ACO-Random Forest Approach to Protect the Kids from Internet Threats through Keystroke

Soumen Roy[#1], Utpal Roy[*2], D. D. Sinha[#3]
[#1,3]Department of Computer Science and Engineering,
University of Calcutta, 92 APC Road, Calcutta -700 009, INDIA.
[*2]Department of Computer & System Sciences, Visva-Bharati, Santiniketan -731235, INDIA
[1]soumen.roy_2007@yahoo.co.in
[2]roy.utpal@gmail.com
[3]devadatta.sinha@gmail.com

Abstract—**Internet users from children group are rapidly increasing. They use the Internet for doing their homework to keep in touch with their friends. But they are vulnerable to unknown threats coming from the Internet. Many Government authorities are actively trying to protect the children from these threats. This study is one approach which can distinguish the children from Internet users by analysing the typing behaviour. The moment a user is identified to be a child or minor, the next stage of protection will be auto sensing firewall appropriate for the users. We have taken two public datasets on keystroke dynamics for experimental purpose and applied Ant Colony Optimization (ACO) technique as search methods and Random Forest as a classifier on each dataset. Obtained results are impressive. As per our study, more than 92% of desktop computer users and 84.22% of touch screen mobile users from children group can be protected from the looming threats from the Internet by analysing the typing behaviour on keyboard or touch screen.**

Keyword - Keystroke Dynamics, Computer Security, Machine Learning, ACO, Child Protection

I. INTRODUCTION

The numbers of users from the age group below 18 are rapidly increasing as the demand of homework and to stay in touch with others necessitates the use of technology. According to a recent survey by McAfee (a Security Technology Firm), more than 62% of children shared their personal information and 39% of their parents were unaware of it and 71% of youth secrete their online activities from parents and 56% of parents are uninformed of it [1]. Another survey in India reveals that 67% of the children under age 10 had Facebook account and 82% of them received inappropriate messages [2]. Many Government authorities are actively trying to protect the children from these types of unknown threats coming from the Internet.

This study is one approach to protect the children from unknown threats coming from the Internet. The science behind this approach is children's physical structure, mentality, knowledge level, experience level on keyboard, neurophysiological, neuropsychological factors, reading style, keyboard position reflect the typing pattern on keyboard or touch screen, which discriminate the children from adults.

Typing pattern is a behavioral biometric characteristics much like our written signature or voice print, by which people can be identified [5-8]. Nevertheless, being non-intrusive and cost effective, this method is now popular field of research. It is totally software based system which can be easily integrated in any existing system with small alternation.

The dataset is collected by Uzun et al. [3] in the year 2014, they have collected the typing pattern samples from 51 children (age below 18) and 49 adults of 100 subjects in one session with 5 repetition for two type of text patterns (".tie5Roanl" and "MercanOtu") through desktop computer keyboard. Second dataset is collected by Abed et. al. [4] in the same year, they have collected the typing pattern samples from 11 children (age below 19) and 40 adults of 51 subjects in three different sessions with the minimum time period of 3 to 30 days separating each session with 15 to 20 repetitions for one text pattern ("rhu.university") through touch screen mobile device, Nokia Lumia 920 (4.5" Multi-Touch, 768×1280 (332ppi), Weight: 185g).

Our objective and contribution of this paper are listed below:

- Provide a novel approach to identify the children group through typing pattern on desktop and touch screen phone.
- Discuss the appropriate area of application where this technique can be fit.
- Comparative analysis of different learning methods, environments in age group identification on different datasets.

DOI: 10.21817/ijet/2017/v9i3/170903S044

II. Related Works

Keystroke dynamics is not new in biometric science. The technique has been started in the year 1980. Many Journal, Conference articles and master thesis have been published. Fig. 1 clearly indicates the increasing trends on keystroke dynamics research. Many datasets have been created considering different type of texts with different lengths from different number of subjects, many methods have been applied and many innovative ideas have been come out from the previous study. But most of the papers focused on user identification or authentication performance through typing pattern.

