Biometric Authentication using Keystroke Dynamics: A Survey

Kavya Puvirajasingam[1] and D.Sangeetha[2]

[1]ME, Department of Computer Science and Engineering, Kongunadu College of Engineering and Technology, Trichy

[2]AP, Department of Computer Science and Engineering, Kongunadu College of Engineering and Technology, Trichy

Abstract

Keystroke dynamics is a behavioral biometric that is used to provide authentication during user working with the computer system. Besides the traditional way of providing authentication through password, which is the static authentication, there is a new method called dynamic authentication which recognizes the user past login. This paper is a survey about various techniques and algorithms used for providing dynamic authentication.

Keywords**—***Keystroke dynamics, biometric, genetic algorithm, ACO, BPNN, FAR, FRR, EER.*

1. INTRODUCTION

Biometrics or biometry is the term that has been used since early 20th century. It is a new door opened to provide security to the data. Biometrics was developed to meet the statistical and mathematical methods for the analysis of data. In the emerging trend based on biometrics the research is contributed to the fields such as bio-medicine, agriculture, environmental studies, etc.

1.1 Need for biometrics

Biometric is used in the field of computer science to provide authentication to access the data in a system. The traditional way of accessing the data is to provide login id and password. With biometrics the character of the password are no more just alphabets, numbers and special characters. Rather the password is the user itself. The identification of the user is done by this characteristics and traits.

1.2 The Biometric system

The two main categorization for biometrics for analysis and providing security are Static and Dynamic biometrics. In the static biometrics, the user identification is stores in a sensor and every time the user login's then it is identified by scanning and matching the scanned data with attribute stored in the database. In dynamic biometrics, the authentication of the user is not just the one-time recognition. The user is identified and checked even after the user log ins. The authentication is collected from various usages of the system at different timing. This dynamic biometric authentication can be coupled with one-time definition of user.

In static biometric, the changes gas to be done or re-entered for a period of time static biometric has expiration constraints. This leads to the birth of dynamic biometric which leads to automated periodic updating.

Some biometric is uniquely used in the computer system to identify the user are fingerprints, palm print, voice print, facial characteristics, retina pattern, infrared signatures, keyboard typing speeds or penmanship.

1.3 Biometric –A new Generation Security

Biometrics has been used in India in the recent years. The universal identification (UID) program provides unique number by identifying every loyal citizens of India. With these data collected from more than 1.15 billion citizens. The Government analysis the data and issues an unique identity card called AADHAR CARD. This program is administered by the Unique Identification Authority of India (UIdAI). This authority aims at the following: using multiple biometrics, support provided by public and private sectors, competitive standard based procurements, card less design. It also deals with compliance such a s unclear jurisdiction, open tech vs. proprietary system and foreign providers.

1.4 Functions of Biometric System

Authentication and identification are two categories based on which the biometric system works. Authentication is a process that is done to check whether someone is the exact person who can claim access rights. This authentication process incorporates factors such as 1. Something you know,2. Something you own,3. Something you have. Something you know is the password, PIN no, etc. something you own is the smartcard, ATM card, etc. something you are the fingerprint, iris scan, etc. the authentication can be provided by either of these factors or combination of these factors. The other way identification is the process that associates the person identity with the static back store, i.e., the database. In this case the system can grasp the identity of the user and match this with pre-defines data or to unknown data, which can be dynamically updated. Researchers work on this shield, which are blooming in recent years. One of such committee is the biometric consortium that is supported by NIST that is focused on biometric technologies for defense, homeland security, identity management and e-commerce. Other such organizations are European Association for Biometrics.

1.5 Types of biometrics

1.5.1 PHYSICAL

Biometrics is measured on human factors and activities. Based on this it is categorized into physical biometric and behavioral biometric. Physical biometric are authentication based on physical attributes such as fingerprint, iris recognition, etc. physical biometric proves its uniqueness, but cannot be claimed to be theft-proof. For example, if a user's polo camera picture can be misused to authenticate unknown hacker to login like a right user. Hence physical biometric uses additional hardware requirement. Fingerprints of a person can be easily available from the dwelling place or working environment. This can be misused by the untrusted user. Though physical biometric is unique, and reduces risk of remembering long passwords for different devices, it is no more secured when it is theft.

