An Efficient Feature Selection Technique for User Authentication using Keystroke Dynamics

D. SHANMUGAPRIYA,† and G. PADMAVATHI ††,
†Dept. of Information Technology
††Prof. and Head, Dept. of Computer science
Avinashilingam Institute for Home Science and Higher Education for women, Coimbatore, India

Summary
Securing the sensitive data and computer systems by allowing ease access to authenticated users and withstanding the attacks of imposters is one of the major challenges in the field of computer security. ID and password are the most widely used method for authenticating the computer systems. But, this method has many loop holes such as password sharing, shoulder surfing, brute force attack, dictionary dttack, guessing, phishing and many more. Keystroke Dynamics is one of the famous and inexpensive behavioral biometric technologies, which identifies the authenticity of a user when the user is working via a keyboard. Keystroke features like dwell time, flight time, di-graph, tri-graph and virtual key force of every user are used in this paper. For the purpose of preprocessing Z-Score method is used. Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Genetic Algorithm (GA) algorithm is used with Extreme Learning Machine (ELM) for feature subset selection. In order to classify the obtained results ELM algorithm is used. Comparison of ACO, PSO and GA with ELM respectively is done to find the best method for feature subset selection. From the results, it is revealed that ACO with ELM is best for feature subset selection.

Keywords
Keystroke Dynamics, Z-Score, Feature Selection, Ant colony Optimization (ACO),Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Extreme Learning Machine (ELM).Virtual Key Force,.

1. Introduction

Authenticating a user’s identity will always be an essential part of a secure system. Many of these systems store highly sensitive, personal, commercial, confidential or financial data. Unauthorized access to such data will lead to loss of money or unwanted disclosure of highly confidential data that threats the security of Information. User Authentication prevents unauthorized access of information for providing information security. This is done for the purpose of performing trusted communications between parties for computing applications. User authentication is based on three categories:

- Knowledge - based,
- Object or Token - based, and
- Biometric - based.

The knowledge-based authentication is based on something one knows and is characterized by secrecy. The object-based authentication relies on something one has and is characterized by possession. The Biometric-based user authentication is based on something you are and depends on behavioral and physiological characteristics of individuals. In knowledge-based and object-based approaches, passwords and tokens can be forgotten, lost or stolen. There are also usability limitations associated with them such as managing multiple passwords / PINs, and memorizing and recalling strong passwords which are not easy tasks. Biometric-based person recognition overcomes the above mentioned difficulties of knowledge-based and object based approaches. Biometric authentication is further classified into Physiological and Behavioral types. Physiological Biometric refer to what the person is, and Behavioral Biometrics are related to what a person does, or how the person uses the body. Keystroke dynamics is considered as a strong behavioral biometric based authentication system [1]. It is a process of analyzing the way a user types at a terminal by monitoring the keyboard in order to identify the users based on habitual typing rhythm patterns. Moreover, unlike other biometric systems, which may be expensive to implement, keystroke dynamics is almost free as the only hardware required is the keyboard. There are two approaches in keystroke authentication: Static and Dynamic. Static approach authenticates the user at logon time and Dynamic methods authenticates after logon. Static approach is used in this paper.

2. Related works

Karnan *et al.*, [1] proposed a personal authentication based on keystroke dynamics using soft computing techniques. Particle Swarm Optimization (PSO), Genetic Algorithm (GA) and t Ant Colony Optimization (ACO) are implemented for feature subset selection. Back Propagation Neural Network (BPNN) is applied for classification. Ant colony optimization provides enhanced performance when compared to PSO and GA and obtained

Manuscript received October 5, 2011
Manuscript revised October 20, 2011

92.8% classification accuracy. Yu and Cho [11] proposed a GA – SVM with Gaussian Kernel for feature selection. Ki-seok Sung and Sungzoon Cho [12] used GA – SVM wrapper ensemble method. GA– PSO wrapper approach was proposed by Azevedo et al [13].

