Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Classification for Intrusion Detection with Different Feature Selection Methods: A Survey (2014-2016)

Rana F. Najeeb*, Ban N. Dhannoon
Computer Science Department, College of Science, AL-Nahrain University
Baghdad, Iraq

DOI: 10.23956/ijarcsse/SV7I5/0154

Abstract— Gigantic development in system based services had brought about the upsurge of web users, security dangers and digital assaults. Intrusion detection systems (IDSs) have turned into a basic segment of every network architecture, in order to safe an IT foundation from the malignant activities of the intruders. A proficient ID ought to have the capacity to detect, recognize and track the malevolent attempts made by the intruders. Intrusion is broadly perceived as an unending and repeating issue of computer systems' security with the persistent changes and expanding volume of hacking systems. The Intrusion detection system identification framework manages gigantic measure of information which contains irrelevant and redundant features producing slow training and testing process, higher asset utilization as well as poor detection rate. The feature selection approach gives enhanced prediction and reduces the computation time. Because the higher numbers of features the comprehension of the data in pattern recognition becomes difficult sometimes. That is the reason analysts have utilized diverse feature selection techniques with the single classifiers in their intrusion detection system framework to develop a model which gives a better accuracy and prediction performance. Feature selection, therefore, is a critical issue in intrusion detection. In this paper we present ideas and algorithms of feature selection used by researchers, survey existing feature selection algorithms intrusion detection system.

Keywords— Intrusion detection, Data mining, feature selection, wrapper, filter.

I. INTRODUCTION

An IDS is a key for security to offer defence of the weaknesses which happen in computer systems through any activities its work on the foundation of these steps. Gather data from a computer system.

1. Study or analyses these data.
2. Find security applicable events.
3. If there are any malevolent events, then make alarm
4. Submit the report to the administrator [1].

In 1987 when Dorothy Denning submitted an intrusion detection model afterward people start working in the arena of network security. This IDS works on the real time data and driven by four issues.

1. For settling the most existing security blemishes which display in the current system.
2. Mainly, numerous attractive and secure features lost in the current system.
3. Development of a great degree secure system is extremely difficult.
4. Most secure system is too misuses in by the insiders who misuse their privileges. Can apply certain security in contradiction of the security attacks on networks it's include essentially three steps.
   a. Prevention: -Prevent before harm.
   b. Detection:-Detect occurrence of every attack through the current.
   c. Mitigation: -Responding to the attack [2].

II. COMPONENTS OF INTRUSION DETECTION SYSTEM

IDS works on the observing of the actions occur in the computer system or network and take the data for the predefined rules from the database and analyzing the events for the malevolent substance or indications of intrusion and keep up the authentication, confidentiality, integrity of the data. For this drive, IDS contains the following components appeared in figure 1 [3].

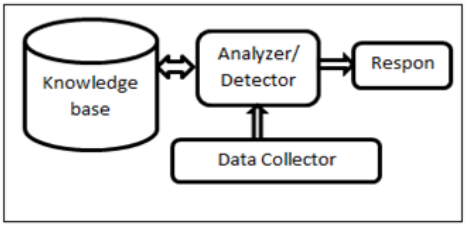

Fig.1 Typical IDS with its component

***Data Pre-processor*:**

Data pre-processor is devoted to the gathering or evaluating of data from favourite sources and gave these data to the following segment which is in charge of more operations. It transmutes the data from user access pattern to network packet level features and this prepared data utilized by the analyzer.

Analyzer (Intrusion Detector):

The analyzer or intrusion detector is the centre part which achieves analysis on the audit data and checked for the attacks. Data mining, pattern matching, soft computing, machine learning and different statistical techniques utilized as an intrusion detector. It is one of the most researched parts which decide the ability or general quality of the system.

System Profile (Database/ Knowledge Base):

The system profile is utilized to depict the ordinary and irregular user conduct. It is the database the audit information, attacks, configuration information about the present state of system and events that that will occur on the system.

Response Engine

The reaction motor contains the response system and chooses how to respond at whatever point analyzer detects an attack. Based on these mechanism systems choose either to raise an alarm without obstructing the source of the attack or block the source address for the specific timeframe. These all activity relies on upon the predefined security guidelines or strategy of the system.