1.5.2 BEHAVIORAL

The other side of biometrics is the behavioral biometric. Behavioral biometric relies on motor skills of the user to accomplish verification motor skill relates to the motion of muscles. Muscle movements are controlled by the functioning of brain, skeleton, joints, nervous system and so on. Behavioral biometric is also called as kinetics. In behavioral biometric, the measurement of physical attributes are not used anymore. Instead how these physical moves and works are considered for authentication and verification. This paper explains about the biometric system used for authentication purpose, with keystroke dynamics as background. Keystroke is the rhythm or movement motion of the user, using the keyboard. Keystroke is simply the typing style and typing speed of a person using a keyboard. Keystroke dynamics is a behavioral biometric, that is used to authenticate users on both pre-login and post-login. This papers is about the authentication and verification across single and multiple applications.

2. MEASURING FACTORS

2.1 Latency Measures

In the research field of keystroke dynamics some measurement criteria has to be followed. The modes the key pressed and released are considered as latency of the keystroke data. Some kinds of key pressing modes are press-to-press (PP), press-to-release(PR), release-to-press(RP), release-to-release(RR) [1]. The literature explains the representation of the latency of keystroke by digraph which is the difference between the two presses. In other words it is the time difference between the first key pressed and the second key pressed. The other type of representation is the trigraph. In other words it is the press and the release of the two consecutive keys or it can be called as the time between the press of first key and the release of the second key. [2]

2.2 Timing Information

The other timing information of the keystroke dynamics are, dwell time or hold time, that defines the pressure of a key pressed. It gives the amount of time a particular key is being pressed. The kind of information is the flight time that is the pressure of key when it is released. It takes the note of RP latency. The researchers extend the feature up to n-graphs for authentication purpose.

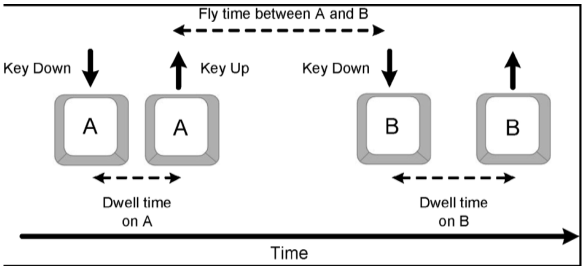


Fig.2.1 Keystroke features and measurements

2.3 Error rate measurement

A decision rule, which depends on a static threshold value, decides whether to accept or reject the user into the system. During such matching, some errors may occur. There are two types of such errors namely False Match Rate (FMR), in which the imposters are wrongly accepted to the system. When a system accepts two or more different users as the right person. The other error rate is the False Non-Match Rate (FNMR), in which the authorized or right person is rejected by the system. This happens when the data above the right person is taken from more than one application is not correlated. The system mistakes that the sample doesn't belong to that right person. In a biometric system, there are possibilities for other kind of errors such as Failure to Enroll Error (FER), which arises when captured sample is not properly enrolled into the system. The next is the failure to Capture Rate (FCR), which occurs when the system tales accidentally pressed key during initial collection of sample. [3]

2.4 Graphical representation

ROC (Receiver Operating Characteristics) or DET(Decision Error Tradeoff). These curves can be used to show the performance at the level of threshold. The curve plots true positives (TP), that is (1-FNMR) and false negatives (FN), which is FNMR. To provide good authentication, low FMR is required to reject the imposters at the maximum.

FNR= no.of.accepted imposters attempts/total no. of imposter attempts

FNMR= no of rejected legitimate users/ total no of legitimate users

The graph represented by the ROC curve gives the clear picture about right attenuation to provide security to the user.

