Defence Science Journal, Vol. 66, No. 6, November 2016, pp. 612-617, DOI : 10.14429/dsj.66.10802

Rough Set-hypergraph-based Feature Selection Approach for Intrusion Detection Systems

M.R. Gauthama Raman[!], K. Kannan[@], S.K. Pal[#], and V. S. Shankar Sriram[!,*]

[!]*Centre for Information Super Highway, School of Computing, SASTRA University, Thanjavur - 613 401, India*
[@]*Discrete Mathematics Research Laboratory, Department of Mathematics, SASTRA University, Thanjavur - 613 401, India*
[#]*Scientific Analysis Group, New Delhi - 110 054, India*
[*]*E-mail: sriram@it.sastra.edu*

ABSTRACT

Immense growth in network-based services had resulted in the upsurge of internet users, security threats and cyber-attacks. Intrusion detection systems (IDSs) have become an essential component of any network architecture, in order to secure an IT infrastructure from the malicious activities of the intruders. An efficient IDS should be able to detect, identify and track the malicious attempts made by the intruders. With many IDSs available in the literature, the most common challenge due to voluminous network traffic patterns is the curse of dimensionality. This scenario emphasises the importance of feature selection algorithm, which can identify the relevant features and ignore the rest without any information loss. A novel rough set κ-Helly property technique (RSKHT) feature selection algorithm had been proposed to identify the key features for network IDSs. Experiments carried using benchmark KDD cup 1999 dataset were found to be promising, when compared with the existing feature selection algorithms with respect to reduct size, classifier's performance and time complexity. RSKHT was found to be computationally attractive and flexible for massive datasets.

Keywords: Intrusion detection systems, rough set theory, hypergraph, feature selection, κ-Helly property

1. INTRODUCTION

Over the past few decades, the advancements in internet applications and communication technology have led to the emergence of a dynamic information society. The phenomenal increase in the number of people connected via Wi-Fi, W iMAX, MANET, 4G mobile networks, etc. has led to the massive generation and exchange of information (sensitive and non-sensitive) over public networks. The importance of network security can be realised in the scenario, where these information needs to be safeguarded against various cyber threats and attacks. Traditional security techniques with respect to data encryption, authentication, firewall, security policies, etc. were proven to be immature since the targeted attacks have become more sophisticated in nature. According to NIST[1], "Intrusion is an attempt to compromise confidentiality, integrity and availability (CIA) or to bypass the security mechanisms of a computer or a network". Intruders tend to manipulate the information by imposing various attacks like phishing attack, password cracking, sniffing, etc. Recent studies show that an average of twenty to forty new vulnerabilities were detected every month and CERT[2] report says that the number of intrusions tend to increase exponentially every year.

This scenario has led to a situation, where an additional layer of security needs to be integrated with the firewall, to strengthen the line of defense against any type of intrusions. According to NIST[1], 'Intrusion detection is the process of monitoring the events occurring in a computer system or network and analysing them for signs of intrusions, defined as attempts to compromise the CIA or to bypass the security mechanisms of a computer or network. Intrusion detection systems (IDSs) are software or hardware products that automate this monitoring and analysis process'. According to Anderson[3], IDSs has been considered as a second line of defense, which plays a significant role in the identification of various kinds of network vulnerabilities. Over the past decades, the design of IDSs has been a major area of research focused by a large variant of researchers. With the existence of a wide range of framework for IDSs, a generic architecture of IDS is shown in Fig. 1, whose major components include information source, analysis engine and decision module.