III. TYPES OF INTRUSION DETECTION SYSTEMS

Intrusion detection is characterized as genuine - time checking and examination of system movement and data for potential vulnerabilities and attacks in progress. One main Restraint of existing intrusion detection system (IDS) technologies is the requirement to filter false alarms lest the operator (system or security administrator) be overcome with data. IDSs are classified in many different ways [4]: -

1. Active and passive IDS.
2. Network-based and host-based IDS.
3. Knowledge-based and behavior-based IDS.

Active Ids:

Presently more usually known as an intrusion prevention system —IPS is a system that's arranged to automatically block suspected attacks in progress with no mediation required by an operator. IPS has the advantage of providing real-time corrective action in response to an attack but however has many weaknesses too. An IPS must be set in-line along a network limit; in this way, the IPS itself is defenseless to attack. Likewise, if false alerts and genuine movement haven't been appropriately identified and filtered, authorized users and applications may be improperly denied access. Finally, the IPS itself may be utilized to impact a Denial of Service(DOS) attack by deliberately flooding the system with alarms that cause it to block connections until no associations or data transmission are accessible.

Passive Ids:

A system that's designed only to screen and analyze network traffic movement action and caution an operator to potential vulnerabilities and attacks. It isn't capable of playing out any defensive or remedial functions on its own. The main advantages of passive IDSs are that these systems can be effortlessly and quickly organized and are not usually susceptible to attack themselves.

Network-Based Ids:

Normally comprises of a network apparatus (or sensor) with a Network Interface Card (NIC) working in promiscuous mode and a discrete management interface. The IDS is put along a network segment or limit and screens all traffic on that segment.

Host-Based Ids:

Requires little programs (or operators) to be introduced on individual systems to be observed. The specialists screen the operating system and write data to log files.

Knowledge-Based (Or Signature-Based) Ids:

References a database of past attack profiles and a known system susceptibility to recognize active intrusion tries. Knowledge-based IDS is presently more common than behavior-based IDS.

Behavior-Based (Or Statistical Anomaly–Based) Ids:

References a benchmark or educated example of normal system action to recognize active intrusion tries. Deviations from this benchmark or example cause an alarm to be triggered.

IV. FEATURE SELECTION

Feature selection is the procedure of selecting relevant features, or a candidate subset of features. The evaluation criteria are utilized for getting best feature subset. In high-dimensional data (number of tests or samples << number of

features), finding the best feature subset is a troublesome errand. There are many related issues that are appeared as NP-hard. The data with number of features, there exists candidate subset of features [5].

Wrapper approaches utilize the feedback received from a particular classifier to assess the quality of the feature subset [6]. Wrapper technique utilizes the predictor and it executes as a function to decide the result. Several of search approaches are utilized which used to expand this function that build up or develop the performance accuracy. Figure 2 general wrapper algorithm[7].

```
INPUT:
D = {X, L}          // a training data set with n number of features where
                    //  X = {f1, f2, f3, ... , fn} and L labels
X'                  // predefined initial feature subset (X' ⊂ X or X' = {φ})
θ                   // a stopping criterion
OUTPUT: X'opt       // an optimal subset
-----------------------------------------------------------------
Begin:
Initialize:
    Xopt = X';
    φopt = E(X', A);     // evaluate X' by using mining algorithm A
do begin
    Xg = generate(X); // Subset generation for evaluation
    φ = E(Xg, A) ;     // Xg current subset evaluation by A
   If (φ > φopt)
      φopt = φ;
      X'opt = Xg;
repeat (until θis not reached );
end
return X'opt;
end;
```

Fig. 2 General Wrapper Algorithms

Filter approaches assess on the statistical characteristics of the training data. Concerning its lower computational drift and decreased time complexity, this strategy is applied on the huge data sets like NSL KDD or KDD CUP 99 datasets [8]. Filter technique utilizes variable position techniques to decrease the irrelevant features and those position techniques are used due to the effortlessness and their application on the pragmatic datasets. In filter methods, the features are considered to irrelevant which are independent of the class label, figure 3 general filter algorithm [9].

```
INPUT:
D = {X, L}          // a training data set with n number of features where
                    //  X = {f1, f2, f3, ... , fn} and L labels
X'                  // predefined initial feature subset (X' ⊂ X or X' = {φ})
θ                   // a stopping criterion
OUTPUT: X'opt       // an optimal subset
-----------------------------------------------------------------

Begin:
Initialize:
     Xopt = X';
     φopt = E(X', Im);     // evaluate X' by using an independent measure Im
do begin
     Xg = generate(X);   // Subset generation for evaluation
     φ = E(Xg, Im) ;     // Xg current subset evaluation by Im
    If (φ > φopt)
       φopt = φ;
       X'opt = Xg;
 repeat (until θis not reached );
 end
 return X'opt;
 end;
```

Fig.3 General Filter Algorithm

Generally, a feature selection algorithm incorporates four sections which are appeared in figure 2, generation, evaluation, stopping criterion and validation [10].