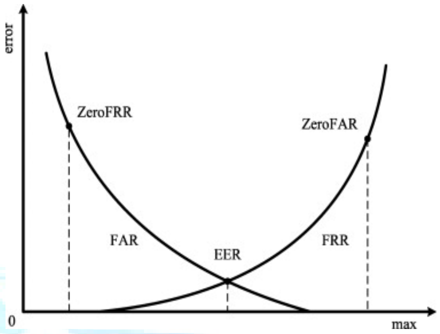


Fig. 2.2 ROC curve for FAR, FRR and EER

3. PHASES

3.1 Data Acquisition

The main approaches in implementing keystroke dynamics are to collect all the necessary data for evaluation. The data are collected from the number of users in their routine working environment which is termed as data from "uncontrolled environment: that relates to dynamic collection of data. From the collected data such as a) latency time, b) dwell time, c) up-to-up time, the data set is places on the research bed.

The two steps in data acquisition are i) data collection, ii) data analysis [4] based on data collected, data analysis is done. In the first stage, a template has to be creates for the users to work. The template must be of application specific. The interesting features are extracted from each application. These features have to be stored in the database. In the second stage, a new version is created. The user keyboard using rhythm is filtered to collect the features. These features are matched with the database that contains the extracted features done before. A matching is performed to analyze whether the matching is performed to analyze whether the matching is succeeded. During then, we do calculations with FMR and FNMR against certain threshold value. Since we are choosing the interesting features against several applications, the process may be of trial and error method of calculating the timing information over the key pressed and key released.

Analysis of data is done in two stages namely, i) statistical analysis, ii) classification of data. The two consecutive steps are the data mining process done to analyze the data and classify them. It is enough if the statistical accuracy is obtained, but in order to generate machine learning approach and more accurate result algorithms such as neural network and perceptron is preferred.

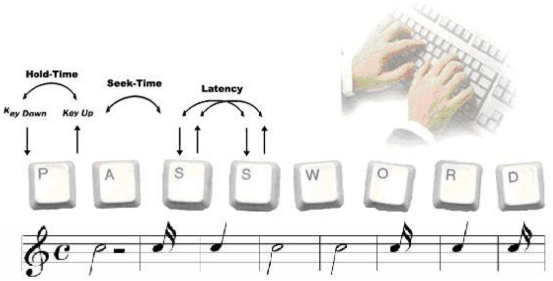


Fig.2.3 Measurements based on key usage

3.2 Template constructions

With the results of the data analysis a template is constructed for each user working on different applications. From the output received we need to generate a template. Before creating a template, the data from the data analysis has to be 'preprocessed'. Data preprocessing is a data mining technique that is used to get better accurate results. The template is developed to finalize that this template is similar to that stored in the database. Concentration on the template designing of the participants is the critical data. Care has to be taken to check, that nothing goes wrong in template construction for a user.

The template creation has been done in two types:

1. General template

2. Personalized template

In general template, the number of features is going to be same for all the user, but not the entries. In personalized data, there would be smaller variations in features from one user to another. The user may have varied keystroke style for every application. Hence personalized template is required additional to generate template. This personalized template is considered as unique template.

3.3 Verification

3.3.1 Realizing captured data

Several attributes are collected regarding the keystroke timing information. The data related to key pressed, key released, pressure of a particular key, time difference among two consecutively pressed(and released) keys. With these data FAR, FRR, EER is determined by working with these data sets in neural network authenticators.

Some of the attributes for realizing the captured data are digraphs, trigraphs, totals username time, total password time total entry time, speed, scan code, edit distance. All related time information data about the keystroke are recorded in nanosecond with 1ns accuracy rate.

3.3.2 Comparison with biometrics

With the available attributes worked on the neural network classification an output is received. This obtained result is matched with the statically stored dynamic keystroke template. If the matches confirms then the right person is continued to work on the system over an application, on the access is timed-out, denied or restricted from the user to use further. This is for what the entire project is concentrating on. Of this verification succeeds,- then the dynamic authentication of users over various applications in a computer system is achieved.