Only few papers described some ancillary information that can be extracted from the typing pattern. Epp et al. [10] show that it is possible to identify the emotional state of the person through the person's way of typing. They reported the accuracy rate 84% to identify the angriness and excitement. Giot et al. [12] show that it is possible to detect the gender and they reported the accuracy rate more than 90% using typing style. Idrus et al. [11] show that it is possible to identify the gender, age group, handedness and one or two hands used while typing and they reported the accuracy rate very close to 90%. Uzun et al. [3] show that it is possible to identify the child group and adults through typing pattern and they obtained the accuracy more than 90% for the simple familiar Turkish text. They have used 13 classification algorithms where SVM (Linear) is achieved minimum Equal Error Rate for familiar text but the performance is not consistent for the other texts.

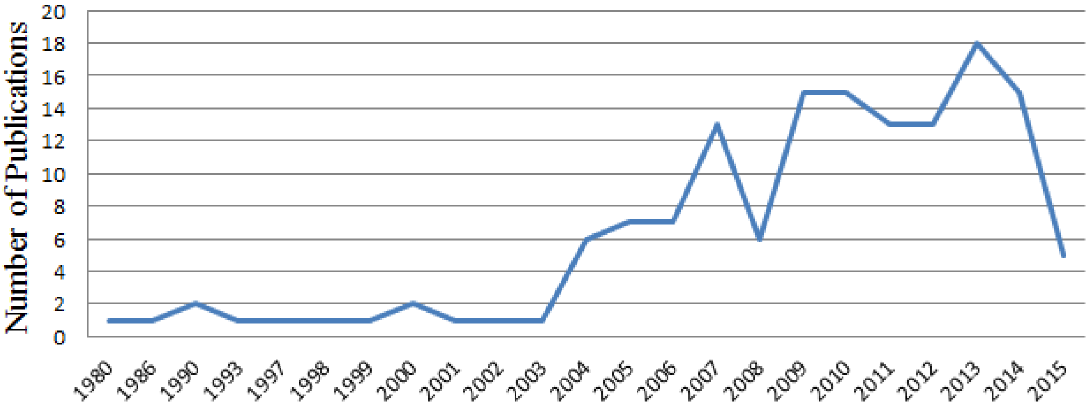

Fig. 1. Published articles on keystroke dynamics by year as per our knowledge

III.Keystroke Dynamics

A. Basic Idea

Keystroke dynamics is a behavioral biometric traits relates the issues in human authentication/identification. But this technique also can be used to recognize the ancillary information. Physical structure, mentality, reading style, hand geometry, weight and length, experience level on keyboard, knowledge level, educational qualification an neuro-physiological are the factors which indirectly effect on keyboard while typing to identify the kids. Since keystroke dynamics is a distance-based measurable pattern it would be the strong alternative which may enable the age group identification.

B. Features

Basic features of keystroke patterns are the time interval between a key pressed and released, the time interval between two subsequent keys pressed and released. Now days, key pressure, finger tips size, finger placement on keyboard and keystroke sound are also considered. The some timing features of the keystroke dynamics are as follow:

$$\text{Key-Duration (KD)}=R_i-P_i \quad (1)$$

$$\text{Up-Up Key Latency (RR)}=R_{i+1}-R_i \quad (2)$$

$$\text{Down-Down Key Latency (PP)}=P_{i+1}-P_i \quad (3)$$

$$\text{Up-Down Key Latency (RP)}=P_{i+1}-R_i \quad (4)$$

$$\text{Down-Up Key Latency (PR)}=R_{i+1}-P_i \quad (5)$$

$$\text{Total-Time Key Latency (T-Time)}=R_n-P_1 \quad (6)$$

$$\text{Tri-graph Latency (Tri-time)}=R_{i+2}-P_i \quad (7)$$

$$\text{Four-graph Latency (F-Time)}=R_{i+3}-P_i \quad (8)$$

Here, P and R represent the key press and release times of entered keys for predefined text.

