GA-SVM Wrapper Approach for Feature Subset Selection in Keystroke Dynamics Identity Veri£cation

Enzhe Yu and Sungzoon Cho
Department of Industrial Engineering, Seoul National University
San 56-1, Shillim Dong, Kwanak-Gu, Seoul 151-744 Korea
Email: {enzhe | zoon}@snu.ac.kr

Abstract— **Password is the most widely used identity veri£cation method in computer security domain. However, due to its simplicity, it is vulnerable to imposter attacks. Keystroke dynamics adds a shield to password. Password typing patterns or timing vectors of a user are measured and used to train a novelty detector model. However, without manual pre-processing to remove noises and outliers resulting from typing inconsistencies, a poor detection accuracy results. Thus, in this paper, we propose an automatic feature subset selection process that can automatically selects a relevant subset of features and ignores the rest, thus producing a better accuracy. Genetic algorithm is employed to implement a randomized search and SVM, an excellent novelty detector with fast learning speed, is employed as a base learner. Preliminary experiments show a promising result.**

I. Introduction

Password is the most widely used identity veri£cation method in computer security domain. However, due to its simplicity, it is vulnerable to imposter attacks. One way to add a shield to password is to employ keystroke dynamics in password typing. Keystroke dynamics is a biometric-based approach that utilizes the manner and rhythm in which each individual types password. It measures the keystroke rhythm of a user in order to develop a template that identi£es the authorized user. When a user types a word, for instance a password, the keystroke dynamics can be characterized by a "timing vector", consisting of the durations of keystrokes and the time intervals between them. The owner's timing vectors are collected and used to build a model that discriminates between the owner and imposters. This idea comes originally from the observations that a user's keystroke pattern is highly repeatable and distinct from others'.

In 1980, Gaines et al. [3] £rst proposed the approach using keystroke dynamics for user authentication. Experiments with a population of 7 candidates were conducted. Later on, Leggett et al. [4] conducted similar experiments by applying a long string of 537 characters, and reported a result of 5.0% False Acceptance Rate (FAR: the rate that an imposter is allowed access) and 5.5% False Rejection Rate (FRR: the rate that a legitimate user is denied access). Recently, through the use of neural networks, a comparable performance of 12% to 21% was achieved using short strings such as real-life names [4]. Obaidat and Sadoun reported a 0% error rate in user veri£cation using 7-character-long login name [6]. However, the assumptions were impractical. First, both the imposter's typing patterns and the owner's patterns were used in training. In case of "password typing," the imposter patterns are not available. Second, a huge training data set of 6,300 owners and 112 negatives were used. Third, the training and test patterns were not chronologically separated. In [1], a novelty detection model, i.e. 2-layer AaMLP, was built by training owner's patterns only, and was used to detect imposters using some sort of a similarity measure, and a 1.0% FRR and 0% FAR was reported. However, 2-layer AaMLP usually works well for linear patterns, while performs bad for nonlinear data. Recently, Yu and Cho [10] applied non-linear models, i.e. 4-layer AaMLP and support vector machine (SVM) [7], [8] and reported improved results.

Novelty detection models are built under the assumption that the owner's typing follows a consistent pattern. But usually there are some problems with the original data due to owner's inconsistency. Irrelevant or redundant features are generated, thus lead the novelty detector to bad performance. One usual way to tackle such problems is by manually preprocessing data, i.e. removing noises or outliers from training patterns [10]. However, data preprocessing is subjective, and is not allowed in automated identity veri£cation. Practically, identity veri£cation allows no human intervention to deal with the owner's raw pattern, and an automated process is preferred. Without manual preprocessing, even the best-so-far method resulted in a relatively poor performance of 15% FRR when FAR was set to 0%. One way to improve the performance is to employ an "automatic" preprocessing process.

In this paper, we propose a feature subset selection process that can automatically selects a relevant subset of features and ignores the rest, thus resulting in a more comprehensive model. In particular, a Genetic Algorithm-Support Vector Machine (GA-SVM) based "wrapper" approach for feature subset selection was applied to keystroke dynamics identity veri£cation problem. Experimental results from the proposed approach were compared with the results from the approach without feature selection.