The candidate subset generation: is a procedure of looking feature subsets, and the got subset will be utilized as input for evaluation function. The selection of the underlying subset is the begin Feature selection algorithm, and the beginning stage of subset generation procedure, which is divided into three categories.

Initial subset is empty. During the time spent searching, the algorithm adds candidate features to candidate subset one by one. This technique is called forward search.

Initial subset is the similar as feature set of a given dataset. What's more it eliminates irrelevant or redundant features from the initial subset step by step in the search procedure, namely the backward search.

The initial subset is created randomly, and then the feature is included or erased one by one in the search process.

The evaluation function: is utilized to assess the qualities of the candidate subset acquired by the search. It will contrast evaluation value with the best ideal value stored before. If the evaluation value is higher, the essential candidate subset will be substituted [11].

Independent Criteria mainly, a filter model is utilized for independent criteria feature subset selection. It does not include any learning algorithm.

It exploits the basic characteristics of the training data to assess the goodness of the feature subset. There are several independent criteria namely Distance measures, Information or uncertainty measures, Probability of error measures, Dependency measures, Interclass distance measures, and Consistency measures.

Dependent Criteria need a predetermined mining algorithm. The performance of the algorithm is utilized to evaluate the goodness of the feature subset to figure out which features are selected. The selected feature subset is best is most appropriate to a fixed algorithm. Hence, the performance of the algorithm is normally better. In any case, it is computationally costly, in light of the fact that each feature subset estimates accuracy. A wrapper model is utilized for dependent criteria [12].

Stopping Criteria*:* the halting model for the feature selection procedure must be characterization. There are some broad stopping criteria: Predefined extreme number of iterations or smallest number of features or minimum classification error rate the search finishes Removal or additions of features to the subset do not produce a huge difference [13].

Validation part: proves the classification efficiency performance of the feature selection results about specific conditions. It is not a part of feature selection process but is needed in the practical application [14]. Approval is for the most part to train and test feature subset in some kind of classifier and contrast prediction results with unique dataset results, or other feature selection results. Contrast may be classification accuracy or computational complexity and so on [15].

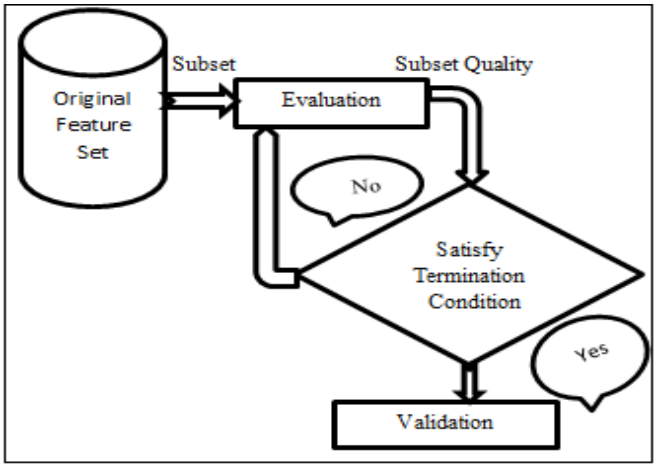

Fig. 4 Feature Selection Structure

V. NSL KDD DATASET DESCRIPTION

Different drawbacks of KDD CUP 99 which was the primary cause to reduction in the performance of different IDS [14] led to the development of NSL KDD dataset [16]. NSL KDD is the refined variant and furthermore called the successor of KDD CUP dataset. It comprises of all the required attributes from KDD CUP dataset. It is an open source data and can be downloaded effortlessly [17].