3. Methodology

3.1 Feature Extraction

Keystroke data can be obtained by measuring the pressing and releasing time of keys. There are many features that can be measured from the keystrokes. They are Duration, Latency, Digraph, Tri-graph, Pressure of keystroke, Force of Keystroke. Difficulties of typing text, Frequency of word errors, Typing rate, etc. All the features are not useful and widely used. For measuring Pressure and Force of keystroke special type of pressure or force sensitive keyboard is required. Difficulties of typing text, frequency of word errors, typing rate are useful for long text. Since user will be providing only password these features are not suitable. Therefore the timing feature such as Duration or Dwell time, Latency or Flight time, Digraph, Tri-graph are frequently measured from keystroke In addition to the above mentioned timing feature a new feature called Virtual Key force has been introduced.

The virtual key force(VKF) is calculated based on the typing speed and behavior of the user on the key board. It measures the time taken by the user between releasing one key and pressing another key. It is based on the fact that each user has different typing speed and each user takes their own time to release and press another key. The usage of keys and the typing speed and force is different for different users. Also the time interval taken for the release of one key and press of another key is different. Consider a user typing a word which consists of ten letters, hence there exists nine time intervals between the release of one key and press of another key. The average typing speed of the user can be calculated based on these time intervals. Virtual key force can be determined from the key complexity. The key complexity can be calculated as follows

- According to the complexity of usage of the keys, key complexity can be determined. It is based on the key position and distance.
- It means that the middle row keys (i.e., the keys from A to L) on the keyboard which are easy to handle by all the users is taken as 0. The key complexity of remaining keys is taken as 1.

In the figure 1, for the keys T,H,E the complexity label is assigned as CL=(0,1).i.e the distance from T and H is nearer(0) and the distance between H and E is longer(1).

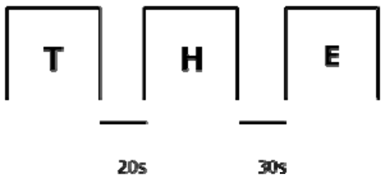

Figure 1. Timing intervals between Keys

Based on the key complexity and the average time interval taken between releasing a key and pressing another key, following algorithm have been formulated:

Algorithm for Virtual Key Force:

```
if (key distance is nearer && time interval is below average)
        VKF=3
else if (key distance is nearer && time interval is above the average)
        VKF =1
else if (keys are longer and the average time interval is below the average)
        VKF =3
else if (keys are longer and the average time interval is above the average)
        VKF =2
end
```

3.2Preprocessing

The extracted feature contains much unnecessary information. Z-Score method is used for Normalization. Preprocessed results are given to the next step called feature subset selection. Given a set of matching scores $\{r_i\}, i = 1, 2, \ldots, n$ the normalized scores are given by

$$n_i = \frac{r_i - \mu}{\sigma}$$

where μ is the arithmetic mean and σ is the standard deviation of the given data.

3.3Feature Subset Selection

Feature subset selection is applied to high dimensional data before preceding the classification step [12]-[15]. Feature subset selection is fundamentally an optimization difficulty, which concerns searching the space of possible features to recognize one that is optimum or near-optimal in accordance with some performance measures, since the objective is to acquire any subset that reduces or to improve a particular measure .

Optimization techniques like PSO, GA and Ant Colony Optimization (ACO) are proposed to select the subset of features from the extracted features which are obtained after preprocessing. These optimization techniques are integrated with Extreme Learning Machine (ELM) for feature subset selection process that can automatically selects an appropriate subset of features and the rest will not be considered, thus resulting in a more comprehensive model.

3.3.1Extreme Machine Learning

ELM technique is used as objective function in GA,PSO,ACO for feature selection and it is described as below [18]:

- At first, the input weights and hidden biases are generated randomly.
- Subsequently, the corresponding output weights are determined with the assistance of ELM algorithm only in first step and randomly produce the output hidden biases.
- Then, the parameters (all weights and biases) are restructured with the help of LM algorithm.

