키스트로크 인식을 위한 패턴분류 방법

조태훈*

Pattern Classification Methods for Keystroke Identification

Tai-Hoon Cho*

본 연구는 산업자원부 지방기술혁신사업(RTI04-01-02) 지원으로 수행되었음

요 약

키스트로크 시간간격은 컴퓨터사용자의 검증 및 인식에서 분별적인 특징이 될 수 있다. 본 논문은 키스트로크 시간간격을 특징으로, 신경망의 역전파 알고리즘과 Bayesian 분류기, 그리고 k-NN을 이용한 분류기의 사용자 인식 성능을 비교 실험하였다. 실험결과, 사용자당 샘플의 개수가 작을 경우에는 k-NN 알고리즘이 가장 성능이 좋았고, 사용자당 샘플의 개수가 많을 경우에는 Bayesian 분류기의 성능이 가장 뛰어난 결과를 보였다. 따라서 웹기반 온라인 사용자인식을 위해서는 사용자별 키스트로크 샘플의 수에 따라 k-NN이나 Bayesian 분류기를 선택적으로 사용하는 것이 바람직할 것으로 보인다.

ABSTRACT

Keystroke time intervals can be a discriminating feature in the verification and identification of computer users. This paper presents a comparison result obtained using several classification methods including k-NN (k-Nearest Neighbor), back-propagation neural networks, and Bayesian classification for keystroke identification. Performance of k-NN classification was best with small data samples available per user, while Bayesian classification was the most superior to others with large data samples per user. Thus, for web-based on-line identification of users, it seems to be appropriate to selectively use either k-NN or Bayesian method according to the number of keystroke samples accumulated by each user.

키워드

keystroke analysis, k-nearest neighbor, neural network, Bayesian classifier

Ⅰ. 서 론

현대 사회에서 빠르게 형성되어가는 정보화에 따라 불법적인 접근 등 정보화 역기능 역시 빠르게 늘어나고 있다. 이 같은 불법행위를 방지하고, 인터넷 공간상에 안전한 정보 교류를 위해 방화벽(firewall), 가상 사설망(Virtual Private Network: VPN) 등과 같은 각종 보안 대책들이 나오고 있다. 그러나 인터넷의 개방성, 소스 개방, 용이한 침입자간의 상호 정보교환 등 어느 솔루션이나 보안상의 취약한 특성을 갖고 있기 때문에 보안 정책에 따라 통합적이며 균형 있게 설치 운영하는, 이른바 보안 솔루션 다각화가 최근의 추세이다.

* 한국기술교육대학교 정보기술공학부

접수일자 : 2006. 1. 31

이러한 다각적인 보안 솔루션 중 하나로 최근 개인의 키 입력 시간패턴을 인식하여 패스워드 등과 같은 키보드 입력 방식(keystroke)의 보안키를 인증하는 새로운 2차 보안 솔루션에 관심이 모아지고 있다. 이것은 해커에 의해 패스워드가 노출되었다 하더라도 원래 개인이 가지고 있던 keystroke 시간패턴과 다르면 로그인(log-in)되지 않도록 하는 일종의 생체인식 방법 중 하나이다. 이에 keystroke를 빠르고 정확하게 인식하기 위한 방법으로 k-NN(nearest neighbor)기반 패턴분류기와 기존에 많이 연구되어왔던 역전파신경망(back-propagation neural network) 과 Bayesian 기반 패턴인식 알고리즘을 비교한다.

Ⅱ. 패턴분류기법

keystroke 보안의 기본적인 개념은 키보드 상에 패스워드를 입력하게 되면, 각각의 문자 혹은 숫자들 간의 시간차(time interval) 패턴을 인식하여 사용자를 인증(authentication)하는 것이다. 이러한 타이핑(typing) 패턴으로 사용자를 인증 할 수 있다고 처음으로 제안한 것은 1980년에 Gaines 등[1]으로, 이중음자 사이 시간(digraph time)의 평균(mean)을 이용하여 7명을 대상으로 실험을 진행했었다. 그 후, Leggett과 Williams [2][3]도 17명의 프로그래머를 상대로 537개의 긴 문자들을 사용하여 Gaines와 유사한 실험을 했었다. 그 결과, Gaines의 경우 FAR과 FRR이 각각 0%, 4%이었던 반면, Leggett등의 경우 5%, 5.5%로 상대적으로 Gaines의 실험이 우수한 성능을 보였다. 최근에는 Ord와 Furnell[4]에 의해 모든 키보드(full keyboard)를 사용했던 이전 연구와는 달리 ATM(Automate Teller Machine)등의 숫자 키패드(keypad)만을 사용하여 6개의 숫자를 가지고 사용자를 인증하는 실험을 했었다. 또한, 'BioPassword'[5]란 이름으로 keystroke를 이용한 패스워드 모니터링(password monitoring) 프로그램이 현재 실용화되고 있다.