The advantage of utilizing this dataset is repetitive record is removed and adequate number of records is available for train and test data. It contains of 41 attributes which is classified under Nominal, Binary and Numeric the training dataset is comprised of 21 different attacks out of the 37 shown in the test dataset [18]. The known attack types are those present in the training dataset while the novel attacks are the extra attacks in the test dataset i.e. not existing in the training datasets. The attack kinds are assembled into four categories: DoS, Probe, U2R and R2L.Table 2 shows the major attacks in both training and testing dataset [19].

Table 1: Major Attacks in both Training and Testing Dataset

Attack Class	Attack Type
DoS	Back, Land, Neptune, Pod, Smurf, Teardrop, Apache2, Udpstorm, Processtable, Worm.
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint.
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpget attack, Httptunnel, Sendmail, Named.
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps.

Table 2: Survey for Most Researchers on Ids with Feature Selection

Author Name	Researcher Name	Year	Data Set Name	Feature Selection Used	Classifier Name Used	Accu racy
Senthilnayaki Balakrishn an, Venkatalakshmi, Kan nan [20]	Intrusion Detection System Using Feature Selection and Classification Technique	2014	KDD Cup dataset	Optimal Feature Selection algorithm based on Informatio n Gain Ratio	Support Vector Machine and Rule Based Classificati on	91% 80%
El-Sayed M. El-Alfya, Feras N. Al-Obeidatb[13]	A Multicriterion Fuzzy Classification Method with Greedy Attribute Selection For Anomaly-Based Intrusion Detection	2014	KDD'99 Dataset	Greedy Attribute Selection	A Multicriterion Fuzzy Classification Method	99.9 %
Jatuphum Juanchaiyaphum , Ngamnij Arch -int, Somjit Arch -int andSaiyan Saiyod[9]	A Novel Lightweight Hybrid Intrusion Detection Method Using a Combination of Data Mining Techniques	2015	NSL-KDD dataset.	Best First and CfsSubsetEval	K- Means clustering, C4.5 decision tree	99.52 %,
Vipin Singh, Himanshu Arora [24]	Network intrusion detection using feature selection and PROAFTN Classification	2015	KDD Cup dataset	particle of swarm optimization	PROAFTN Classification (fuzzy logic and protein cell classification technique.)	96%- -- 98%
Sedigheh Khajouei Nejad, Sam Jabbehdari, Mohammad Hossein Moattar[19]	A Hybrid Intrusion Detection System Using Particle Swarm Optimization for Feature Selection	2015	KDD-CUP 99	Accelerated Particle Swarm Optimization (APSO) algorithm	KNN, Decision Tree and Neural Network	97|%
GuiPing Wang, ShuYu Chen and Jun Liu [15]	Anomaly-based Intrusion Detection using Multiclass - SVM with Parameter/Optimized by PSO	2015	KDD'99 Dataset	particle swarm optimization	multiclass support vector machine	97%
Adel Sabry Eesa, Zeynep Orman, Adnan Mohsin Abdulazeez Brifcani [17].	A Novel Feature-Selection Approach Based On The Cuttlefish Optimization Algorithm For Intrusion Detection Systems	2015	KDD Cup 99 Dataset	Cuttlefish Algorithm	Decision Tree	91%.
Sumaiya Thaseen Ikram ,Aswani Kumar Cherukuri [21]	Intrusion Detection Model Using Fusion Of Chi-Square Feature Selection And Multi Class SVM	2016	NSL-KDD Dataset Which Is an Enhanced Version of KDD Cup 1999 Dataset	Chi-Square Feature Selection	Multiclass SVM	95%