C. Public Datasets

Many datasets on keystroke dynamics have been created in the last 30 years but some of them listed below are available in the Internet, we can download it or we can download on request. This datasets are collected from both child and adult users. Details are given in the Table 1. We have given the name of each dataset for this paper where Dataset A and B are created through keyboard where Dataset C is created through touch screen.

TABLE I. Publicly availablekeystroke dynamics soft biometric datasets

Dataset Name	**Considered Text**	**Number of Subjects**	**Session**	**Repetition**	**Features**	**Diversity [* C=Child and A=Adult]**
Dataset A [3]	".tie5Roanl"	100	1	500	KD, PP, RP	*C=51, *A=49
Dataset B [3]	"MercanOtu"	100	1	500	KD, PP, RP	*C=51, *A=49
Dataset C [4]	"rhu.university"	51	3	15-20	KD, PP, RP, RR	*C=11, *A=40

IV. Experimental Results

Details of the experimental results are described in the Table 2, Table 3 and Table 4. Eight popular and recognized classification algorithms were used on each dataset described in Table 1. The accuracy rate is calculated by the weka environment version Weka 3.7.2 [9]. Two test options were used in our experiment. First one is 10 fold cross validation where total sample of instances is divided into 10 groups, each group will be treated as testing and remaining training groups will be treated as training. In second test option we have divided the total training data into 2 groups with 66% of training and 34% of testing instances. Only test accuracy of each learning processes were listed. The table shows that Random Forest methods achieved highest accuracy before and after optimization with ACO technique consistently for each dataset.

TABLE III. Recorded accuracy before and after optimization on [1]dataset A

Classifiers	**Accuracy before optimization**		**Accuracy after optimization by ACO**	
	10 fold cross validation	**66% of training**	**10 fold cross validation**	**66% of training**
Random Forest [14]	90.4	88.82	92.2	88.82
Fuzzy Rough NN [15]	91.8	86.47	90	86.47
Fuzzy NN [15]	91.6	85.88	91.8	85.88
SVM(Linear) [13]	63.2	73.53	51	71.76
MLP [16]	90.2	87.06	92	87.06
Naïve Bayes [17]	85.6	81.18	88.8	81.18
K- NN [18]	90	82.35	88	82.35
J48 [19]	89.8	84.71	90.2	84.71

TABLE IIIII. Recorded accuracy before and after optimization on [1]dataset B

Classifiers	**Accuracy before optimization**		**Accuracy after optimization by ACO**	
	10 fold cross validation	**66% of training**	**10 fold cross validation**	**66% of training**
Random Forest [14]	90.2	90.58	87.8	86.47
Fuzzy Rough NN [15]	90.4	89.41	83.4	86.47
Fuzzy NN [15]	88.4	89.41	89	89.41
SVM(Linear) [13]	63.6	86.47	64.4	67.06
MLP [16]	87.2	84.71	88	82.35
Naïve Bayes [17]	83.4	85.88	85.8	85.29
K- NN [18]	88.8	88.24	84.2	87.06
J48 [19]	85.6	88.24	84.4	88.24

TABLE IVV. Recorded accuracy before and after optimization on ^{2}dataset C

Classifiers	Accuracy before optimization		Accuracy after optimization by ACO	
	10 fold cross validation	66% of training	10 fold cross validation	66% of training
Random Forest [14]	83.91	81.73	84.22	77.4
Fuzzy Rough NN [15]	87.28	85.14	80.23	69.97
Fuzzy NN [15]	83.18	80.81	83.91	78.02
SVM(Linear) [13]	71.52	71.51	77.71	76.16
MLP [16]	80.76	79.26	77.86	76.16
Naïve Bayes [17]	42.48	44.27	48.89	75.54
K- NN [18]	84.96	81.42	80.96	69.04
J48 [19]	79.18	75.85	81.49	78.02

V. Comparative Analyses

Among the 8 algorithms Random Forest and Fuzzy Rough NN achieved the highest accuracy on desktop and android environments both. We compared the performance by paired T test and results shows in the Table 5 that Fuzzy Rough NN is always better than Random Forest for all datasets used in our experiments, but after optimization we observed that Random Forest is proved the suitable methods in this domain.