keystroke 분석을 위해 주로 사용되는 방식에는 신경망을 이용한 접근방법[6][7]과 Bayesian 이론을 이용한 통계적 접근방법이 있다[8]. 신경망에서는 주로 지도 학습 방식(supervised learning)의 역전파 알고리즘(Back-propagation algorithm)이 많이 사용되고 있다. 이 알고리즘은 keystroke 시간간격 데이터를 입력 데이터로 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)의 3개 계층(layer)으로 이루어진 망(network)에 입력하고, 각 가중치(weight)와 곱하고 더하는 과정을 통해 결과 값을 출력하게 된다. 이 결과 값은 우리가 원하는 목표 값(target value)과 다르기 때문에 오차(error)가 발생하게 된다. 이 오차를 줄이는 방향으로 가중치가 갱신되고, 갱신된 값과 출력값을 다시 입력층으로 입력하여 같은 계산을 반복하게 된다. 이 같은 순환 계산과정에서 결국에는 결과 값과 목표 값의 오차가 줄어들게 되고, 각각의 입력에 따른 목표 값의 출력으로 사용자를 구분할 수 있게 된다[9]. (그림 1)

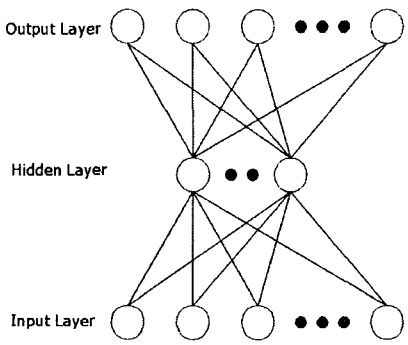

그림 1. 역전파 신경망
Fig. 1 Back-propagation neural network

그러나, 이 알고리즘은 몇 가지 단점을 가지고 있다. 먼저, 약간의 데이터 변화가 있더라도 전체를 다시 학습시켜야 한다는 점이다. 이 과정에서 데이터의 양이 많아질수록 학습시간은 그 만큼 더 오래 걸리게 된다. 또한, 오차함수의 모양이 복잡해 질 경우, 전역적 최소점(global minimum)이 아닌 지역적 최소점(local minima)에 머무를 가능성이 있다. 이렇게 되면 더 이상 학습을 시켜도 오차가 감소되는 학습을 수행 할 수 없게 된다. 이를 보완하기 위한 방법으로 수정 모멘텀, 일괄 수정법, 선택적 재학습 방법 등이 제안되고는 있지만, 이 역시 학습단계에서 얼마간의 시간이 소요된다는 점은 피할 수 없다.

한편, 이와는 다른 통계적 접근방식인 Bayesian 정리를 이용할 경우 학습과정에 걸리는 시간이 신경망에 비해 매우 짧기 때문에 빠른 keystroke 데이터 처리를 할 수 있다는 장점이 있다. 일반적으로 Bayesian 분류기(classifier)는 그림 2와 같은 블록 다이어그램으로 표현할 수 있다[8].