Author Name	Researcher Name	Year	Data Set Name	Feature Selection Used	Classifier Name Used	Accuracy
Dhuha I. Mahmood, Sarab M. Hameed[4]	A Feature Selection Model based on Genetic Algorithm for Intrusion Detection	2016	NSL-KDD	genetic algorithm	Naive Bayes	97%
M.R. Gauthama Rama, K. Kannan, S.K. Pal, and V. S. Shankar Sriram [14]	A novel rough set κ-Helly property technique (RSKHT) feature selection algorithm	2016	KDD cup 1999 dataset	rough set κ-Helly property technique	Bayes Net, RBF, BF tree Classifier, SVM, K Star, J48, and random forest	96% 76% 96% 96% 95% 96% 97%
Kajal Rai ,M. Syamala Devi, Ajay Guleria[11]	Decision Tree Based Algorithm for Intrusion Detection	2016	NSL-KDD 99	Information Gain feature	C4.5	76%
Rajinder Kaur, Monika Sachdeva, Gulshan Kumar[2]	An Empirical Analysis of Classification Approaches for Feature Selection in Intrusion Detection	2016	NSL-KDD 99	Chi Squared Eva, Correlation-based Feature Selection (CFS), Gain ratio	Bayes Net, Naive Bayes, J48, Random Forest and Decision Tree	59% 73% 61% 60% 57%
Koushal Kumar , Jaspreet Singh Batth[12]	Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms	2016	NSL-KDD data set	Correlation-based Feature Selection (CFS), Information Gain feature evaluator (IGF), Gain ratio	Naive Bayes	93% 94% 97%
Jingping Song[7]	Feature Selection for Intrusion Detection System	2016	KDD 99 dataset	Modified Mutual Information based Feature Selection algorithm	C4.5	94%
Vaishali Chahar, Rita Chhikara, Yogita Gigras and	Significance of Hybrid Feature Selection Technique for Intrusion Detection	2016	KDDCup9 9	particle swarm optimization	Support Vector Machine	98%

Latika Singh [23]	Systems					
Mehdi Hosseinzadeh Aghdam, and Peyman Kabiri [10]	Feature Selection for Intrusion Detection System Using Ant Colony Optimization	2016	KDD Cup 99 and NSL-KDD	Ant Colony Optimization	Support Vector Machine	98%
Udaya Sampath K. Perera Miriya Thanthrige, Jagath Samarabandu, Xianbin Wang[22]	Machine Learning Techniques For Intrusion Detection On Public Dataset	2016	Aegean Wi-Fi Intrusion Dataset (Awid)	Information Gain And Chi Squared Statistics Based Feature Selection.	OneR J48 Random Forest Random Tree Ada Boost	92%92%92% 91% 91%
Harvinder Pal Singh Sasan and Meenakshi Sharma[8]	Intrusion Detection Using Feature Selection and Machine Learning Algorithm With Misuse Detection	2016	NSL-KDD	gain ratio	Classification & regression trees (CART) algorithm	88%
Chidananda Murthy P., Dr.A. S. Manjunatha, An ku Jaiswal and Madhu B. R. [3]	Building Efficient Classifiers for Intrusion Detection with Reduction of Features	2016	NSL KDD	Search method and ranker method	Random Tree	99%

VI. CONCLUSION

Intrusion Detection Systems have turned out to be imperative and an essential segment of practically every computer and network security. As network speed turns out to be quicker, there is a develop requirement for IDS to be lightweight, efficient and accurate with high detection rates (DR) and low false positive rates (FAR).

Different challenges confronted by intrusion detection systems are curse of feature dimensionality and rising data complexities. Feature selection selects a subset of relevant features, removes irrelevant and redundant features from the dataset to construct strong, proficient, precise and lightweight intrusion detection system to guarantee timeliness for real time.

A lot of feature selection approaches have been proposed by researchers in intrusion detection system to manage these problems. This paper has offered to survey this fast emerging field and addresses the primary contribution of feature selection research proposed for intrusion detection. We presented that why feature selection method is vital in IDS. We surveyed the existing feature selection methods for IDS categorized as filter, wrapper and hybrid. We also presented the performance of these methods based on different metric on KDD Cup'99 and NSL KDD dataset.

REFERENCE

[1] Lundin, E., Jonsson, E," Survey of Intrusion Detection Research", Chalmers Publication Library, 2002.

[2] Rajinder Kaur, Monika Sachdeva, Gulshan Kumar," An Empirical Analysis of Classification Approaches for Feature Selection in Intrusion Detection", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 9, September 2016.

[3] Chidananda Murthy P., Dr.A. S. Manjunatha, An ku Jaiswaland Madhu B. R.," Building Efficient Classifiers for Intrusion Detection with Reduction of Features", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 6, pp. 4590-4596,2016.

[4] Dhuha I. Mahmood, Sarab M. Hameed," A Feature Selection Model based on Genetic Algorithm for Intrusion Detection", Iraqi Journal of Science, pp:168-175 ,2016.

[5] Levent Koc," Application of a Hidden Naïve Bayes Multiclass Classifier in Network Intrusion Detection", Ph.D .Thesis, January 31, 2013.

[6] Abdur Rahman Onik, NutanFarahHaq, Lamia Alam ,Tauseef Ibne Mamun," An Analytical Comparison on Filter Feature Extraction Method in Data Mining using J48 Classifier", International Journal of Computer Applications ,Volume 124 ,No.13, pp. (0975 –8887),August 2015.