On the basis of the intrusion detection approach adopted by the analysis engine, IDS can be categorised into signature based detection and anomaly based detection. Signature based detection techniques compare the network traffic patterns against a signature database, which contain the patterns of known intrusions[4]. As the accuracy of detection depends on the signatures, frequent updates to the signature database become essential. Anomaly based detection techniques depend on the

Received : 31 January 2016, Revised : 13 April 2016
Accepted : 18 July 2016, Online published : 31 October 2016

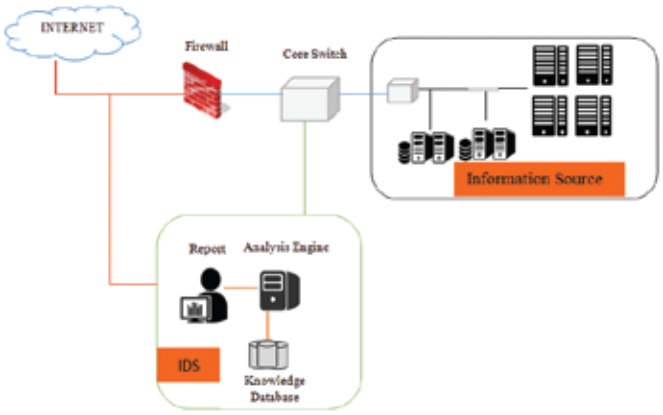

Figure 1. A generic IDS architecture.

baseline (normal user behaviour) designed by the network administrators, such that any traffic patterns that deviate from the configured baseline were considered to be an attack[4]. The former approach detect the attacks which have signatures similar to that found in the signature database, while the latter can identify new or unknown security breaches without any signature database, but possess a high false positive rate in the absence of intellectual baseline profile configuration.

As mentioned above, numerous research works have been carried out in the field of network security with respect to design of IDS using machine learning and statistical analysis techniques, to discover knowledge from massive datasets. Improving the performance of the IDSs has been considered to be a challenging task due to the volatility, incompleteness and redundancy in the voluminous network traffic patterns. This emphasise the importance of feature selection algorithms in IDS, to identify the informative features and over look the irreverent or redundant features that degrade the performance of the IDSs in terms of computational complexity and detection rate. In recent years, rough set theory (RST) has emerged as an intelligent tool for knowledge discovery from imprecise, uncertain datasets through the identification of redacts (feature subset) which represents the maximum information of the system. Finding all the possible redacts using RST is found to be an NP - Hard problem[5]. Many heuristic based approaches modeled to find the near optimal solution failed to guarantee an optimal reduct. Several challenges like the curse of dimensionality, classification accuracy, etc. have insisted researchers to develop a feature selection algorithm which combine the benefits of RST and a generic data representation framework. In this paper, a novel feature selection algorithm based on RST and κ-Helly property of hypergraph had been proposed to identify key features, which improves the efficiency of the IDS. RSKHT obtains the optimal reduct, which enhances the performance of the classifier in terms of accuracy with reduced time complexity.

2. PROPOSED FEATURE SELECTION ALGORITHM

To identify the minimal feature subset which consists of indispensable attributes from the dataset a novel feature selection algorithm (RSKHT) is put forth. In general, dataset consists of attributes, which can be either conditional or decisional. Let us consider an information system (Table 1) which consists of n samples $\{X_1, X_2, ..., X_n\}$ characterised by *m* conditional attributes $\{C_1, C_2, ..., C_m\}$ and a decision attribute D that represent the classes to which each sample belong to. The dimensionality reductions rely on the degree of dependency that exist among the attributes. An attribute (conditional) or group of attributes is found to be irrelevant, if its association degree with other attributes is high or it is unproductive of the decision class. Ignoring such kind of attributes not only enhances the classification accuracy, but also reduces the complexity of the classifier in terms of time and computation. RST, the choice of researchers for the problem of dimensionality reduction, obtain the reducts (minimal set of attributes) with respect to the degree of dependency among the attributes. One of the simplest methods to find the reduct using RST is to generate all possible combinations of the feature subsets and best subset is identified based on the evaluation function (dependency degree). This method had been suitable for the dataset with minimum number of samples or attributes, whereas in the case of voluminous dataset, the computational complexity increases exponentially. To overcome this problem, quick reduct (QR) algorithm[6] was used to find the reduct, which starts with an empty set and iteratively add up the attributes which gives maximum dependency value.