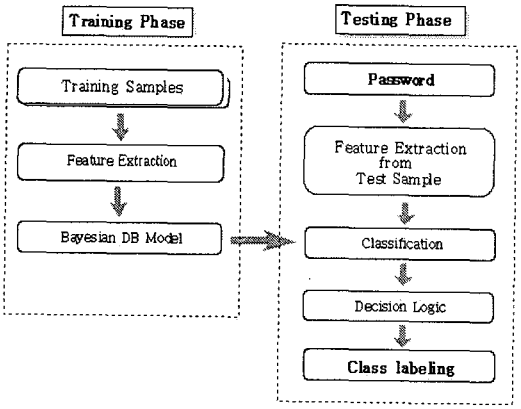

그림 2. 베이지안 분류기 블럭도
Fig. 2 Bayesian classifier block diagram

기존의 패스워드 입력 시 각 단어사이의 시간차 개수를 n개라고 했을 때, n개의 시간 간격 데이터는 n차원의 특징 공간(feature space)안에서 패턴벡터(pattern vector)를 형성하게 된다. 동일한 사용자의 입력패턴들이 가우시안(Gaussian) 분포를 따른다고 가정하면, 사용자 i에 대한 패턴벡터 X의 가우시안 확률밀도함수는 식 (1)로 주어진다.

$$p_i(X)=(2\pi)^{-\frac{n}{2}}|C_i|^{-\frac{1}{2}}\exp[-\frac{(X-m_i)^t C_i^{-1}(X-m_i)}{2}] \quad (1)$$

여기서, m_i는 n차원의 평균벡터(mean vector)이고, C_i는 $n\times n$의 공분산행렬(covariance matrix)로 학습 집합으로부터 각각 식 (2)와 식 (3)에 의해 계산된다.

$$m_i=(1/N_i)\sum_{j=1}^{N_i} x_{ij} \quad (2)$$

$$C_i=(1/N_i)\sum_{j=1}^{N_i} x_{ij}x_{ij}^t - m_i m_i^t \quad (3)$$

N_i는 평균벡터(mean vector)와 공분산의 계산을 위해 사용된 패턴의 개수이다. 모든 사용자에 대한 사전확률(a priori probability)이 같다고 가정하면, 식 (1)을 이용한 Bayesian 분류는 식 (4)와 같이 더욱 간단히 표현할 수 있다.

$$d_i(X)=(X-m_i)^t C_i^{-1}(X-m_i)+\ln|C_i| \quad (4)$$

여기서, 모든 사용자중 식 (4)의 값이 최소가 되는 사용자로 판정하게 된다. 통계적 접근방식의 Bayesian 방법을 이용할 경우 각 사용자의 패턴의 개수가 최소 20개 이상이 되어야 가우시안 분포의 가정이 유효하고, 많은 클래스와 데이터를 사용할 경우 복잡한 산술연산 때문에 다소 처리 시간이 지연될 수 있다는 단점을 가지고 있다. 또한, 샘플들이 비교적 정확한 가우시안 분포를 따르지 않을 경우 정확한 분류를 하기 힘들게 된다.

keystroke의 시간패턴을 분류하는 또 다른 알고리즘으로 non-parametric 패턴분류 방법으로 많이 사용되는 k-NN(k-nearest neighbor)을 사용하였다. k-NN은 n개의 문자간 시간간격의 패스워드로 구성된 n차원의 공간에 사용자당 m개의 패턴을 학습 모델로 분포시켜 놓고, 임의의 사용자에 의해 입력된 패턴과 비교하여 가장 가까운 k개의 패턴을 선택한 후, 이중, 가장 많은 패턴에 해당하는 사용자를 입력패턴에 할당하여 사용자를 구분하게 된다[9].($k\leq m$, k: 홀수)

k-NN의 경우, 신경망처럼 데이터의 일부가 변경된다 하더라도 전체를 다시 학습시켜야 할 필요도 없을 뿐만 아니라, Bayesian 분류기와 같이 가우시안 분포를 따른다는 통계적 가정을 적용시킬 필요가 없다. 또한, 적은 양의 데이터 샘플로도 분류가 가능하다는 장점이 있다.

Ⅲ. 실험 결과

본 연구에서는 'classification'이라는 14개의 문자로 구성된 패스워드를 사용하였는데, 실험자의 키스트로크 입력 특징이 잘 나타날 수 있도록 실험 전 5~10회 가량 연습을 한 후에 본 실험에 임했다. 실험자는 크게 30명과 5명 두 그룹으로 나눠 각각 10개와 60개씩의 샘플을 추출하여 사용했다(표 1). 데이터를 효율적으로 사용하여, 분류기의 성능을 평가하기 위해, *m-fold cross-validation*[9]방법을 사용하였다. 즉, 전체 데이터 세트를 m개의 서로 다른 (disjoint) 같은 크기의 데이터세트들로 임의로 나눈 후, 분류기는 m개중의 어느 하나를 택해 테스트 세트(test set)

