ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Hybrid Intrusion Detection Model Based on Clustering and Association

Manish Somani[1], Roshni Dubey[2]

M.Tech Research Scholar, SRIT, Jabalpur, M.P, India[1]

Assistant Professor, SRIT, Jabalpur, M.P, India[2]

ABSTRACT: There are several intrusion detection model are presented till now. But there is still need of betterment in this direction. Our paper focuses on the limitation faced in the traditional approach. In this paper we suggest a hybrid framework based on clustering and association. Clustering is used for separate it on the basis of various classes and on the bases of classes we can classify it. FP growth algorithm is then used as the association classifier which can classify the data accordingly on the same set of category which can be proof to be better in terms of intrusion detection.

KEYWORDS: FP-growth, Association, Cluster, Classifier

I.INTRODUCTION

With the continuous development of the intrusion technology of the network, the intrusion behaviors are characterized with uncertainty, complexity, diversity and dynamic tendency etc. [1]. Early detection is an important concern in intrusion detection. Intrusion Detection System (IDS) [2] is an important detection used to preserve data integrity. Intrusion Detection Systems (IDS) is a combination of software and hardware that attempts to perform intrusion detection [3]. The main motivation behind using intrusion detection in data mining [4][5][6][7][8][9] is automation. Pattern of the normal behavior and pattern of the intrusion can be computed using data mining. To apply data mining techniques in intrusion detection, first, the collected monitoring data needs to be preprocessed and converted to the format suitable for mining processing.

KDD99Cup and DARPA98 datasets [10],[11] provided by MIT Lincoln Laboratories are widely used as training and testing datasets for the evaluation of IDSs KDDCUP'99 intrusion detection dataset contains a standard dataset and training data to be audited .It analyzes the connections of 41features. All the connections can be divided into 5 categories including normal network connection, and other four categories are Denial of Service Attack (DOS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing Attack. The KDD dataset has various forms of data such as nominal, continuous, discrete, and symbolic, with significant resolution and range.

Clustering is the method of grouping objects into meaningful subclasses so that members from the same cluster are quite similar and members from different clusters are quite different from each other[12]. Therefore, clustering methods can be useful for classifying log data and detecting intrusions. Clustering intrusion detection is based on two assumptions suggested in [12]. The first assumption is that the number of normal action is far greater than the number of intrusion action. The second assumption is that the intrusion action makes a difference with the normal action. However, these cluster-based IDS have many drawbacks: k-means is used for intrusion detection to detect unknown attacks and partition large data space effectively [13]. But it has two shortcomings: number of cluster dependence and degeneracy. To gain an optimal k is a NP hard problem. Ali and Yu guan presented a heuristic k-means algorithm called Ymeans [14]. Otherwise, Wei used improved FCM algorithms [15] to obtain an optimized k. A. Prabakar [16] combined with C4.5 decision tree to improve performance of k-means intrusion detection system. The superset and subset association will improve the searching [17] and mining will be improved the classification accuracy [18].

The remaining of this paper is organized as follows. In Section 2 we discuss about intrusion detection tools. Literature Survey in section 3. In section 4 we discuss about the proposed framework. The conclusions are given in Section 5.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Finally references are given.

II.INTRUSION DETECTIONTOOLS[17][18]

In [19] authors provided a list of most common IDS tools [20] describing their features. SNORT - This lightweight network intrusion detection and prevention system excels at traffic analysis and packet logging on IP networks. It detects threats, such as buffer overflows, stealth port scans, CGI attacks, SMB probes and NetBIOS queries, NMAP and other port scanners and DDoS clients, and alerts the user about them. It develops a new signature to find vulnerabilities. It records packets in their human-readable form from the IP address.
OSSEC – HIDS – It is scalable, multi-platform, open source Host-based Intrusion Detection System (HIDS). It has a powerful correlation and analysis engine, integrating log analysis; file integrity checking; Windows registry monitoring; centralized policy enforcement; rootkit detection; real-time alerting and active response.
FRAGROUTE – It is a one-way fragmenting router -IP packets get sent from the attacker to the Fragrouter, which transforms them into a fragmented data stream to forward to the victim. Fragrouter helps an attacker launch IP-based attacks while avoiding detection. METASPLOIT - It is an advanced open-source platform for developing, testing, and using exploit code. It ships with hundreds of exploits, as you can see in their online exploit building demo. This makes writing your own exploits easier, and it certainly beats scouring the darkest corners of the Internet for illicit shell code of dubious quality. TRIPWIRE – It Detects Improper Change, including additions to, deletions from and modifications of file systems and identifies the source. It Simplifies and Eases Management of Change Monitoring Policies. But instead of several methodology in [19] they suggest that there are still many challenges Like Improving, mining, and reducing intrusion detection data are critical to dealing with multisensory architectures of the future. Fast and flexible detection techniques are necessary to identify the vast variety of clever and unusual attacks undoubtedly encounter are suggested in [19].