4. RELATED WORK

4.1 Timing vector based user verification

When a user types a password on the keyboard, the typing dynamics or timing pattern are measured. Timing vectors is the duration of keystroke time interleaved with keystroke interval time. For a password with 'n' number of characters it has 'n' number of keystroke duration time and 'n-1' keystroke interval time. The sum of these two gives the (n+ (n-1))- dimensional timing vector. The time unit is calculated in millisecond when the next key is pressed before the release of previous key, then the negative time interval is recorded. It is based on the belief that every individual has characteristics and distinctive typing dynamics. A pattern classifier is built to distinguish and identify the right user. to provide a good protection security to the system, the combination of simple password scheme along with pattern classifier is used in spite of negligible increase in cost and processing time.

Result:

A password of 7 characters, results in timing vector of 15-dim, since a strike of enter key is also pressed. Example timing vector is 120,60,120,90,120,60,150,-

60,120,-30,120,-60,120,120,90,60,150. Where each element was measures in ms. Total of 25 subjects were asked to enter with a new password. For longer password, more number of input and output layers is required in neural network. A 75 vector set of timing information are taken for training set and remaining vectors was taken to train the system. With 10% as threshold acceptance rate, the result was obtained. 4% error rate and 1% of average error rate.

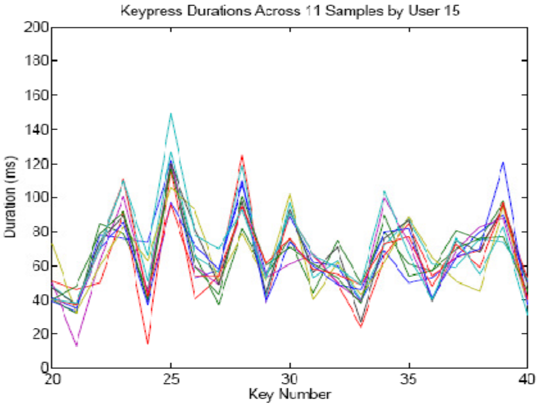


Fig.4.1 keystroke duration across 11 samples user

4.2 Identity authorization based on keystroke latencies

The training set above the keystroke information is collected at different times. Users are allowed to participate under unsupervised conditions. The reference profile collected were representation of n-dimensional feature vector. The data sets were separated into learning and testing sets. These datasets were fed into different classifier techniques such as Euclidean distance, Non-weighed Probability measures, and Weighed Probability measures.

4.2.1 Euclidean Distance measure

Similarity can be calculated on pattern vectors using Euclidean distance. Let R=[r1,r2,r3…,rn] and U=[u1,u2,u3,…,un] now the Euclidean distance between the two n- dimensional vector U and R is given by

$$D(R,U) = \left[\sum_{i=1}^{N}(r_i - u_i)^2\right]^{1/2}$$

For unknown U, pairwise Euclidean distance is calculated.

4.2.2 Non-weighed probability

Along with the n- dimensional pattern vector R & U, the additional quadrupts components such as mean, SD, no.of.occurances and data value of $i^{th}$are considered. The score between the reference profiles is calculated by

$$Score(R,U) = \sum_{i=1}^{N} \mathcal{S}_{v_i}$$

where

$$\mathcal{S}_{u_i} = \frac{1}{o_{u_i}}\left[\sum_{j=1}^{o_{v_i}} Prob\left(\frac{X_{ij}^{(u)} - \mu_{r_i}}{\sigma_{r_i}}\right)\right]$$

and $X_{ij}^{(u)}$ is the j^{th} occurrence of the i^{th} feature of U.

