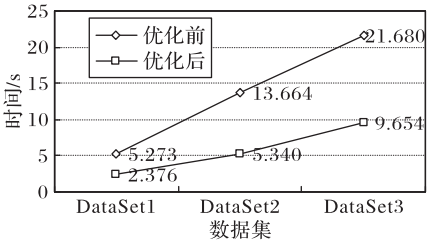

图 2 算法优化前后运行时间对比

由图 2 可以看出,引入聚类算法后,整体算法的运行时间有明显的减少。这意味着该算法在实际应用中,能提高运行效率,提升用户体验。

在效率提高的同时,算法的准确性变化见表 2。从表 2 可以看出在提升了效率的同时,算法的准确率也有着较为明显的提升,所以用几个聚类中心代替整体数据在理论和应用上是可行的。

表 2 准确率对比 %

情况	20 个样本	50 个样本	100 个样本
优化前	81.08	82.54	82.02
优化后	81.41	88.13	88.21

随着样本集的增大,算法准确性有着明显的提高。可以预见的是,在投入实际应用后随着样本集的增大算法的准确率还可以进一步提升。综合算法二在从算法执行效率和准确率上的表现,本文提出的算法有较强的应用价值。

4 结语

本文针对常用分类算法无法应用于真实击键数据的问题,面向真实数据提出了一种基于击键行为的用户身份识别方法,该方法通过距离方法和聚类方法的结合,使识别算法在执行效率和准确率上都达到了令人满意的效果,已达到可实际应用的条件。今后将在以下问题上进一步深入研究:1)如何对机器学习算法如神经网络和支撑向量机等算法进行改进,使其可以适用于真实数据,以进一步提高击键行为识别的精度和效率;2)通过对真实数据的处理,解决真实数据正常行为和非正常行为样本不平衡的问题。

参考文献:

[1] 张治元, 田国忠. 基于击键韵律的身份认证模型设计与实现[J]. 计算机应用, 2009, 29(10): 2799 - 2801.

[2] 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.

[3] MONROSE F, REITER M, WETZEL S. Password hardening based on keystroke dynamics[C]// Proceedings of the 6th ACM Conference on Computer and Communications Security. New York: ACM, 1999: 73 - 82.

[4] MONROSE F, RUBIN A D. Keystroke dynamics as a biometric for authentication[J]. Future Generation Computer Systems, 2000, 16: 351 - 359.

[5] KILLOURHY K S, MAXION R A. Comparing anomaly-detection algorithms for keystroke dynamics[C]// DSN '09: Proceedings of the 2009 IEEE/IFIP International Conference on Dependable Systems & Networks. Piscataway: IEEE, 2009: 125 - 134.

[6] GIOT R, EI-ABED M, ROSENBERGER C. Keystroke dynamics with low constraints SVM based passphrase enrollment[C]// Proceedings of the IEEE 3rd International Conference on Biometrics: Theory, Applications and Systems. Washington DC, IEEE Computer Society, 2009: 425 - 430.

[7] DOUHOU S, MAGNUS J R. The reliability of user authentication through keystroke dynamics[J]. Statistica Neerlandica, 2009, 63(4): 432 - 449.

[8] ZHONG Y, DENG Y, JAIN A K. Keystroke dynamics for user authentication[C]// Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2012: 117 - 123.

[9] FRANCESCO B, GUNETTI D, PICARDI C. User authentication through keystroke dynamics[J]. ACM Transactions on Information and System Security, 2002, 5(4): 367 - 397.

[10] AHMED A A, TRAORE I. Biometric recognition based on free-text keystroke dynamics[J]. IEEE Transactions on Cybernetics, 2014, 44(4): 458 - 472.

(下转第 129 页)

最小作为判据完成识别,并输出结果[12]。表 1 是对本次实验重复进行三次统计以后的结果。从表中可以看出以下三点:1)利用 MFCC 识别率比 LPCC 高;2)MFCC 鲁棒性比 LPCC 好;3)经过主成分分析和 K-means 聚类的混合特征参数识别率比单一参数的识别率有小幅度提升。