로 하고, 나머지((m-1)개)는 학습세트(training set)로 사용하여, m번을 학습하고, 각각을 테스트한 총 결과로서 성능을 평가한다. keystroke 데이터는 각 문자의 키가 눌러진 최초 시각들의 차이를 측정하여, 특징값으로 사용했다.(그림 3 참조) 이 값들을 원소로 하는 13차원의 패턴벡터를 구성하였다. 실험을 위해 PC Pentium 1.7GHz, Memory 256Mbyte, Windows XP에서 Visual C++ 환경을 사용하여 알고리즘을 구현하였다.

표 1. 데이터 세트
Table 1. Data set

	실험자 수	Data 샘플 수/인	Data 샘플 총 개수
Data set #1	30	10	300
Data set #2	5	60	300

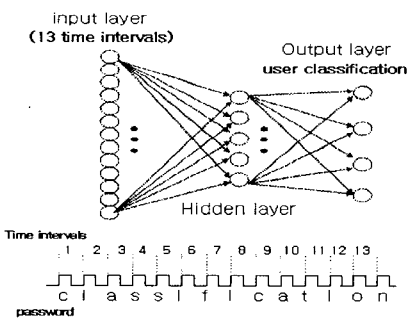

그림 3. 키스트로크 타이밍 데이터
Fig. 3 Keystroke timing data

먼저, 신경망 알고리즘에서는 학습세트 샘플들의 학습이 진행됨에 따라 에러율(error rate)이 감소하게 된다. 에러율은 학습세트 샘플들의 에러(error)의 합으로, 에러율이 0.1이하로 줄어들면 학습을 종료하였다. (에러율이 너무 작아질 때까지 학습을 시키면 과최적화에 빠질 위험이 있다.) 입력층의 노드(node) 개수는 keystroke 시간 간격 개수에 해당하는 13개로, 은닉층의 노드는 23개, 출력층의 노드는 Data #1과 #2의 클래스(class)수, 즉 사용자 수와 같도록 각각 30개, 5개로 설정했다(그림 3). 은닉층의 노드수는 실험적으로 결정되었다. 학습이 끝난 후, 테스트 샘플을 입력하여, 출력값이 가장 큰 노드(사용자)를 택하였다.

표 2에 신경망알고리즘을 적용한 실험결과를 보인다. 오인식률은 오인식된 샘플개수 / 총 테스트 샘플개수(여기서는, 300)로 계산된다. 여기서, m-fold cross validation의 m이 커짐에 따라 오인식률이 감소하는 것을 알 수 있다. 이는 m이 커지면, 학습세트의 샘플수가 증가하기 때문이다. 예를 들어, 10-fold인 경우, 10번의 각 테스트세트의 샘플개수는 300/10=30개이고, 학습세트의 샘플개수는 270개가 된다. 신경망알고리즘의 성능은 양호하나, 학습데이타 샘플의 개수가 많아지면, 학습이 오래 걸리는 단점이 있다.

표 2. 신경망알고리즘을 적용한 결과
Table 2. Result using back-propagation algorithm

Data #1 \ m-fold	2	5	10
오인식 패턴수	12	9	7
오인식률(%)	4.0	3.0	2.3

Data #2 \ m-fold	2	5	10
오인식 패턴수	17	14	12
오인식률(%)	5.7	4.7	4.0

Bayesian 분류의 경우, 먼저, 학습샘플들로부터 각 사용자에 대한 평균벡터와 공분산 행렬을 구하였다. 일반적인 공분산행렬의 요소개수가 13x13=169로 너무 많아, 대각요소(diagonal elements)만 사용하고, 나머지는 0으로 설정하였다. (즉, 각 특징간에 서로 독립적이라고 가정.) 테스트 단계에서, 샘플벡터 X가 입력되면, 각 사용자 i에 대해서, 거리 $d_i(X)$를 구한 후, 이 값이 가장 최소인 사용자로 인식하였다.