Table 1. Information table

Samples	Conditional attributes						Decision attributes (D)
	C_1	C_2	.	.	.	C_m	
X_1	u_{11}	u_{21}	.	.	.	u_{m1}	C_1
X_2	u_{12}	u_{22}	.	.	.	u_{m2}	C_2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
X_n	u_{1n}	u_{2n}	.	.	.	u_{mn}	C_n

However, the major drawback in QR algorithm is the volatility in the dependency metric due to the addition or removal of attributes from the reduct, which induces complexity in the process of attaining the optimal reduct. In addition, it may trap at local minimum, since the dependency metric will discriminate the attributes of the dataset. To overcome these challenges, a novel feature selection algorithm based on RST and *K*-Helly property of hypergraph had been proposed to find the optimal feature subset. Consider a hypergraph *H* whose vertices represent the conditional attributes $C = \{C_1, C_2, ..., C_m\}$ and the hyper edges represent the reducts obtained through RST. $R = \{R_1, R_2, ..., R_r\}$ The κ-Helly property[7] of hypergraph was recursively applied on the pair wise intersection of the neighborhood hyper edges to construct the optimal feature subset. Thus, the rough set theory hybridised with the κ-Helly property of hypergraph provides a novel solution to the problem of optimal feature subset identification.

To facilitate reader's understandability, the working of RSKHT had been elaborated using the hypergraph representation of the reducts obtained using RST (Fig. 2) and the application of *K*-Helly property on the hyper edges of the

hypergraph (Fig. 3). Let us consider a simple hypergraph H constructed using the information table (Table 1). Each hyper edge of the hypergraph represents the reducts obtained from RST

$$R_1=\{C_2,C_5,C_7\}, R_2=\{C_2,C_3,C_6\}, R_3=\{C_1,C_5,C_7\}, R_4=\{C_1,C_2,C_7\}, R_5=\{C_1,C_6,C_7\},$$
$$. R_6=\{C_3,C_7,C_8\}, R_7=\{C_2,C_6,C_8\}, R_8=\{C_3,C_6,C_7\}, R_9=\{C_3,C_5,C_7\}, R_{10}=\{C_4,C_6,C_7\}$$

The vertices of the hyper edge represent the conditional attributes $C=\{C_1,C_2,C_3,...,C_{12}\}$ contained in the reducts. The recursive application of κ-Helly property on the hyper edges of the hypergraph resulted in the identification of the optimal feature subset. Various stages in the identification of the optimal feature subset based on recursive κ-Helly property of hypergraph is as shown in Fig. 3.

Algorithm: Rough Set based κ-Helly property Input

$C \leftarrow$ Set of conditional attributes

$D \leftarrow$ Set of decisional attributes

Output

$f_s \leftarrow$ Optimal reduct

RSKHT (*c*, *d*)

1. $R \leftarrow \{\ \}$
2. while $\left(\left(\Upsilon_{R(D)}-\Upsilon_{C(D)}\right)\neq 0\right)$ do
3. $T \leftarrow R$
4. for each $Y \in (C-R)$ do
5. calculate $\left|POS_{R\cup\{Y\}}(D)\right|$
6. $\Upsilon_{R\cup\{Y\}(D)} \leftarrow \dfrac{\left|POS_{R\cup\{Y\}}(D)\right|}{|U|}$
7. if $\left(\left(\gamma_{R\cup\{Y\}(D)}\right)> T(D)\right)$ then
8. $T \leftarrow R\cup\{Y\}$
9. end
10. end
11. $R \leftarrow T$
12. end
13. $H=\{V,R\}, V=\{C_1,C_2,...,C_m\}, R=\{R_1,R_2,...R_R\}$ // Construction of hypergraph structure
14. for each $i \subset \{1,2,...,R\}$ do
15. $f_s \leftarrow \bigcap_{j\in i} R_j$ // κ-Helly property
16. end