This paper is structured as follows. In session 2, descriptions on the proposed GA-SVM wrapper approach are presented. Session 3 explains the data and experimental settings, and experimental results. A conclusion and limitation of the current work then follows.

II. GA-SVM based Wrapper Approach

Feature subset selection is essentially an optimization problem, which involves searching the space of possible

0-7803-7898-9/03/$17.00 ©2003 IEEE

features to identify one that is optimum or near-optimal with respect to certain performance measures (e.g., accuracy, learning time, etc.) Various ways to perform feature subset selection exist.

Let us consider them in terms of two different perspectives. First, according to the characteristics of the search strategy, feature subset selection algorithms can broadly be classi£ed into three categories [9]: (a) *Exhaustive search*, which is computationally infeasible in practice, except in those rare instances where the total number of features is quite small; (b) *Heuristic search*, which is often used in conjunction with branch and bound search. But heuristic search assumes that the performance criterion is monotone, therefore it only works well with linear classi£ers, and shows bad results with non-linear classi£ers such neural network; (c) *Randomized search*, which uses randomized or probabilistic steps or sampling processes. Prominent among the randomized search algorithms is genetic algorithm, which does not require the restrictive monotonicity assumption.

Second, feature subset selection algorithms can be classi£ed into two categories based on whether or not feature selection is done independently of the learning algorithm used to construct the classi£er. If feature selection is performed independently of the learning algorithm, the technique is said to follow a £lter approach. Otherwise, it is said to follow a wrapper approach. Filter approach is generally computationally more ef£cient than the wrapper approach, its major drawback is that an optimal selection of features may not be independent of the inductive and representational biases of the learning algorithm that is used to construct the classi£er. The wrapper approach on the other hand, involves the computational overhead of evaluating candidate feature subsets by executing a selected learning algorithm on the dataset represented using each feature subset under consideration. This is feasible only if the learning algorithm used to train the classi£er is relatively fast [9].

A. SVM for Novelty Detection

SVM is commonly used to solve two-class classi£cation. But, recently, Schölkopf et al. [7], [8] extended the support vector machine methodology to "one-class" classi£cation, i.e. novelty detection problem. Other support vector data description (SVDD) approaches such as [12] used a concept of *balls* to describe the data in a feature space. For Gaussian kernels, these two approaches can be proved to be equivalent.

Our approach is based on the one-class classi£cation algorithm which was proposed by Schölkopf et al.[7], [8]. The idea is to map the data into the feature space corresponding to the kernel, and to separate them from the origin with a maximum margin. The algorithm returns a decision function f that takes the value +1 in a 'small' region capturing most of the normal data, and -1 elsewhere. For a new point x, the value $f(x)$ is determined by evaluating which side of the hyperplane it falls on in the feature space.

Let $x_1, x_2, ..., x_l \in$ X, where $l \in$ N denotes the number of normal data, and X a compact subset of R^N corresponding to the one class. Let Φ be a feature map X $\rightarrow$ F, which transforms the training data to a dot product space F such that the dot product in the image of Φ can be computed by evaluating some simple kernel

$$k(x, y) = \big(\Phi(x) \cdot \Phi(y)\big). \tag{1}$$

In case of Gaussian kernel,

$$k(x, y) = e^{-\|x-y\|^2/s}. \tag{2}$$

To separate the normal data set from the origin, one needs to solve the following quadratic programming problem:

$$\min_{w \in F, \xi \in R^l, \rho \in R} \frac{1}{2} \| w \|^2 + \frac{1}{\nu l} \sum_i \xi_i - \rho. \tag{3}$$

$$\text{subject to } \big(w \cdot \Phi(x_i)\big) \geq \rho - \xi_i, \ \xi_i \geq 0.$$

Since nonzero slack variables ξ_i are penalized in the objective function, it can be expected that if w and ρ solve this problem, then the decision function

$$f(x) = sgn\big(w \cdot \Phi(x) - \rho\big). \tag{4}$$

will be positive for most examples x_i contained in the training set, while the SV type regularization term $\| w \|$ will still be small. The trade-off between these two goals is controlled by $\nu \in (0, 1)$.