4.2.3 Weighed probability measures

The larger sample set with high frequency in written language are measured, example er, th, sm, et. The score between R&U is calculated as

$$Score(R,U) = \sum_{i=1}^{N}\left(\mathcal{S}_{u_i} * w_{u_i}\right)$$

Result:

The dataset was collected from 63 users. The correct identification rate was 87.18%. The performance of Euclidean distance is 83.22%. The non-weight scoring approach was 85.63%. When examined using Bayesian classifier, it was approximated up to 92.14%, which was almost 5% over the weighed classifier. [6]

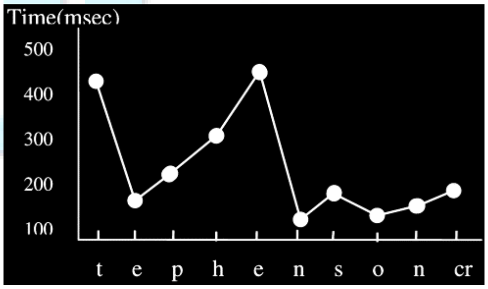


Fig 4.2 Result graph for word Stephenson

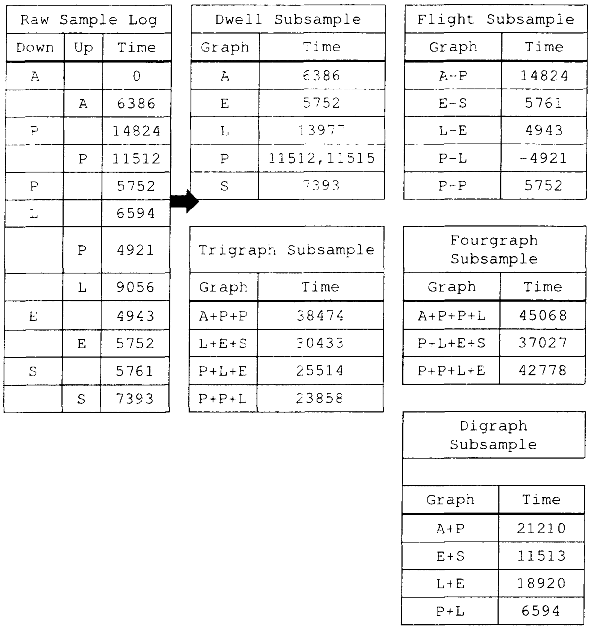
4.3 Trigraph features used for identification of keystroke dynamics

The three conseqitive keys typed are called trigraphs. Trigraph duration is the time between the 1st key pressed and the 3rd key released. For example if the user types indie, then the sequence of trigraphs and duration (msec).

S1: Ind: 277, ndi: 255, dia: 297, ia + enter key: 326. Now the feature vector is sorted in ascending order. The vector is measured for various vectors S2, S3, etc. with varied timing information. Then the distance is calculated and the value is normalized.

Results:

The genuine users and 110 imposters are made to enter the text. 5 samples of genuine users are taken. The experiment was made up to 350 different trigraphs. Let the template for user A is [A1, A2, A3]. The interclass variability of user A is determined using the vectors. The distance is found between the two vectors and the normalized value is obtained, which is between 0 and 1.



Raw Sample Log		
Down	Up	Time
A		0
	A	6386
P		14824
	P	11512
P		5752
L		6594
	P	4921
	L	9056
E		4943
	E	5752
S		5761
	S	7393

Dwell Subsample	
Graph	Time
A	6386
E	5752
L	13977
P	11512,11515
S	7393

Flight Subsample	
Graph	Time
A-P	14824
E-S	5761
L-E	4943
P-L	-4921
P-P	5752

Trigraph Subsample	
Graph	Time
A+P+P	38474
L+E+S	30433
P+L+E	25514
P+P+L	23858

Fourgraph Subsample	
Graph	Time
A+P+P+L	45068
P+L+E+S	37027
P+P+L+E	42778

Digraph Subsample	
Graph	Time
A+P	21210
E+S	11513
L+E	18920
P+L	6594

Table 4.1.Trigraph features

4.4 Using Ant Colony Optimization for feature subset selection

It is essential to select the optimized feature from the obtained sample set. Tough there are lot of feature subset selection such as genetic algorithm, artificial intelligence, pattern recognition, neural network, nearest neighbor algorithm, greedy attribute selection, hill climbing algorithm. One such optimization technique used to select the feature sunset is the Ant Colony Optimization. The various steps followed in ACO are,

Step 1: Get the feature values from duration, latency, digraphs of keystroke.