TABLE VII. Paired T test results on each dataset at the significant level 0.05

Dataset	SVM (Linear Kernel)	Fuzzy Rough NN	Random Forest
Dataset A [3]	63.20(23.71)	90.40(3.24) V	90.20(3.33) V
Dataset B [3]	69.40(6.75)	87.27(3.84) V	83.91(2.90) V
Dataset C [4]	63.60(28.04)	91.80(6.49) V	90.40(4.79)
	(V/ /*)	(3/0/0)	(2/1/0)

Fig. 2 represents the accuracy rates before and after optimization achieved by different classification algorithms on dataset A. Where Random Forest (RF) algorithm is achieved 92.2% of accuracy after optimization instead of 90.4%.

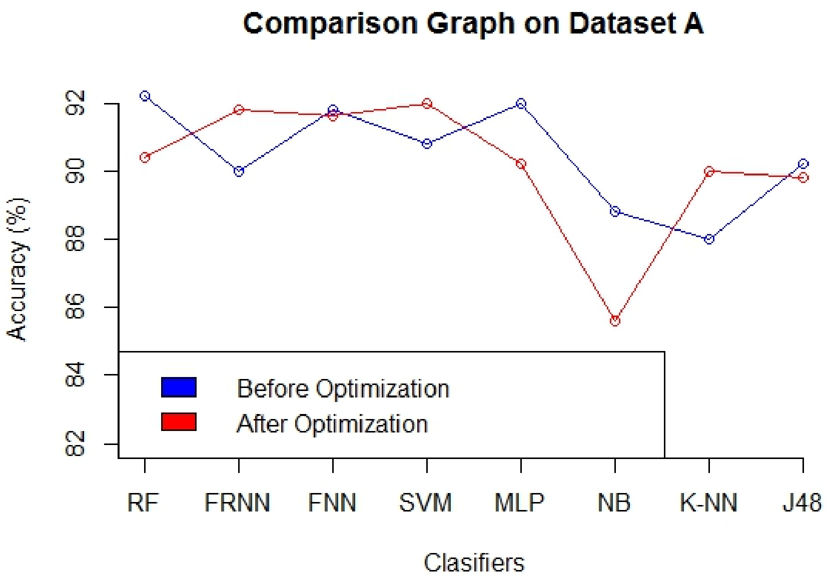

Fig. 2. Accuracy before and after optimization by ACO on Dataset A

Fig. 3 represents the accuracy rate before and after optimization by ACO on dataset B.