III.LITERATURE SURVEY

In 2010,G. Schaffrath et al. [21] provide a survey of current research in the area of flow-based intrusion detection. The survey starts with a motivation why flow-based intrusion detection is needed. The concept of flows is explained, and relevant standards are identified. The paper provides a classification of attacks and defense techniques and shows how flow-based techniques can be used to detect scans, worms, Botnets and (DoS) attacks.
In 2012, R.Venkatesan et al. [22] survey and analysis that data mining techniques have been successfully applied in many fields like Network Management, Education, Science, Business, Manufacturing, Process control, and Fraud Detection. Data Mining for IDS is the technique which can be used mainly to identify unknown attacks and to raise alarms when security violations are detected.
In 2012, Sneha Kumari et al. [23] suggest that the Over the past several years, the Internet environment has become more complex and untrusted. Enterprise networked systems are inevitably exposed to the increasing threats posed by hackers as well as malicious users internal to a network. IDS technology is one of the important tools used now-a-days, to counter such threats. Authors also provide the comparison study which is based on artificial neural network (ANN), Bayesian network classifier (BNC), Support Vector Machine (SVM) and Decision Tree (DT).
In 2012, Vineet Richariya et al. [24] analyzed the performance and applicability of the well knows IDS system based on mobile agent with their pros and cons. Mobile agent is efficient way to find out the intruder in distributed system. The main features of mobile agents are intelligence and mobility which is the core motivation to us to designed cost. The aim of their review work is to help to select appropriate IDS systems as per their requirement and application.
In 2012, Deepak Rathore et al. [25] proposed an ensemble Cluster Classification technique using SOM network for detection of mixed variable data generated by malicious software for attack purpose in host system. In their methodology SOM network control the iteration of distance of different parameters of ensemble their experimental result which show better empirical evaluation on KDD data set 99 in comparison of existing ensemble classifier.
In 2012, LI Yin–huan [26] focuses on an improved FP-Growth algorithm. According to author Preprocessing of data mining can increase efficiency on searching the common prefix of node and reduce the time complexity of building FP-tree. Based on the improved FP Growth algorithm and other data mining techniques, an intrusion detection model is carried out by authors. Their experimental results are effective and feasible.
In 2012, P. Prasenna et al. [27] suggested that in conventional network security simply relies on mathematical algorithms and low counter measures to taken to prevent intrusion detection system, although most of this approaches

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

in terms of theoretically challenged to implement. Authors suggest that instead of generating large number of rules the evolution optimization techniques like Genetic Network Programming (GNP) can be used .The GNP is based on directed graph. They focus on the security issues related to deploy a data mining-based IDS in a real time environment. They generalize the problem of GNP with association rule mining and propose a fuzzy weighted association rule mining with GNP framework suitable for both continuous and discrete attributes.
In 2013, Manish et al. [28] present an efficient framework for intrusion detection which is based on Association Rule Mining (ARM) and K-Means Clustering. K- Means clustering is use for separation of similar elements and after that association rule mining is used for better detection. Detection Rate (DR), False Positive Rate (FPR) and False Negative Rate (FNR) are used to measure performance and analysis experimental results.

IV.PROPOSED WORK

Our proposed work is better explained with the flowchart shown in figure1. According to figure 1 we first accept the data from the dataset. Based on the dataset it is divided into number of classes separately for this we apply K-means clustering algorithm.