Deriving the dual problem, and (1), the solution can be shown to have an SV expansion

$$f(x) = sgn\Big(\sum_i a_i k(x_i, x) - \rho\Big). \tag{5}$$

(patterns x_i with nonzero a_i are called support vectors), where the coef£cients are found as the solution of the dual problem:

$$\min_a \frac{1}{2} \sum_{ij} a_i a_j k(x_i, x_j), \tag{6}$$

$$\text{subject to } 0 \leq a_i \leq \tfrac{1}{\nu l}, \ \ \textstyle\sum_i a_i = 1.$$

If ν approaches 0, the upper boundaries on the Lagrange multipliers tend to in£nity, thus the problem then resembles the corresponding hard margin algorithm. If ν approaches 1, then the constrains only allow one solution, that where all a_i are at the upper bound $1/(\nu l)$. In this case, for kernels with integral 1, i.e. Gaussian kernel, the decision function corresponds to a Parzen windows estimator with threshold [5], [8], [12].

In our research, we used used Chang and Lin's toolbox LIBSVM[2], which was based on the "one-class" SVM algorithm developed by Schökolpf et al. As for the parameters, since we evaluate the feature subset and training model according to FRR when FAR=0, therefore a heuristic search was applied for tuning parameters.

B. Genetic Algorithm

GAs are stochastic search techniques based on the mechanism of natural selection and genetics. A typical GA starts with an initial set of random solutions called *population* and each individual in the population is called a *chromosome*. A chromosome is usually, but not necessarily, a binary string and represents a solution to the problem on hand. Chromosomes evolve through successive iterations, called *generations*. During each generation, the chromosomes are evaluated, using some measures of £tness. To create the next generation, new chromosomes, called *offspring*, are formed either by *(a)* merging two chromosomes from the current generation using a *crossover* operator, or *(b)* modifying a chromosome using a *mutation* operator. A new generation is formed by: *(a)* selecting some of the parents and offspring according to their £tness values, and *(b)* rejecting the rest so that the population size is kept constant. In the process, £tter chromosomes have a higher chance of being selected. After several generations, the algorithm converges to the best set of chromosomes, which hopefully represent the optimum or near optimal solution to the problem. Table I describes the general structure of GA, where *P(t)* and *C(t)* denote the parents and offspring or children set in the current generation.

GAs are becoming popular as a technique for solving optimisation and search problems mainly because of three distinct advantages they have over their competition [11]: (1) GAs do not involve sophisticated mathematics; (2) The operators' ergodicity of evolution makes GAs very effective at performing global search; (3) GAs are ¤exible in that they readily allow for hybridization with domain dependent heuristics, which can result in a more powerful search routine for a speci£c problem.

TABLE I
GELETIC ALGORITHM

```
Procedure: Genetic Algorithm

  Begin
    t ← 0;
    Initialization P(t);
    Evaluation P(t);
    While (not termination condition)
      Do
      Recombine P(t) to yield C(t);
      Evaluation C(t);
      Selection P(t+1) from P(t) or P(t)+C(t);
      t ← t+1;
    End
  End
```

C. GA-SVM Wrapper

Our choice is the randomized wrapper approach. Speci£cally, we chose genetic algorithm paradigm for randomization and SVM as a base learner in wrapper approach. In other words, a population of feature subsets are evolved through the mechanism of genetic algorithm and a feature subset is evaluated through training and testing a SVM with the data set. GAs are stochastic search techniques based on the mechanism of natural selection and genetics, and are generally quite effective for rapid global search of large search spaces in dif£cult optimization problems. Previous researches have reported the feasibility of GA for wrapper approach to feature subset selection [9]. SVM also suits as a base learner well due to its quick training capability. In our previous study, SVM novelty detector was found to result in a comparable performance with that of neural network, but the learning time is much faster than that of neural network, i.e. less than 1/1000 times of neural network's learning time [10]. An initial population is made up of diversi£ed binary strings indicating the features selected.