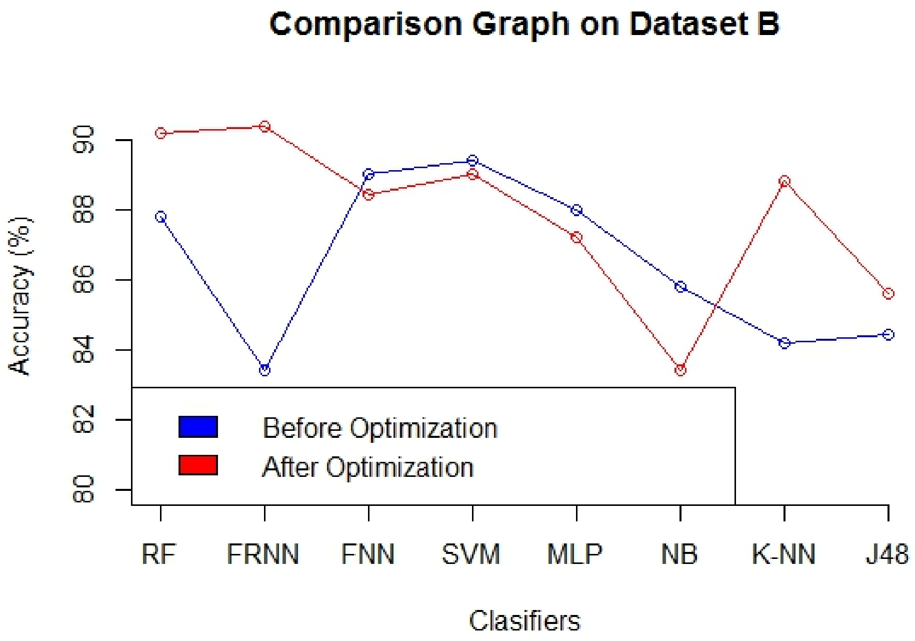

Fig. 3. Accuracy before and after Optimization by ACO on Dataset B

Fig. 4 represents the accuracy rate before and after optimization by ACO on dataset C.

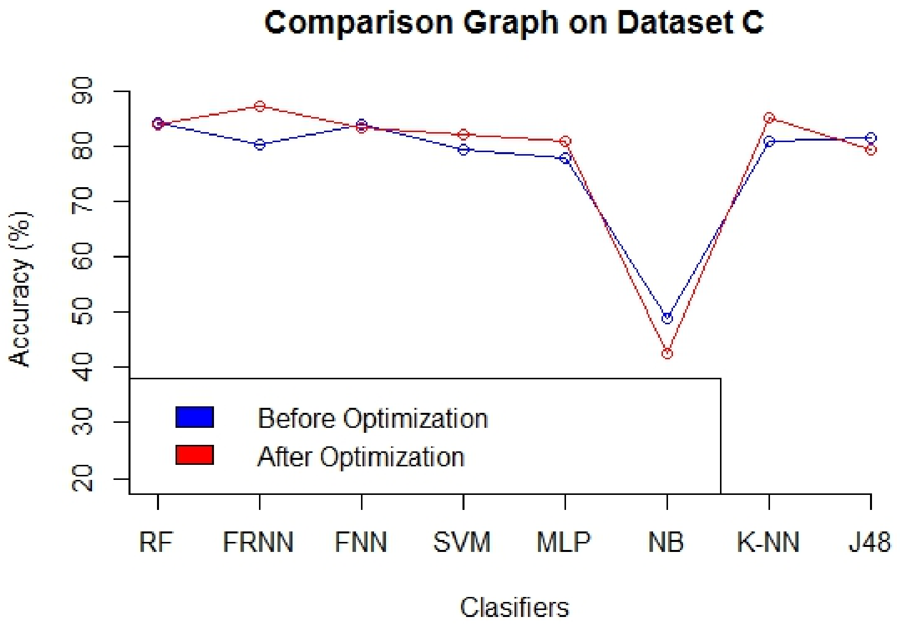

Fig. 4. Accuracy before and after Optimization by ACO on Dataset C

VI. APPROACHES

A. ACO-RF Approach

Our proposed model is ACO-RF. The searched input is the key parameters to check is it kids or not. The moment a user is identified to be a child or minor, the next stage of protection will be auto sensing firewall appropriate for the users and it will be continued whenever user types the search inputs, the graphical representation is presented in the Fig. 5.

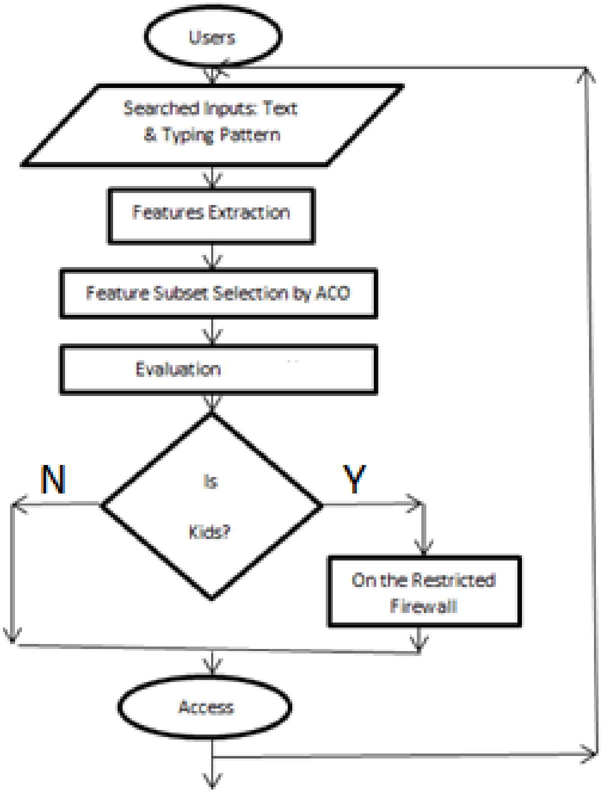
Users
Searched Inputs: Text & Typing Pattern
Features Extraction
Feature Subset Selection by ACO
Evaluation
N
Is Kids?
Y
On the Restricted Firewall
Access