3.3.2Particle Swarm Optimization with ELM

Particle swarm optimization is a population dependent stochastic optimization approach, motivated by social activities of bird gathering or fish schooling [9, 11]. The system is initialized with a population of arbitrary solutions and looks for optima by updating generations.

```
Input: Feature values - Dwell Time, Flight Time, Di-Graph, Tri-Graph and Virtual Key Force of every user.
Output: Subset feature values.
For all particle
  Initialize particle
END
Do
  For all particle
    Compute fitness value using ELM
    If the fitness value is greater than the best fitness value (pBest) of ELM in history set current value as the new pBest
  End
  Select the particle with the best fitness value of all the particles which satisfies the fitness value of ELM as the gBest
  For all particle
    Compute particle velocity
    Update particle position
```

Figure 2: PSO Algorithm with ELM

The possible solutions in PSO are called particles. The entire particles follow its coordinates in the trouble space which are related with the best possible solution (fitness). An additional “best” value that is tracked by the particle swarm optimizer is the best value, achieved at any point by any particle in the neighbors of the particle and this location is called Pbest. If a particle considers all the population as its topological neighbors, then the best value is a global best and is called Gbest.

Flow chart for PSO with ELM Approach is given in the following figure. The fitness function of the PSO is determined using ELM.

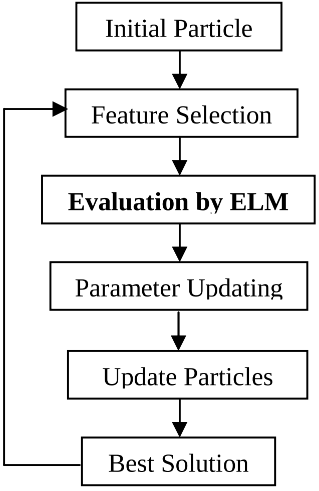

Figure 3: PSO – ELM

3.3.3Genetic Algorithm with ELM

```
Input: Feature values - Dwell Time, Flight Time, Di-Graph, Tri-Graph and Virtual Key Force of every user.
Output: Subset feature values.
START
  Produce initial population.
  Allocate fitness function to each individual.
DO UNTIL best solution is found
  Select individuals from current generation
  Create new off springs with mutation and/or breeding
  Compute new fitness for all individuals using ELM
  Kill all the unfit individuals to give space to new off springs
  Based on ELM check if the best solution is found
  LOOP
END
```

Figure 4: Genetic Algorithm with ELM

Genetic Algorithm is employed as a computer model in which a population of conceptual representations (called chromosomes) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization complexity progresses near better solutions.

The fitness of all individuals in the population is calculated in every generation, several individuals are stochastically chosen from the existing population (according to their fitness value), and updated (recombined and perhaps arbitrarily mutated) to generate a new population. When the algorithm has completed because of a highest number of generations, a reasonable solution may or may not have been obtained. Genetic Algorithm is illustrated in Figure 4. Flow chart for GA with ELM Approach is given in the following figure. Fitness function of the GA is determined by the ELM.

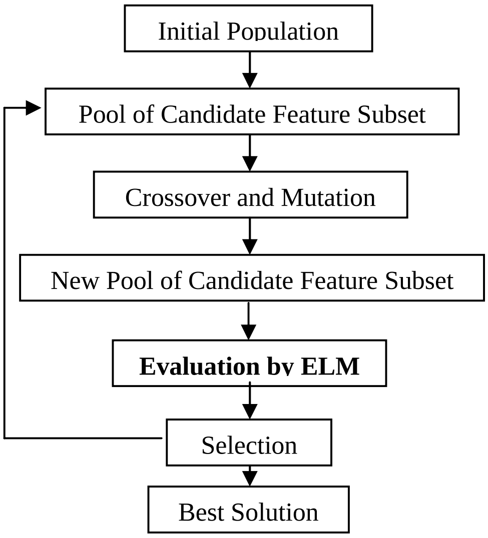

Figure 5: GA – ELM

3.3.4Ant Colony Algorithm with ELM

Ant colony optimization technique is used in this paper for feature subset selection. The ant colony optimization technique has been motivated by the examination on real ant colony's foraging activities, and on that ants can frequently identify the shortest path between food source and their nest. Ants will broadcast information with the help of volatile chemical substances that ants left in its crossing way and also called as the "pheromone" and then reach the intention of identifying the shortest path to identify food sources. An ant identifying an already laid trail can identify the thickness of pheromone trail. It chooses with high probability to follow a shortest path and strengthen that trail with its own pheromone.