Step2: Calculate the fitness function

Step 3: initialize the no.ofiterations,no.of ants, initial pheromone values and rate of pheromone evaporation.

Step 4: calculate the local and global optimization value.

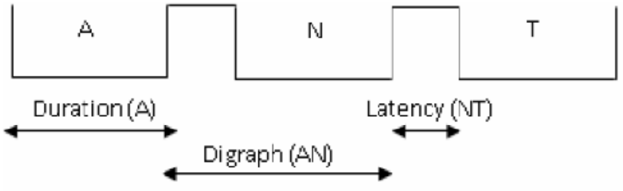
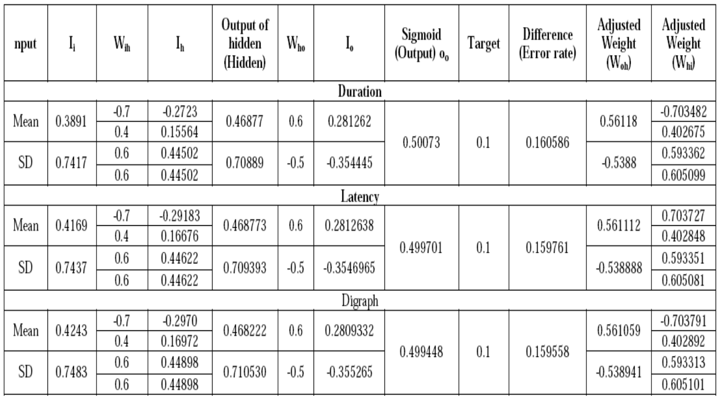


Fig.4.3 Duration,Digraph and Latency

Result:

The fitness value for the calculated duration is 0.425, local minimum for duration is 0.41689, local pheromone update for duration is 0.001. the global duration was calculated as 0.4168 and the global pheromone updating was 0.00225. the remaining ant pheromone update for duration was 0.0001. ant colony optimization can be verified by comparing with BPNN algorithm. The classification error was 0.059% and accuracy was nearly 92.2%. [9]

Table 4.2. Results for duration, digraph and latency



nput	I_i	W_{ih}	I_h	Output of hidden (Hidden)	W_{ho}	I_o	Sigmoid (Output) o_o	Target	Difference (Error rate)	Adjusted Weight (W_{oh})	Adjusted Weight (W_{hi})
Duration											
Mean	0.3891	-0.7	-0.2723	0.46877	0.6	0.281262	0.50073	0.1	0.160586	0.56118	-0.703482
		0.4	0.15564								0.402675
SD	0.7417	0.6	0.44502	0.70889	-0.5	-0.354445				-0.5388	0.593362
		0.6	0.44502								0.605099
Latency											
Mean	0.4169	-0.7	-0.29183	0.468773	0.6	0.2812638	0.499701	0.1	0.159761	0.561112	0.703727
		0.4	0.16676								0.402848
SD	0.7437	0.6	0.44622	0.709393	-0.5	-0.3546965				-0.538888	0.593351
		0.6	0.44622								0.605081
Digraph											
Mean	0.4243	-0.7	-0.2970	0.468222	0.6	0.2809332	0.499448	0.1	0.159558	0.561059	-0.703791
		0.4	0.16972								0.402892
SD	0.7483	0.6	0.44898	0.710530	-0.5	-0.355265				-0.538941	0.593313
		0.6	0.44898								0.605101

5. ALGORITHM THAT WORKS/ APPROACHES

Once the feature extraction process is done, then the templates are created. The users are classified based on the similarity measures. There are some statistical algorithms that are used to classify the users. Sometimes the combinatorial algorithm can also be used.