Fig. 5. ACO-RF Model to distinguish the kids from Internet Users

VII. Comparisons with other methods

Bicakciet. al. [3] showed that the accuracy rate to distinguish the children group from adults is 91.2%. This is the optimum accuracy recorded in literature for the simple text in Turkey, where our proposed approach achieved 92.2% of accuracy on same dataset. They also applied their classification algorithms on password type text but only achieved 87.2% of accuracy where our approach achieved 90.2% of accuracy. Therefore, our approach is more consistent that previously proposed methods.

VIII. Discussions

It is true that performance of keystroke dynamics is not much promising due to high failure to enroll rate or intra class variation. So this technique can be applied where this error rates can be compromise instead the use in user identification / authentication. In this paper, we have tried to segregate the children from adults through the way of typing and obtained promising results.

The experiments have been done in both environments. In desktop environment, we achieved 92.2% of accuracy and we achieved 84.22% of accuracy in android environment using ACO-Random Forest. It is very hard to achieve these results in practice where there are more chances to high FTE rate due to external factors like cross device validation.

IX. Conclusions

Keystroke dynamics and mouse movements are two common measurable distance-based activities to use the Internet through keyboard/touch screen. It is enough to identify the age group which can protect the kids or minor from looming threats coming from the Internet. We have collected datasets only contain keystroke pattern and applied 8 machine learning algorithms on each and also we have applied optimization techniques (ACO) to select feature subset. Random Forest machine learning models are proved a suitable classification method, where ACO is achieved optimum solution as optimization technique. Machine learning algorithms were used in our experiment, where we obtained up to 92% of accuracy in desktop environment. This accuracy rate is impressive for single familiar fixed text, if enrolment phase is extremely accurate. But it is very hard to achieve in practice. There are many factors which may affect the process and increases the failure to enroll rate. It means the technology is not much efficient. More research work has to be done and many factors have to be included like mouse dynamic, pressure which is proportional to force, depends on mass of hand weight may be the good factor in desktop environment. In android platform, key pressure, acceleration, and finger tips size may be included where advance sensing device, accelerometer are embedded in each smart phone, So this technique get achieved acceptable accuracy and can be used to protect the children from looming Internet threats.

Adulthood is ascertained by attainment to 18 years of age legally. The knowledge level, IQ and ability may not always follow this suit. Exceptionally there are retarded adults as well highly proficient minors. The treatment in this paper does not discriminate the biological age. But indication is on mental age and efficiency.