These candidates undergo crossover and mutation, evaluated by the SVM base learner. Only those that are selected according to the speci£ed multi-criteria £tness are put back into the population and the process is repeated for a £xed number of generations. The best solutions are achieved in the end (see *Fig. 1*).

In the proposed GA-SVM wrapper approach, a Gaussian kernel is used for the induction algorithm, i.e. SVM, and the parameters were tuned through some heuristic method. The GA was implemented with the following settings. The chromosome is a binary string where each bit denotes whether the corresponding feature is present (1) or absent (0). The population size was generally set at 30, but when the population diversity resulted in an unsatisfactory performance, it was modi£ed up to 50. The crossover rate of 0.6, and the mutation rate of 0.01~0.02 were adopted with corresponding mechanisms being two-point crossover and uniform mutation, respectively. Selection provides the driving force in the evo-

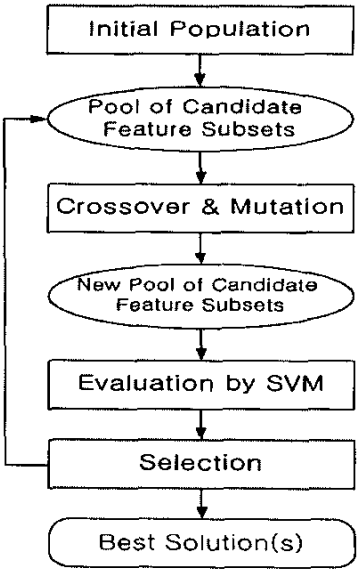

Fig. 1. GA-SVM Wrapper Feature Subset Selection

lutionary process, and the selection pressure is critical. At the early stage of evolution, a low selection pressure is preferred for a wide exploration of the search space. At the end of evolution, where the population is near convergence, however, a high selection pressure is taken to exploit the most promising regions of the search space [11]. As for the sampling space, a regular one was chosen, which has the size of the specified population and is made up of all the offspring and only part of parents. The sampling mechanism follows the probabilistic roulette wheel selection. In order to discriminate among the similar strong individuals in the last 10%~20% generations, a linear scaling method was applied to deal with the selection probability.

The fitness function combined three different criteria, i.e. the accuracy of the novelty detector, the learning time used, and the dimension reduction ratio. One definitions of the fitness function emphasized more on the accuracy:

$$Fitness(x) = \frac{1}{DimRat(x)} + \frac{1}{100 \times LrnT(x)} + 10 \times Acc(x). \quad (7)$$

where $Fitness(x)$ is the fitness of the feature subset represented by x, $Acc(x)$ is the test accuracy of the SVM novelty detector using the feature subset represented by x, and $LrnT(x)$ is the time taken to train the SVM.

Although usually the test accuracy is the only most important criterion, we also include dimension reduction ratio and training time into the fitness function in that when the model show comparable results, the model with least training time which is important in practical application, and the feature subset with the smaller dimension which is less suspectable to introduce irrelevant or redundant features, are more preferred. In fact, proper tradeoff values among the multiple objectives have to be based on the knowledge of the problem domain or the experimental results.

III. Experiments and Results

Experiments were implemented using the GA-SVM wrapper approach with the data set identical to that of [10]. The goal was to observe the effect of the feature subset selection. A comparison of the model performances was made between the results of before- and after- feature subset selection.

A. Data Collection

The data was captured by a program in X window environment on a Sun Sparc-Station, in which the keystroke duration times and interval times was measured. The keystroke duration and interval times were captured at the accuracy of milliseconds (*ms*). A timing vector consists of keystroke duration times and interval times. A password with *n*-character long would result in the timing vector of dimension (*2n+1*), with the Enter key included. For instance, a password *abcd*, which is 4-character long (n = 4), together with the *Enter* key, results in a timing vector of 9 dimension.