3. EXPERIMENTAL ANALYSIS

The implementation of the RSKHT had been carried out using MATLAB 6.4 on core i5 processor with 8GB RAM running on windows 7 operating system. To highlight the performance of RSKHT, the benchmark KDD cup 1999 intrusion detection dataset[8] had been used for experimentation and validation. The dataset consists of 41 conditional attributes, one decisional attribute with respect to the various types of attacks. Since the KDD cup 99 dataset is substantial, two data subsets namely, IDS_{Train} and IDS_{Test} have been prepared. The IDS_{Train} and IDS_{Test} had been obtained from 10 per cent of the KDD cup dataset[8] and corrected cup KDD dataset[8] respectively.

Table 2 shows the number of samples in each subset with respect to the various intrusion types. In order to assess

Figure 2. Hyper graph representation.

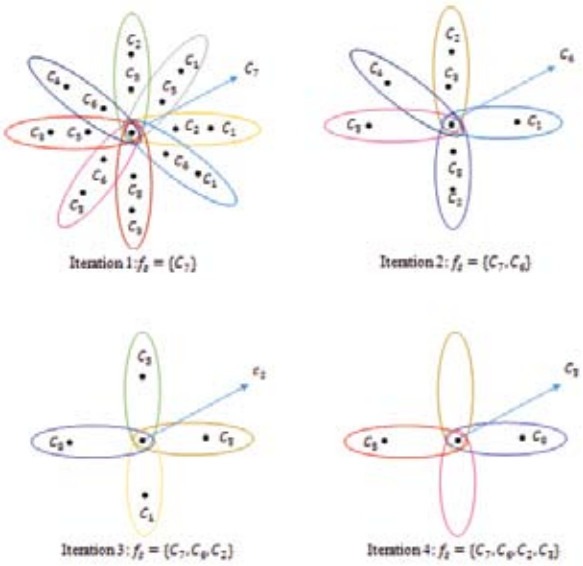

Figure 3. Identification of the optimal reduct using κ-Helly property of hyper graph.

Table 2. Samples in IDS_{Train} and IDS_{Test}

Class	Samples (IDS_{Train})	Samples (IDS_{Test})
Normal	973	606
DOS	3915	2233
U2R	5	3
R2L	13	59
Probe	41	24
Total	4947	2925

the performance of RSKHT, it had been compared with the existing feature selection techniques with respect to the reduct size, classification accuracy and time complexity. In context with existing feature selection algorithms, few attribute evaluators like Cfs subset eval (CF), consistency subset eval (CO), RSAR subset eval (RST), etc. have been hybridised with search techniques like best first (BF), genetic search (GS), particle swarm optimisation (PSO), etc. It has to be noted that, as these techniques were available as libraries in WEKA tool[9], they had not been implemented separately for the purpose of validation.

Figure 4 shows the comparative analysis of RSKHT with the existing feature selection algorithms with respect to the reduct size. From the analysis, it can be found that the

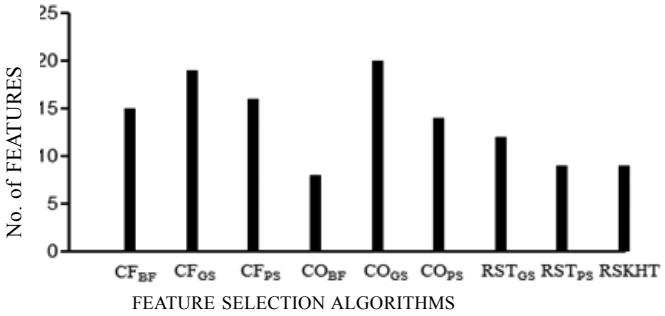

Figure 4. Reduct size of various feature selection algorithms.

Performance of RSKHT was similar to RSAR Subset Eval _PSO (RST_PS), Consistency Subset Eval_Best First (CO_ BF) as they have similar reduct size. Features in the optimal feature

Subset obtained using RSKHT for various attacks in the KDD cup 1999 dataset are given in Table 3.