5.1 Statistical algorithm

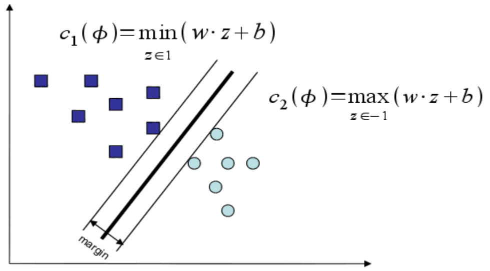
To compute mean and SD of the features. Then the values are compared against the threshold. The comparisons can be done by using hypothesis test, T-test, distance measures such as absolute distance, Euclidean distanced, Manhattan distance, etc. since keystroke dynamics is continuous authentication and non-linear in nature. It is not appreciable to use the linear, statistical methods to compute the features. Moreover, training the datasets is not encouraged by the statistical method to great extent. Hence there is a necessity for more appropriate approaches.

5.2 Neural networks

Neural network is also called Artificial Neural Network (ANN). It is a non-linear statistical data modeling tool. There are basically two different ways of learning the training data sets. They are supervised learning and unsupervised learning. The most popular supervised learning is back propagation [10]. The other supervised learning algorithm examples are train a decision tree, cross validation, neural networks, transduction, ensembles. The popular unsupervised learning are Hopfield Neural Network (HNN). The other unsupervised learning algorithm examples are clustering; dimensionality reduction using PCA, independent component analysis, etc. neural networks is suggested by many researches to give best results. Neural network can handle many parameters. Due to the black box feature of neural network, it is considered as a problem during continuous keystroke authentication [11]

5.3 Support vector Machine Algorithm

Keystroke dynamics concentrates on identifying the correct users. On the other side imposters are also identified. Support vector machine (SVM) is one such algorithm used to detect the imposters. It is considered as consistent and low complexity algorithm. The approach is carried out in two ways, 1. One class svm (OC-SVM), 2. Two class svm (TC-SVM). Ocscm is used to capture data with probable values. Tcsvm provides training data with overall coverage of objects. (i.e. imposters). The FAR and FRR calculated using two approaches are compared and best performance is evaluated. [12].


$c_1(\phi)=\min_{z\in 1}(w\cdot z+b)$
$c_2(\phi)=\max_{z\in -1}(w\cdot z+b)$
margin

$$c_0=\frac{c_1(\phi_0)-c_2(\phi_0)}{2}$$

Fig.5.1 SVM vector used for classification of users

5.4 Back propagation neural network

Back propagation neural network has a forward pass and a backward pass. The features extracted are fed as input to the input layers. It propagates to a value to the hidden layer. The values generated by the hidden layer are fed as input to the output layer. The output layer in turn calculates the output value for the given inputs since the weight are random values sometimes the output vector is not related. Hence there requires a backward pass. This is achieved by back propagation. The steps involved here are 1. To compute error in the output layer, 2.To compute error in hidden layer, 3. Adjust the weight values to improve the performance, 4. Sum up the total error [13]

5.5 Genetic algorithm

Genetic algorithm is a class of probability optimization algorithm, inspired by the biological evaluation process. It uses the concept of natural selection and genetic inheritance (Darwin 1859). It was originally developed by John Holland (1975). The feature extractions are considered as population. There are several steps involved sequentially after the population is selected.

1. The populations are ranked according to their fitness.
2. The population is made to reproduce by two steps such as crossover and mutation. It is based on the concept of 'a pair of parents produces two children'.
3. The steps are repeated until the desired fitness level is reached

5.5.1 PSEUDOCODE:

```
Simple genetic algorithm
Produce initial population of intervals
Evaluate the fitness of all individuals
While (termination not met)
{
Do
{
Select fittest individual for reproduction;
Recombine (i.e. crossover individuals;
Mutation individuals;
Evaluated the fittest modified individual;

Generate new population;
}
}
End while [14]
```

Genetic algorithm can be preceded in travelling salesman problem (TSP). The travelling salesman must visit every city at least once and return to the starting point at the minimum total cost of the entire travel. The TSP can be approximately a genetic algorithm. The advantage of GA is

1. It can solve every optimization problem
2. It is easy to understand and simulate
   The disadvantage is if the fitness value is poorly functioned, then the optimization will be at risk.