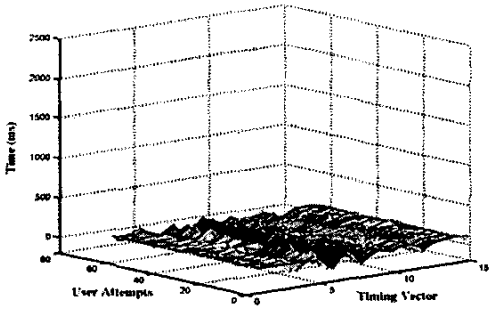

Fig. 2. Example of Owner's Patterns

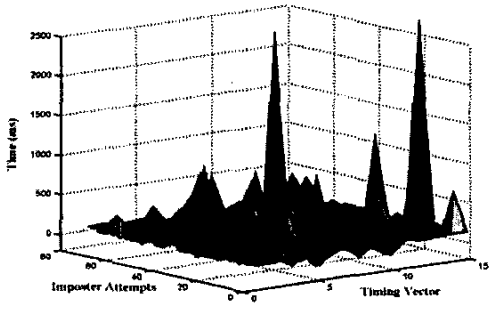

Fig. 3. Example of Impoters' Patterns

An example of a timing vector is [30, 60, 70, 135, 60, -35, 75, 40, 55]. A negative keystroke interval time results when a next key is stroked before a previous key is released.

The owners' data was collected from 21 participants with different passwords, whose length ranges from 6 to 10. Each participant was asked to type his password 150 to 400 times, and the last 75 timing vectors were collected for testing, whereas the remaining ones were used as training patterns.

As for the novelty data, 15 imposters were given passwords beforehand, and were asked to practice typing these passwords. After that, they type each of the given 21 passwords 5 times, resulting in 75 impostor timing vectors for each password. We call these imposters as "*imposters with practice*." Together with the owners' test patterns mentioned above, two groups of 75 patterns, i.e. normal and novelty, are collected for each password.

When compared to other problems, the test data set is quite small, this is due to the limited participants in the data collection process. However, these imposters practiced typing the given password in many ways and can be regarded as somewhat representative of the practical situations, although these test sets still need to be extended in terms of diversity. Fig. 2 and Fig. 3 illustrated timing vectors of a certain password for the owner and "imposters with practice", respectively.

B. Experimental Results

Experiments were carried out on the 21 datasets as was described in the previous section. Since it can be assumed that the "imposters with practice" resemble the owners more than those "imposters without practice" in the way they type, and are more representative of practical situations, we evaluated the model performance by using the data collected from the owners and the corresponding "imposters with practice".

In the proposed GA-SVM approach, the novelty detector, i.e. SVM, is built by using the owner's pattern only, and GA makes a large number of evaluations in the evolutionary process, a certain number of validation sets for evaluation are required. This was solved by conducting cross-validation method in the experiments.

Table 2 compared model performance with raw data and the data after feature subset selection. Raw data contained noise or outliers, and resulted in bad performance without data cleaning or feature selection. Experiments reported an average FRR of 15.78%, with minimum FRR of 5.3% and maximum FRR of 20.38%. Wrapper approach greatly reduced the dimension of the timing vector, and improved the model accuracy as well.The average dimension of the feature subset is 5.86, the smallest feature subset was reduced even down to 3. Different features were selected for different passwords. For "atom" password dataset, for instance, the dimension was reduced from 15 to 5, with corresponding feature subset being "100100001100100." Only three keystroke durations and two keystroke intervals were selected. For other passwords, different features were selected. Model accuracies were all improved with the selected feature subset. The average FRR was reduced from 15.78% down to 3.54% .