The algorithm shown below:
The k-means algorithm is one of the widely recognized clustering tools that are applied in a variety of scientific and industrial applications. K-means groups the data in accordance with their characteristic values into K distinct clusters. Data categorized into the same cluster have identical feature values. K, the positive integer denoting the number of clusters, needs to be provided in advance.

The steps involved in a K-means algorithm are given subsequently:

1. K points denoting the data to be clustered are placed into the space. These points denote the primary group centroids.
2. The data are assigned to the group that is adjacent to the centroid.
3. The positions of all the K centroids are recalculated as soon as all the data are assigned.

Steps 2 and 3 are reiterated until the centroids stop moving any further. This results in the Segregation of data into groups from which the metric to be minimized can be deliberated. The preprocessed software estimation data warehouse is clustered using the K-means algorithm with K value as 4. Because we need the separation based on four different object oriented parameters that is class, object, inheritance and dynamic behavior.

In our proposed work we first access the data from the valid database like KDD CUP 99 database is accessed.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

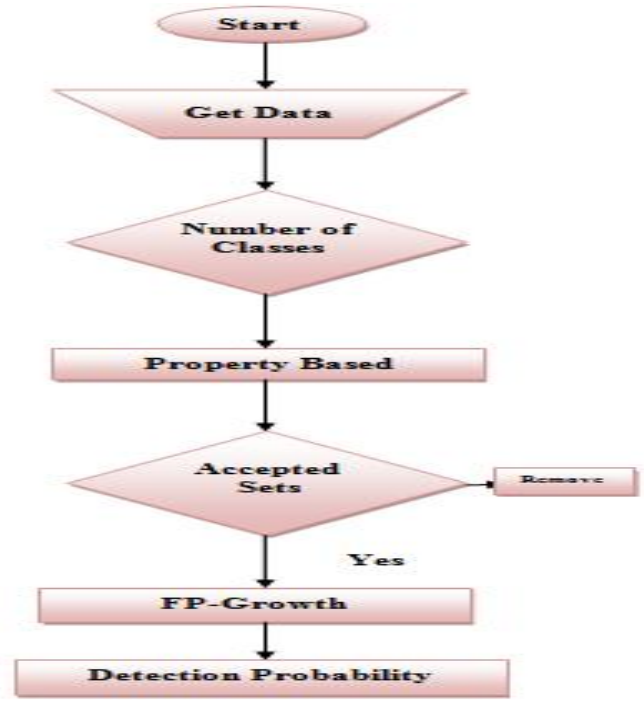

Figure 1: Flowchart of working Process

Then the data is pre-processed. Pre-process phase is like the data audit phase. Because the data we taken are not necessary support the properties our framework. So in pre-processing phase we first make it compatible to the framework we used. Then we also check the redundant data so that we only process the meaningful data, means the set contains unique items, so the processing overhead is reduced. Then we apply FP growth algorithm.

FP-Growth Algorithm

We can divide the construction in three different parts:

1. The dataset is first scan based on frequency or the occurrences we then select only frequent item set and other infrequent items are deleted.
2. For each transaction we scan the database for the formation of tree:
    a. If the transaction is unique from one path then we set the initial counter value as 1.
    b. But if it shares the common attributes increment all the parts and needed create a new initial then.
3. The above process is continued till all transactions are mapped.

The above algorithm is used to generate the frequent set from the database we taken. Here frequent means the higher support value achieved. It provides the data generation efficiently forming the trees in a hierarchical structure which stares the data in descending order. If uses divide and conquer technique to achieve the above phenomena. The next procedure is defined below:

1. Set out exordium paths for a particular suffix node. This is superior by aggregation all the paths containing a particular suffix node. Peasant-like approach that ends with this suffix is examined.
2. Turn to account the begin path tree determine whether the suffix is frequent. This is rank by uniting the dormant counts attached around the bulge and if the number is greater than or equals to the minsup the node is frequent. If the node isn't frequent the analysis ends for this suffix.
3. Convert the prefix paths into a conditional FP-tree.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

1. Advance the in a holding pattern counts relate the prefix paths to reflex the actual number of transactions containing the item set.
2. Shorten the introduce paths by removing the nodes of the chosen suffix
3. ick as a matter of actual fact rove may merge longer be deal with if the support count of a particular node is less than min-sup it is no longer frequent and should be pruned.
4. Repeat I →III for all prefix paths for the chosen suffix.