Table 3. Features in the optimal feature subset obtained from RSKHT for various intrusions in KDD cup 1999 dataset

Attacks	Features
Denial of Service (DoS)	2,3,5,23,33,36,39
Users to Root (U2R)	6,9,12,16,17,32
Remote to Local user(R2L)	5,6,12,14,37,40
Probe	3,4,23,26,29,38,40
Normal	5,10,14,17,33,36

Along with the reduct size, classifier accuracy was considered as the most important criteria for the validation of RSKHT. The reducts obtained from RSKHT and various feature selection techniques were used to train and validate (10 fold cross validation) the classifiers namely Bayes Net, RBF Classifier, SVM, K Star, BF Tree, J48, and random forest[9] in an independent manner. From Table 4, it was clear that the overall accuracy of the classifiers trained using RSKHT was comparatively high with those trained with the other feature selection algorithms. Although the performance of RSKHT was found to be similar with few existing techniques (Fig. 4), RSKHT identifies the optimal feature subset that extemporise the classifier's performance in terms of accuracy. Apart from these assessment criteria, time complexity plays a vital role in determining the performance of the feature selection algorithms[10]. Time complexity can be defined as the time taken by a feature selection algorithm to identify the optimal feature subset. Tables 4 and 5 highlights the significance of RSKHT with respect to the time complexity incurred during the search for an optimal reduct. Based on the time complexity, various feature selection techniques can be ordered as follows:

$$TC_{GAFS} > TC_{PSORSFS} > TC_{CEFS} > TC_{DISMFS} > TC_{POSFS} > TC_{FSRDE} > TC_{FSMRDE} > TC_{RSKHT}$$

The feature selection techniques based on gxenetic algorithm and PSO incurs high time complexity, since they generate all possible combinations of population for each generation. In the case of CEFS algorithm, the time complexity was due to search for the equivalent class. Similarly, construction of discernibility matrix becomes the major reason behind the time complexity of DISMFS. For RSKHT, the time complexity was computed to be $O\left((A_n x e_n) x Sn\right)$, which depends on number of hyper edge construction, such that $e < A_n$, hence $(A_n x e_n) \ll A_n^2$. In most of the cases, $A_n <<< S_n$, thus time complexity of RSKHT had been minimum compared with the existing feature selection algorithms (Table 4).

Table 5. Time complexity of various feature selection algorithms

Feature selection algorithms	Time complexity*
Genetic algorithm based feature selection (GAFS)[11]	$O\left(A_n x S_n^2\right)$
Particle swarm optimisation based rough set for feature selection (PSORSFS)[12]	$O\left(A_n x S_n^2\right)$
Conditional entropy based feature selection (CEFS)[13]	$O\left(\left(A_n x S_n^2\right) + S_n^2\right)$
Positive region-based feature selection (POSFS)[14]	$O\left(A_n x S_n^2\right)$
Discrnibility matrix-based feature selection (DISMFS)[15]	$O\left((A_n + \log S_n) x S_n^2\right)$
relative decision entropy-based feature selection (FSRDE)[16]	$O\left(A_n^3 x S_n\right)$
Modified relative decision entropy-based feature selection (FSMRDE)[16]	$O\left(A_n^2 x S_n\right)$
Rough set -Helly property technique (RSKHT)	$O\left((A_n x e_n) x S_n\right)$

* $A_n \leftarrow$ *Number of Attributes;* $S_n \leftarrow$ *Number of Samples;* $e_n \leftarrow$ *Number of Hyper edges*

Table 4. Classifier's performance and run time analysis of various feature selection algorithms