The huge quantity of pheromone is on the certain path, the higher probability is that an ant chooses that path and the paths pheromone trail will become harder. Finally, the ant colony together plots the shortest path that has the higher pheromone density. Such easy indirect communication way between ants embodies really a type of collective leaning technique. As in the Figure 6, based on ACO - ELM technique, the optimum feature is choose from every group and only those chosen features is further employed in the classification phase.

The flow chart of ACO with ELM is provided in Figure 6

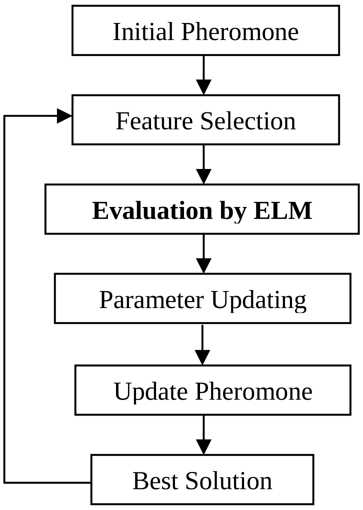

Figure 6: ACO – ELM

4. Experimental Results

The proposed system is experimented with KSP Dataset [19]. This dataset is a Public Typing Database Created and collected by Jeffrey D. Allen Southern Methodist University. This dataset represents the typing of 103 individuals on three different words. Each user typed anywhere from 7 to 503 entries (with an average of 26 entries per user). Dwell time, flight time, di-graph, tri-graph and virtual key force of every user were calculated for all the samples typed by every user. This obtained sample is helpful for learning phase. Then by using PSO, GA, and ACO with ELM the feature subset selection is carried out.

GA, PSO and ACO with ELM respectively are applied for feature subset selection. For evaluating the three proposed approaches, considered a password 'drizzle' from the KSP Dataset. After preprocessing the extracted features, the number of features obtained is 43. The following table illustrates the number of selected after the execution of the proposed GA, PSO and ACO with ELM respectively.

TABLE 1: COMPARISON OF NUMBER OF FEATURES SELECTED

Method	***No. of Features***	***No. of Features Selected***	***Feature Reduction %***
GA with ELM	*43*	*28*	34.88
PSO with ELM	*43*	*30*	*30.23*
ACO with ELM	*43*	*23*	46.51

From the table, it is clear that the proposed ACO with ELM reduces 46.51% of the total features and selects only less number of appropriate features when compared with the other two proposed approaches GA with ELM and PSO with ELM.

5. Conclusion

Security and authentication are the most considerable problem in computer systems or networks. There are many techniques exists based on different biometrics features like fingerprint, iris, etc. But many of these can be easily cracked and also very expensive for getting the scanning device. To overcome these drawbacks, keystroke pattern is used in this paper. Keystroke features like dwell time, flight time, di-graph, tri-graph and virtual key force of every user are used in this paper. ACO, PSO and GA with ELM algorithm respectively is used for feature subset selection. Experimental result compares ACO, PSO and GA with ELM respectively and revealed that the ACO with ELM is the best method for feature subset selection, since the number of features selected is very low compared with other methods.

References

[1] Karnan, M. and Akila, M., "Personal Authentication Based on Keystroke Dynamics Using Soft Computing Techniques", Second International Conference on Communication Software and Networks, Pp.334-338, 2010.

[2] Giroux, S., Wachowiak-Smolikova, R. and Wachowiak, M.P., "Keystroke-Based Authentication by Key Press Intervals as a Complementary Behavioral Biometric", IEEE International Conference on Systems, Man and Cybernetics, Pp. 80-85, 2009.