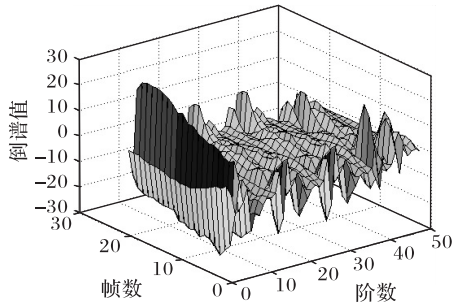

图 1　MFCC 和一阶差分 MFCC 参数

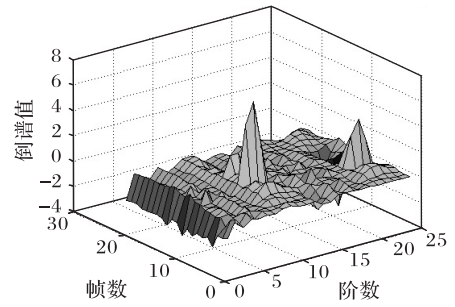

图 2　LPCC 和一阶差分 LPCC 参数

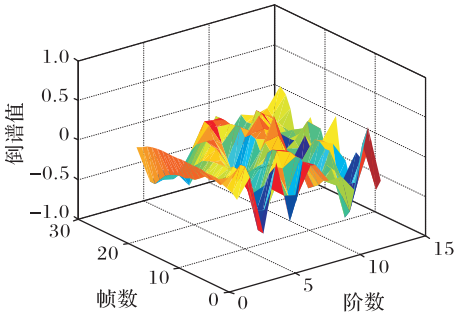

图 3　经过主成分分析后混合特征参数

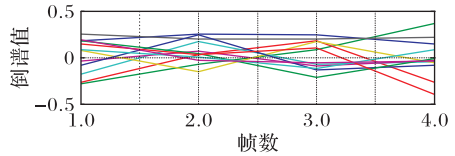

图 4　K-means 聚类(4×12)

4　结语

本文详细介绍了混合特征参数的提取算法,以及主成分分析和 K-means 聚类的计算过程。首先分别将先前提取的 LPCC 和 MFCC 以及差分参数经过主成分分析降低每一帧阶数,然后对计算结果组合后利用K-means算法降低帧数,最后以此作为识别参数进行实验。通过实验表明 MFCC 参数识别率比 LPCC 高,鲁棒性好,同时混合后参数正确识别率更高。在 K-means 聚类后,使得计算量降低,收敛速度更快,更适合硬件开发和实时测试。在本文研究的基础上,下一步重点研究 LPCC、MFCC 和其他参数如何更合理地混合,更加有效全面地表征语音信号的特征,以及针对不同年龄和不同语种的语音提取不一样的参数。

表 1　不同特征参数识别率对比

方法	识别率/%	
	测试 1	测试 2
LPCC	78.82	45.00
MFCC	83.33	66.70
LPCC + △LPCC	85.11	72.35
MFCC + △MFCC	90.78	77.78
混合参数	92.35	81.33
混合参数 K-means 聚类	92.75	88.42

参考文献:

[1] 尉洪, 周浩, 杨鉴. 基于矢量量化的组合参数法说话人识别[J]. 云南大学学报: 自然科学版, 2002, 24(2): 96 – 100.

[2] 余建潮, 张瑞林. 基于 MFCC 和 LPCC 的说话人识别[J]. 计算机工程与设计, 2009, 30(5): 1189 – 1191.

[3] ERGUN Y, VASIF V N. Comparison of MFCC, LPCC and PLP features for the determination of a speaker' s gender [C]// Proceedings of the 2014 IEEE 22nd Signal Processing and Communications Applications Conference. Piscataway: IEEE Press, 2014: 321 – 324.

[4] 于明, 袁玉倩, 董浩, 等. 一种基于 MFCC 和 LPCC 文本相关说话人识别方法[J]. 计算机应用, 2006, 26(4): 883 – 885.

[5] 刘雅琴, 智爱娟. 几种语音识别特征参数的研究[J]. 计算机技术与发展, 2009, 19(12): 67 – 70.

[6] YUAN Y J, ZHAO P H, ZHOU Q. Research of speaker recognition based on combination of LPCC and MFCC[C]// Proceedings of the 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems. Piscataway: IEEE, 2010: 765 – 767.

[7] 孙强, 叶玉堂, 邢同举, 等. 基于主成分分析法的人脸识别的探讨与研究[J]. 电子设计工程, 2011, 19(20): 101 – 104.

[8] 储雯, 李银国, 徐洋, 等. 基于段级特征主成分分析的说话人识别算法[J]. 计算机应用, 2013, 33(7): 1935 – 1937, 1968.

[9] SALDANHA J C, ANANTHAKRISHNA T, PINTO R. Vocal fold pathology assessment using Mel-frequency cepstral coefficients and linear predictive cepstral coefficients features[J]. Journal of Medical Imaging and Health Informatics, 2014, 4(2): 168 – 173.

[10] 彭湘陵, 钱盛友, 赵新民. 基于混合特征参数和 BP-Adaboost 的方言辨识[J]. 计算机工程与应用, 2013, 49(3): 152 – 155.

[11] 杨行峻, 迟惠生. 语音信号数字处理[M]. 北京: 电子工业出版社, 1996: 128 – 131.

[12] 李泽, 崔宣, 马雨廷, 等. MFCC 和 LPCC 特征参数在说话人识别中的研究[J]. 河南工程学院学报: 自然科学版, 2010, 22(2): 51 – 54.

(上接第 112 页)

[11] CHAUHAN S, PREMA K V. Effect of dimensionality reduction on performance in artificial neural network for user authentication [C]// Proceedings of the 3rd IEEE International Advance Computing Conference. Washington, DC: IEEE Computer Society, 2013: 788 – 793.

[12] BARTLOW N, CUKIC B. Evaluating the reliability of credential hardening through keystroke dynamics[C]// Proceedings of the 17th International Symposium on Software Reliability Engineering. Washington, DC: IEEE Computer Society, 2006: 117 – 126.

[13] CHANG J M, FANG C-C, HO K-H, *et al*. Capturing cognitive fingerprints from keystroke dynamics[J]. IT Professional, 2013, 15(4): 24 – 28.