4. Repeat Steps 1-3 for all suffix nodes to determine the frequent item set for the dataset.

Then we apply any optimization technique for better detection in the possible sets.

V.RESULT ANALYSIS

For showing the effectiveness of our algorithm we first consider the dataset from 70000 to 80000 from the KDD dataset.

Table 1: Data Set Classification 1

Normal Data	Attack Nodes	Attack Node From Normal Data
6000	4001	884

Then by applying our support based classification we will obtain the classification accuracy as follows.

Min-Support-60%

Table 2: Obtained attack Nodes

DoS	U2R	R2L	Probe
3589	5	72	897

Table 3: Real attack Nodes

DoS	U2R	R2L	Probe
3589	5	74	898

As shown in the table 3 we miss some values in R2L and Probe. So the classification accuracy is less in R2L and probe as shown in figure 2.

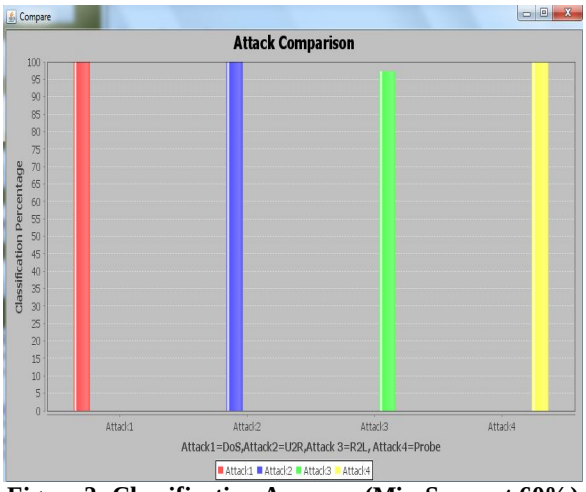

Figure 2: Classification Accuracy(Min-Support 60%)

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Min-Support-70%
Table 4: Obtained attack Nodes

DoS	U2R	R2L	Probe
3261	0	2	439

Table 5: Real attack Nodes

DoS	U2R	R2L	Probe
3589	5	74	898

As shown in the table 4 we miss values in DOS,R2L and Probe. So the classification accuracy is less in DOS,R2L and probe as shown in figure 3.

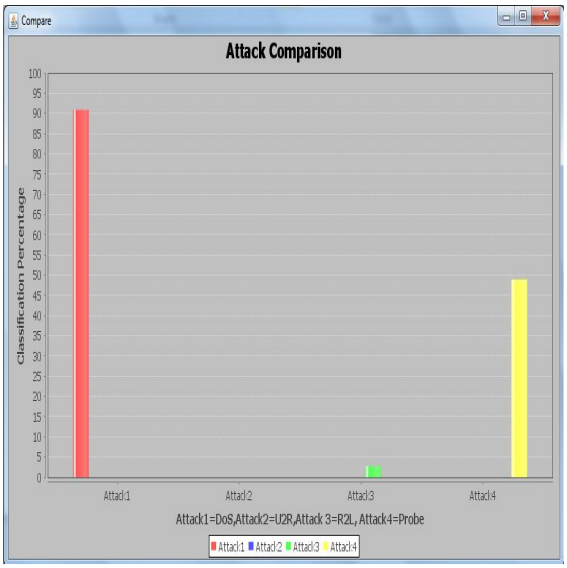

Figure 3: Classification Accuracy(Min-Support 70%)

For showing the effectiveness of our algorithm we first consider the dataset from 100773 to 108650 from the KDD dataset.

Table 6: Data Set Classification 1

Normal Data	Attack Nodes	Attack Node From Normal Data
4642	3236	737

Then by applying our support based classification we will obtain the classification accuracy as follows.

Min-Support-60%
Table 7: Obtained attack Nodes

DoS	U2R	R2L	Probe
2913	2	66	737

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Table 8: Real attack Nodes

DoS	U2R	R2L	Probe
2913	3	68	737

As shown in the table 8 we miss some values in U2R and R2L. So the classification accuracy is less in U2R and R2L as shown in figure 4.