Feature selection algorithms	*Classification Accuracy per cent*							*Run time (s)*
	Bayes net	**RBF classifier**	**SVM**	**KStar**	**BF tree**	**J48**	**Random forest**	
CF_BF	95.48	75.63	95.58	96.00	95.16	95.48	97.47	3.85
CF_GS	94.85	75.63	95.16	95.79	95.16	95.48	97.79	3.75
CF_PSO	95.06	75.44	95.85	95.79	95.27	95.05	97.26	3.53
CO_BF	95.48	75.52	93.27	94.85	95.27	96.42	97.58	14.60
CO_GS	94.22	70.37	90.75	96.11	94.53	94.85	97.79	14.75
CO_PSO	95.27	70.06	93.86	95.37	94.85	95.27	96.95	14.20
RST_GS	90.88	71.13	96.17	89.61	93.19	93.50	91.82	30.45
RST_PSO	91.95	73.65	96.73	91.84	93.10	93.63	93.00	27.73
RSKHT	**96.32**	**76.58**	**96.89**	**96.85**	**95.62**	**96.25**	**97.63**	**2.75**

4. CONCLUSION

IDS has become an obligatory component after firewall for many organisations, in order to protect their IT infrastructure from external or internal security breaches to trade off the security goals. The curse of high dimensionality has been found to be a significant challenge which degrades the performance of IDS due to massive volume of live network traffic with redundant and irrelevant features. To address this problem, a novel feature selection algorithm (RSKHT) had been proposed to identify the optimal feature subset using RST and κ-Helly property of the hypergraph. Reducts were obtained through RST and their

respective hypergraph representations were built. The recursive application of κ-Helly property on the pair wise intersecting neighborhood hyper edges of the hypergraph resulted in the identification of optimal feature subset. Experiments conducted using the KDD cup 1999 dataset highlights the performance of the RSKHT with respect to the reduct size, classifier accuracy and time complexity. RSKHT was found to computationally attractive, scalable and applicable over a wide range of applications such as weather forecasting, DNA sequencing, ranking cloud services[17], medical diagnosis[18], etc.. where feature selection plays a significant role.

ACKNOWLEDGEMENTS

Authors thank the IBM, India for their financial support towards this work; Authors also thank the Department of Science and Technology - Fund for Improvement of S&T Infrastructure in Universities and Higher Educational Institutions Government of India (SR/FST/MSI-107/2015 and SR/FST/ETI-349/2013) for their financial support. We would like to express our gratitude towards the unknown potential reviewers who have agreed to review this paper and provided valuable suggestions to improve the quality of the paper.