Baysian 분류기를 이용한 실험결과가 표 3에 있다. 300-fold의 경우는 전체 데이터 샘플수가 300개이기 때문에, 하나의 샘플을 테스트용으로 하고, 나머지 샘플들을 학습용으로 사용하는 leave-one-out 방법을 나타낸다. Data #1의 경우 한 사람당 sample의 개수가 통계적 가치를 가질 수 있는 20~30개에 크게 못 미치기 때문에 공분산행렬을 구하는 게 무리이지만, 2-fold의 경우를 제외하고는 비교적 양호한 성능을 보여주었다. (2-fold의 경우에는 분산이 0인 경우가 발생하였다.) Data #2는 각 사용자에 대한 샘플의 개수가 충분하여, 가우시안 분포 가정이 잘 맞아, Data #1보다 오인식률이 낮게 나타났다.

표 3. 베이지안 분류기를 적용한 결과
Table 3. Result using Bayesian classification

Data #1 \ m-fold	2	5	10	30	300
오인식 패턴수	-	14	13	11	11
오인식률(%)	-	4.7	4.3	3.7	3.7

Data #2 \ m-fold	2	5	10	30	300
오인식 패턴수	12	10	10	10	10
오인식률(%)	4.0	3.3	3.3	3.3	3.3

표 4. k-NN 알고리즘을 사용한 결과 (k=1)
Table 4. Result using k-NN algorithm (k=1)

Data #1 \ m-fold	2	5	10	30	300
오인식 패턴수	9	6	6	6	6
오인식률(%)	3.0	2.0	2.0	2.0	2.0

Data #2 \ m-fold	2	5	10	30	300
오인식 패턴수	14	13	14	14	14
오인식률(%)	4.7	4.3	4.7	4.7	4.7

k-NN (k=1)로 분류한 실험결과는 표 4에 보인다, 50-fold 이상의 경우(k=5일때, 오인식률이 최저)를 제외하고는 k=1일때, 오인식률이 가장 낮았다. 특이한 점은, 한 사람당 샘플의 수가 작은 Data #1의 경우, 신경망이나 베이지안 분류기보다 성능이 우수하다는 점이다. 하지만, Data #2와 같이, 한 사람당 샘플의 개수가 많은 경우, 신경망이나 베이지안 분류기의 성능보다 못함을 알 수 있다. 다른 특별한 학습과정이나 통계적인 가정없이 손쉽게 적용할 수 있는 점은 k-NN의 장점으로 들 수 있다.

Ⅳ. 결 론

keystroke 데이터를 가지고 신경망의 역전파 알고리즘과 Bayesian 분류기, 그리고 *k*-NN을 사용하여 사용자의 인식성능을 비교 실험하였다. 실험결과, 신경망은 성능은 비교적 양호하였지만, 시간이 걸리는 학습이라는 선행과정이 필요한 단점이 있다. 사용자당 샘플의 개수가 작을 경우에는 *k*-NN 알고리즘이 가장 성능이 좋았고, 사용자당 샘플의 개수가 많을 경우에는 Bayesian 분류기의 성능이 가장 뛰어난 결과를 보였다. 따라서, keystroke 데이터를 이용하여, 사용자를 인식하고자 하는 경우, 각 사용자별 데이터가 일정량(약 20-30개) 축적되기 전에는 *k*-NN을 적용하고, 그 후에는 베이지안 분류기를 사용하는 것이 바람직할 것으로 보인다. 향후, keystroke 분석을 이용하여, 인터넷 웹상에서 사용자 인증시스템을 구현할 계획이다.

참고문헌

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

[2] J. Leggett, G. Williams, "Verifying identity via keystroke characteristics," Int. J. Man-Mach. Stud. vol.28, no.1, pp. 67-76, 1988

[3] J. Leggett, G. Williams, D. Umphress, "Verification of user identity via keystroke characteristics," Human Factors in management Information System, Ablex Publishing Corp., Norwood, NJ, 1988.

[4] T. Ord, S.M. Furnell, "User

저자소개

조 태 훈 (Tai-Hoon Cho)

1981년 서울대학교 전자공학과 학사
1983년 한국과학기술원 전기 및 전 자공학과 석사
1991년 Virginia Polytechnic Institute & State University 박사
1992년~1998년 LG산전 연구소 책임/수석 연구원
1998년~현재 한국기술교육대학교 정보기술공학부 조교수/부교수
※관심분야 : 컴퓨터비전, 영상처리 및 해석, 패턴인식, 신경망