[3] Mandujano, S. and Soto, R., "Deterring Password Sharing: User Authentication via Fuzzy C-Means Clustering Applied to Keystroke Biometric Data", Proceedings of the Fifth Mexican International Conference in Computer Science, Pp. 181-187, 2004.

[4] Giroux, S., Wachowiak-Smolikova, R. and Wachowiak, M.P., "Keypress Interval Timing Ratios as Behavioral Biometrics for Authentication in Computer Security", First International Conference on Networked Digital Technologies, Pp. 195-200, 2009.

[5] Karnan, M. and Akila, M., "Identity Authentication Based on Keystroke Dynamics using Genetic Algorithm and Particle Swarm Optimization" , 2nd IEEE International Conference on Computer Science and Information Technology, Pp. 203-207, 2009.

[6] Kenneth Revett, Florin Gorunescu, Marina Gorunescu, Marius Ene, Sergio Tenreiro de Magalhaes and Henrique M. Dinis Santos, "A Machine Learning Approach to Keystroke Dynamics Based User Authentication", International Journal of Electronic Security and Digital Forensics, Vol. 1, No. 1, 2007.

[7] Saevanee, H. and Bhattarakosol, P., “Authenticating User Using Keystroke Dynamics and Finger Pressure”, Proceedings of 6th IEEE Conference on Consumer Communications and Networking, Pp. 1-2, 2009.

[8] Leggett, J. and Williams, G., “Verifying Identity via Keystroke Characteristics”, International Journal of Man-Machine Studies, Vol. 28, No. 1, Pp. 67-76, 1998.

[9] Mroczkowski, P. and Choras, M., “Keystroke Dynamics in Biometrics Client-Server Password Hardening System”, Proceedings of Advanced Computer Systems, Vol. 2, Pp. 75-82, 2006.

[10] Yang, J. and Honavar, V., “Feature Subset Selection using a Genetic algorithm”, IEEE Expert: Intelligent Systems and their Applications, Vol. 13, Pp. 44-49, 1998.

[11] Enzhe Yu and Sungzoon Cho, “Keystroke Dynamics Identity Verification - Its Problems and Practical Solutions”, Computers and Security, Vol. 23, Pp. 428-440, 2004.

[12] Ki-seok Sung and Sungzoon Cho, “GA SVM Wrapper Ensemble for Keystroke Dynamics Authentication”, Proceedings of International Conference on Biometrics, Vol. 3832, pp. 654-660, 2006.

[13] Gabriel L. F. B. G. Azevedo, George D. C. Cavalcanti and E.C.B. Carvalho Filho, “An Approach to Feature Extraction for Keystroke Dynamics Systems based on PSO and Feature Weighting”, IEEE Congress on Evolutionary Computation, Pp. 3577-3584, 2007.

[14] Saleh Bleha, Charles Slivinsky and Bassam Hussein, “Computer-Access Security Systems Using Keystroke Dynamics”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Pp.1217-1222, 1990.

[15] Marcus Brown and Samuel Joe Rogers, “User Identification Via Keystroke Characteristics of Typed Names Using Neural Networks”, International Journal of Man-Machine Studies, Pp. 999-1014, 1993.

[16] Rick Joyce and Gopal Gupta, “Identity Authorization Based on Keystroke Latencies”, Communications of the ACM, Pp.168-176, 1990.

[17] John G.H, Kohavi .R and Pfleger .K, ‘Irrelevant Features and the Subset Selection Problem’, Proceedings of Eleventh International Conference on Machine Learning, Morgan Kaufmann, San Francisco, pp. 121-129, 1994.

[18] A.Bharathi and A.M.Natarajan, “Cancer Classification using Modified Extreme Learning Machine based on ANOVA Features”, European Journal of Scientific Research, pp.156-165, 2011.

[19] http://jdadesign.net/2010/04/pressure-sensitive-keystroke-dynamics-dataset/