TABLE II

FRR WHEN FAR=0 (DIMENSION) BEFORE- AND AFTER-FSS

Owner ID	FRR (Dimension)	
	Before-FSS	After-FSS
Atom	19.32 (15)	2.68 (5)
Bubugi	19.75 (17)	4.39 (4)
Celavie	14.87 (17)	4.55 (4)
Crapas	19.82 (19)	3.60 (5)
Dry	19.88 (19)	4.23 (7)
Flower	15.15 (13)	3.25 (7)
Gmother	19.80 (17)	1.25 (11)
Gusegi	20.26 (15)	3.45 (6)
Jmin	14.85 (17)	3.25 (7)
June	5.30 (17)	1.97 (6)
Jywoo	12.40 (15)	4.35 (4)
Megadeth	10.53 (17)	4.68 (5)
Oscar	10.14 (17)	4.61 (6)
Perfect	12.79 (17)	3.49 (8)
Shlee	12.62 (17)	3.18 (7)
Sjlee	11.71 (13)	4.39 (3)
Woo	13.29 (13)	2.81 (5)
Wooks	19.74 (17)	3.95 (5)
Yanwenry	19.44 (17)	2.78 (4)
Ysoya	20.38 (17)	3.85 (6)
Zeronine	19.26 (21)	3.71 (8)
Minimum	5.30 (13)	1.25 (3)
Maximum	20.38 (21)	4.68 (11)
Average	15.78 (16.52)	3.54 (5.86)

IV. CONCLUSION

In this paper, we proposed GA-SVM based wrapper approach for feature subset selection for keystroke dynamics identity veri£cation. SVM showed its excellence in both accuracy and learning speed, and was proved to be a suitable learner in the wrapper approach. A comparison was made between the performances of before- and after- feature subset selection and promising results were achieved.

Further investigation is necessary regarding the following issues: £rst, practically the users are unwilling to type passwords hundreds of times, thus result in insuf£cient training data. In such a situation, existing approaches may not work so well. A new method has to be studied to deal with such circumstances. Second, in order for a benchmark study, standard keystroke dynamics data like UCI data is called for.

ACKNOWLEDGMENT

This research was supported by Brain Science and Engineering Research Program sponsored by Korean Ministry of Science and Technology, and (in part) by KOSEF through Statistical Research Center for Complex Systems at Seoul National University. The authors would like to thank...

REFERENCES

[1] S. Cho, C. Han, D. Han, and H. Kim, Web-based keystroke dynamics identity veri£cation using neural network, *Journal of organizational computing and electronic commerce 10(4)*, 295-307, 2000.

[2] C. Chang, C. Lin, LIBSVM - A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[3] R. Gaines, W. Lisowski, S. Press, and N. Shapiro. Authentication by keystroke timing: some preliminary results. Rand Report R-256-NSF. Rand Corporation, 1980.

[4] J. Leggett, G. Williams, M. Usnick, and M. Longnecker, Dynamic identity veri£cation via keystroke characteristics, *International Journal of Man-Machine Studies, vol. 35*, pp. 859-870, 1991.

[5] L. M. Manevitz, Malik Yousef, One-Class SVMs for document classi£cation, *Journal of Machine Learning Research 2, 139-154*, 2001.

[6] M. Obaidat and S. Sadoun, Veri£cation of computer users using keystroke dynamics, *IEEE Transactions on Systems, Man and Cybernetics, Part B:P Cybernetics, vol. 27, no. 2, pp. 261-269*, 1997.

[7] B. Schölkopf, J. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution, Technical Report MSR-TR-99-87. Microsoft Research, Redmond, WA, 1999.

[8] B. Schölkopf, R. C. Williamson, A.J. Smola, J. Shawe-Taylor and J.C. Platt. Support vector method for novelty detection. *Advances in Neural Information Processing Systems 12, 582-588. (Eds.) S.A. Solla, T.K. Leen and K.-R. Mller, MIT Press*, 2000.

[9] J. Yang, V. Honavar, Feature subset selection using a genetic algorithm, Feature Extraction, *In Construction, and Subset Selection: A Data Mining Perspective. Motoda, H. and Liu, H. (Ed.) New York: Kluwer*, 1998.

[10] E.Yu, S.Cho, Novelty detection approach for keystroke dynamics identity veri£cation, *Fourth International Conference on Intelligent Data Engineering and Automated Learning*, 2003.

[11] Gen, M., Cheng, R., Genetic algorithm & engineering design, John Wiley & Sons, Inc.

[12] D.M.J. Tax and R.P.W. Duin, Outliers and data descriptions. *Proc. ASCI 2001, 7th annual conference of the advanced school for computing and imaging (Heijen, NL, May 30-June 1), ASCI, Delft, 234-241*, 2001.