REFERENCES

1. Scarfone, K. & Mell, P. Guide to intrusion detection and prevention systems (IDPS). NIST special publication, 2007, 800, 94.
2. Stouffer, K.; Joe, F. & Karen, S. Guide to industrial control systems (ICS) security. NIST special publication, 2011, **800**(82), 16-16. doi: 10.6028/NIST.SP.800-82
3. Anderson, J.P. Computer security threat monitoring and surveillance. Vol. 17. Technical report, James P. Anderson Company, Fort Washington, Pennsylvania. 1980.
4. Aydın, M.A.; Zaim, A.H., & Ceylan, K.G. A hybrid intrusion detection system design for computer network security. *Comp. Elect. Eng.*, 2009, **35**(3), 517-526. doi: 10.1016/j.compeleceng.2008.12.005
5. Gheyas, I.A. & Smith, L.S. Feature subset selection in large dimensionality domains. *Pattern Recognition*, 2010, **43**(1), 5-13. doi: 10.1016/j.patcog.2009.06.009
6. Anaraki, J.R. & Eftekhari, M. Improving fuzzy-rough quick reduct for feature selection. *In* 19th Iranian Conference on Electrical Engineering Tehran, Iran, 2011.
7. Bretto, A.; Cherifi, H. & Ubéda, S. An efficient algorithm for *K*-Helly property recognition in a linear hypergraph. *In* 8th International Workshop on Combinatorial Image Analysis. Electronic Notes in Theoretical Computer Science, 2001. doi:10.1016/S1571-0661(04)80985-X
8. KDD Cup 1999 data set. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (Accessed on 27 January 2016)
9. Witten, I.H. & Frank, E. Data mining: Practical machine learning tools and techniques. Morgan Kaufmann. 2005. 365.
10. Alelyani, S. On feature selection stability: A data perspective. Arizona State University, Tempe, Arizona, 2013. (PhD Thesis).
11. Wroblewski, J. Finding minimal reducts using genetic algorithms. *In* Proceedings of the 2nd Annual Joint Conference on Information Science, 1995.
12. Wang, X.; Yang, J; Teng, X.; Xia, W. & Jensen, R. Feature selection based on rough sets and particle swarm optimisation. *Pattern Recognition Letters*, 2007, **28**(4), 459-471. doi: 10.1016/j.patrec.2006.09.003
13. Wang, G.Y.; Yu, H. & Yang, D. C. Decision table reduction based on conditional information entropy. *Chinese J. Comp.,* (Chinese), 2002, **25**(7), 759-766.
14. Hu, X.; Cercone, N. & Han, J. An attribute-oriented rough set approach for knowledge discovery in databases. *In* Rough sets, Fuzzy sets and knowledge discovery. Springer, London,1994
15. Hu, K.; Lu, Y. & Shi, C. Feature ranking in rough sets. *Artificial Intelligence Comm.*, 2003, **16**(1), 41-50. doi : 10.1.1.14.9476
16. Jiang, F.; Sui, Y. & Zhou, L.A. Relative decision entropy-based feature selection approach. *Pattern Recognition*, 2015, **48**(7), 2151-2163. doi : 10.1016/j.patcog.2015.01.023
17. Nivethitha, S.; Kannan K. & Shankar Sriram, V. S. A computational model for ranking cloud service providers using hypergraph based techniques. *Future Generation Comput. Sys.*, 2017, **68**,14-30. doi: 10.1016/j.future.2016.08.014
18. Nivethitha, S.; Gauthama Raman, M.R.; Kannan, K. & Shankar Sriram, V. S. Hypergraph based feature selection technique for medical diagnosis. *J. Medical Sys.*, 2016, **40**(11), pp. 239. doi: 10.1007/s10916-016-0600-8

CONTRIBUTORS

Mr M.R. Gauthama Raman obtained his BE from Anna University, Chennai, in 2012 and MTech from B.S. Abdur Rahman University, Chennai, in 2014. Currently pursuing his PhD at School of Computing, SASTRA University, Thanjavur. He is a member of Ramanujan Mathematical Society (Mem. No: 1113). His current research interests include : Network and information security, computational intelligence, data mining and big data analytics.

Mr K. Kannan is a Professor in the Department of Mathematics, SARTRA University, Thanjavur, INDIA. He obtained his Bachelor's and Master's from the University of Madras, India, in 1980 and 1982, respectively. He was conferred PhD in Mathematics in the area of Computational Fluid Dynamics by Alagappa University, Karaikudi, India, in 2000. He is a member of Ramanujan Mathematical Society (Mem. No: 1042). His specific areas of interest include : Combinatorial optimisation, artificial neural networks and hypergraph-based image processing.

Dr S.K. Pal did his post-graduation in Computer Science from the J.K. Institute of Applied Physics, Electronics & Communications, University of Allahabad in 1990 and PhD from University of Delhi in the area of Information Security. Presently working as Scientist G, Scientific Analysis Group, Delhi. His research interests include: Digital signal processing, cryptology, multimedia and network security, information hiding and soft computing.

Dr Shankar Sriram V.S. is an Associate Professor in School of Computing, SASTRA University, Thanjavur, Tamil Nadu. He received his BSc from Madurai Kamaraj University, Madurai, India in 1997. He obtained his MCA from Madurai Kamaraj University, Madurai, in 2000. He received his PhD from Birla Institute of Technology, Mesra, in 2010. His current area of research includes : Information and network security, cryptography, MANETs, steganography, and cloud computing.