REFERENCES

[1] McAfee, “The Digital Divide: How the Online Behavior of Teens is Getting Past Parents”, http://www.mcafee.com/in/resources/misc/digital-divide-study.pdf, June, 2012.
[2] V. Mugdha, “82% children on Facebook get vulgar messages”, Hindustan Times, Mumbai, http://www.hindustantimes.com/mumbai/82-children-on-facebook-get-vulgar-messages/story-0d532SUH4E2kYDN4o1ja8H.html, Feb, 2013.
[3] Bicakci, Yuzun, “Distinguishing Child Users from Adults Using Keystroke Dynamics”, http://bil.etu.edu.tr/bicakci/dagkd/dagkd.htm, 2014.
[4] El-Abed M., Dafer M., El Khayat R., "RHU Keystroke : A Mobile-based Benchmark for Keystroke Dynamics Systems", 48th IEEE International Carnahan Conference on Security Technology (ICCST), Rome, Italy, 2014.
[5] S. Roy, U. Roy, D.D. Sinha, “Password Recovery Mechanism Based on Keystroke Dynamics”, Proceedings of Second International Conference INDIA 2015, Volume 1, Springer India, Advances in Intelligent Systems and Computing, ISSN 2194-5357, pp 245-257.
[6] S. Roy, U. Roy, D.D. Sinha, “Rhythmic Password-based Cryptosystem”, 2nd International Conf. on Computing and System, University of Burdwan, West Bengal, India, 2013, 303-307.
[7] S. Roy, U. Roy, D.D. Sinha, "Performance Perspective of Different Classifiers on Different Keystroke Datasets ", International Journal of New Technologies in Science and Engineering (IJNTSE), Volume-02, Issue-04, Page No (64-73), Oct -2015
[8] S. Roy, U. Roy, D.D. Sinha, " Distance Based Models of Keystroke Dynamics User Authentication ", International Journal of Advanced Engineering Research and Science (IJAERS), Volume-02, Issue-09, Page No (89-94), Sep -2015.
[9] http://www.cs.waikato.ac.nz/ml/weka/
[10] Epp, C., Lippold, M., Mandryk, R.: Identifying emotional states using keystroke dynamics. In: Proceedings of the 2011 annual conference on human factors in computing systems. (2011) 715–724
[11] Idrus S.Z.S., Cherrier E., Rosenberger C., Bours P., Soft Bio-metrics For Keystroke Dynamics. International Conference on Image Analysis and Recognition (ICIAR), 2013, Povoa de Varzim, Portugal. 8 p., 2013.
[12] Giot, R., Rosenberger, C.: A new soft biometric approach for keystroke dynamics based on gender recognition. Int. J. Info. Tech. and Manag., Special Issue on ”Advances and Trends in Biometrics by Dr Lidong Wang 11(1/2) (2012) 35–49
[13] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY, 1995.
[14] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
[15] Jensen, Richard, and Chris Cornelis. "Fuzzy-rough nearest neighbour classification." Transactions on rough sets XIII. Springer Berlin Heidelberg, 2011. 56-72.
[16] Gardner, Matt W., and S. R. Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." Atmospheric environment 32.14 (1998): 2627-2636.
[17] Koch, Karl-Rudolf. Bayes’ Theorem. Springer Berlin Heidelberg, 1990.
[18] Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4.2 (2009): 1883.
[19] Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

AUTHOR PROFILE

Soumen Roy is with University of Calcutta since last 4 years as a research scholar. He is also working with Bagnan College, Bagnan, Howrah, India for last 7 years as a teacher. He has also one year of experience in software development. He has published more than 15 international papers in the form of journal, conference proceeding and book chapter.

Utpal Roy was with School of Technology (SOT), Assam University, Silchar 788011, India. He is now with Visva-Bharati, Santiniketan -731235 acting as a head of the department of Computer and System Science. He has published more than 64 articles in the form of journal and conference proceedings.

Devadatta Sinha has More than thirty eight of experience in the field of Computer Science and Engineering in research and teaching. He worked as faculty member in BIT Mesra, Ranchi, Jadavpur University, and University of Calcutta. He has written a number of research papers in National and International Journals,Conference Proceedings. He has also written a number of expository articles in periodicals, books, monographs. He has research interests include Software Engineering, Parallel and Distributed Algorithms, Bioinformatics, Computational Intelligence, Computer Education, Mathematical Ecology,Networking. He guided Research students for Ph.D, in Computer Science and Engineering and M.Tech, B.Tech and M.Sc students for their dissertation. He was Sectional President, Section of Computer Science, Indian Science Congress Association (1994), Fellow,Computer Society of India.