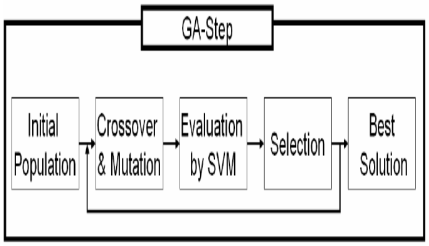


Fig.5.2 Crossover and mutation process

5.6 Ant colony optimization used to solve TSP

Ant colony optimization was introduced by Marco Dorigo in Italy in the year 1992 in his doctoral thesis. It is used to solve TSP ants go through the food by laying down the pheromone traits. The shortest path is found via pheromone traits.

1. The ant move in random
2. After some time, ants follow the traits which have more amount of pheromone.
3. Meanwhile, all the ants will follow the pheromone traits
4. The previous path is evaporated.
   Each ant located in city I has to move to city j. d (I,j) is the attractiveness, which is the function that gives the inverse of cost. T(I,j) is the trait level, detecting the amount of pheromone trait. The set of cities not visited by the ant k in city I is $T_k(i)$. the probability that ant k $P_k(I,j)$ in city i will go to city j, is calculated.

5.6.1 PSEUDOCODE:for general ant colony

```
Initialize the base attractiveness $\tau$ for each edge
For (each ant) do
P( choose the edge)
Add and move to table list of each ant
Repeat until each ant complete solution
End;
For (each ant that completes a soln)
Update $\tau$ for each edge the ant traversed
End;
If (local better than global) save local
End;
End;
[15]
```

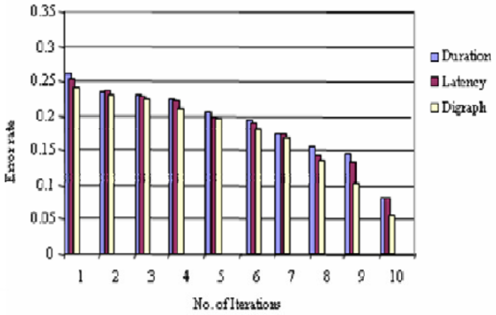


Fig.5.3 Results based on ACO and BPNN

The benefit of ACO is that it can solve NP-Hard problem in short time.It balances the previous solution and new exploring solution. Optimal solution is obtained.

The limitations of ACO are the coding differs for different applications. Ineffective utilization of previous solution affects global solution

REFERENCES

[1] Biometric authentication and identification using keystroke dynamics

[2] Bengadano keystroke dynamics

[3] Hafez Barghoutti,keystroke dynamics, how typing differ from one application to another.

[4] Hafez Barghoutti, keystroke dynamics, chapter 4, choice of method

[5] Sungzooncho, chigeunhan,daeheehan,'Web based keystroke dynamics identity verification using neural network', journal of organizational computing and electronic commerce,vol 10,no.40 pp 295-307,2000

[6] Fabian Monrose and Aviel di rubin, AT&T Laboratory Research. Florham Park, N.J,'keystroke dynamics as a biometric for authentication'.

[7] 'Biometric Authentication based on keystroke dynamics',http://biometrics.cse.msu.edu

[8] citeseerx.ist.psu.

[9] Marcus akrnan,M.Akila,'Personal Authentication based on keystroke dynamics using ACO & BPNN','JCNS.Vol-no.2,2009,Nov

[10] Waffles. Source forge.net/tutorial/supervised-examples.html

[11] Salil.P.banerjee, damon.L.woodard,'biometric authentication & identification using keystroke dynamics', Journal of pattern recognition research 7, 2012(116-169)

[12] Yingpengsang,hongs hen, pingzhi fan, 'Novel imposters detection in keystroke dynamics by svm'.

[13] – [9]

[14] Assafzaritsky, ben gurion university, Israel,' Introduction to genetic algorithm.