[7] Jingping Song," Feature Selection for Intrusion Detection System", Ph.D. Thesis, March 3, 2016.

[8] Rajni Tewatia, Asha Mishra," Introduction To Intrusion Detection System: Review", International Journal Of Scientific & Technology Research Volume 4, Issue 05, May 2015.

[9] Jatuphum Juanchaiyaphum, Ngamnij Arch-int, Somjit Arch-int andSaiyan Saiyod," A Novel Lightweight Hybrid Intrusion Detection Method Using a Combination of Data Mining Techniques", International Journal of Security and Its Applications Vol. 9, No. 4, pp. 91-106, 2015.

[10] Mehdi Hosseinzadeh Aghdam, and Peyman Kabiri," Feature Selection for Intrusion Detection System Using Ant Colony Optimization", international Journal of Network Security, Vol.18, No.3, PP.420-432, May 2016.

[11] Kajal Rai, M. Syamala Devi, Ajay Guleria," Decision Tree Based Algorithm for Intrusion Detection", Int. J. Advanced Networking and Applications Volume: 07 Issue: 04 Pages: 2828-2834 (2016).

[12] Koushal Kumar , Jaspreet Singh Batth," Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 150 – No.12, September 2016.

[13] El-Sayed M. El-Alfya, Feras N. Al-Obeidatb "A Multicriterion Fuzzy Classification Method with Greedy Attribute selection For Anomaly-Based Intrusion Detection", Procedia Computer Science Vol.34 ,pp.55 – 62,2014.

[14] M.R. Gauthama Rama, K. Kannan, S.K. Pal, and V. S. Shankar Sriram," A novel rough set κ-Helly property technique (RSKHT) feature selection algorithm", defense Science Journal, Vol. 66, No. 6, pp. 612-617, November 2016.

[15] GuiPing Wang, ShuYu Chen and Jun Liu," Anomaly-based Intrusion Detection using Multiclass - SVM with Parameters Optimized by PSO", International Journal of Security and Its Applications Vol. 9, No. 6, pp. 227-242,2015.

[16] Harvinder Pal Singh Sasan and Meenakshi Sharma," Intrusion Detection Using Feature Selection and Machine Learning Algorithm With Misuse Detection", International Journal of Computer Science & Information Technology (IJCSIT) ,Vol. 8, No 1, February 2016.

[17] Adel Sabry Eesa, Zeynep Orman, Adnan Mohsin Abdulazeez Brifcani," A Novel Feature-Selection Approach Based On the Cuttlefish Optimization Algorithm For Intrusion Detection Systems" Expert Systems with Applications Vol.42, pp. 2670–2679, 2015.

[18] Sakshi Sharma and Manish Dixit," A Review on Network Intrusion Detection System Using Open Source Snort", International Journal of Database Theory and Application Vol.9, No.4, pp.61-70, 2016.

[19] Sedigheh Khajouei Nejad, sam Jabbehdari, mohammad Hossein Moattar "A Hybrid Intrusion Detection System Using Particle Swarm Optimization for Feature Selection", International Journal of Soft Computing and Artificial Intelligence, Volume-3, Issue-2, Nov-2015.

[20] Senthilnayaki Balakrishnan, Venkatalakshmi, Kannn," Intrusion Detection System Using Feature Selection and Classification Technique", International Journal of ComputeScienceand Application (IJCSA) Volume 3 Issue 4, November 2014.

[21] Sumaiya Thaseen Ikram , Aswani Kumar Cherukuri ," Intrusion Detection Model Using Fusion Of Chi-Square Feature Selection", Journal of King Saud University – Computer and Information Sciences ,2016.

[22] Udaya Sampath K. Perera Miriya Thanthrige, Jagath Samarabandu, Xianbin Wang," Machine Learning Techniques for Intrusion Detection on Public Dataset", IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2016.

[23] Vaishali Chahar, Rita Chhikara, Yogita Gigras and Latika Singh," Significance of Hybrid Feature Selection Technique for Intrusion Detection Systems", Indian Journal of Science and Technology, Vol 9(48), December 2016.

[24] Vipin Singh, Himanshu Arora," Network intrusion detection using feature selection and PROAFTN Classification", International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015.