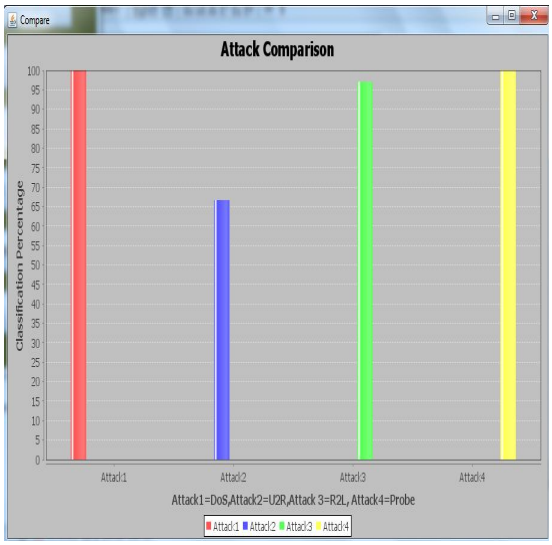

Figure 4: Classification Accuracy(Min-Support 60%)

Min-Support-70%

Table 9: Obtained attack Nodes

DoS	U2R	R2L	Probe
3203	4	127	1125

Table 10: Real attack Nodes

DoS	U2R	R2L	Probe
2913	3	68	737

As shown in the table 4 we miss values in DOS, U2R, R2L and Probe. So the classification accuracy is less in DOS, U2R, R2L and probe as shown in figure 5.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

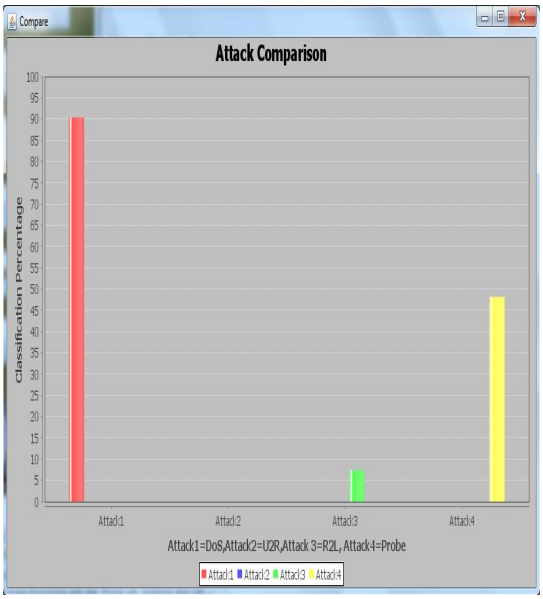

Figure 5: Classification Accuracy(Min-Support 70%)

V.CONCLUSION

Habituated Evidence Mining techniques perform on business like observations such as corporate databases; this has been an active area of research for many years. It evokes the security problem so that any unavoidable situations not arise in the future. Intrusion detection is an area growing in relevance as more and more sensitive data are stored and processed in networked systems. The system of Intrusion detection also monitors their networks node for any malicious behavior. Selecting appropriate data mining algorithms and designing IDS model are effective measures in order to improve system detection performance. In this paper we discuss an efficient framework for intrusion detection based on FP growth.

REFERENCES

[1] Wen Jie Tian; Ji Cheng Liu, "A new network intrusion detection identification model research," Informatics in Control, Automation and Robotics (CAR), 2010 2nd International Asia Conference on , vol.2, no., pp.9,12, 6-7 March 2010.
[2] R. Bane, N. Shivsharan, "Network intrusion detection system (NIDS)", pp. 1272-1277, 2008.
[3] Gudadhe, M.; Prasad, P.; Wankhade, K., "A new data mining based network Intrusion Detection model," Computer and Communication Technology (ICCCT), 2010 International Conference on , vol., no., pp.731,735, 17-19 Sept. 2010.
[4] S. T. Brugger, "Data mining methods for network intrusion detection",pp. 1-65, 2004.
[5] W. Lee, S. J. Stolfo, "Data Mining Approaches for Intrusion Detection",Proceedings of the 1998 USENIX Security Symposium, 1998.
[6] W. Lee, S. J. Stolfo, "Data mining approaches for intrusion detection" Proc. of the 7th USENIX Security Symp.. San Antonio, TX, 1998.
[7] W. Lee, S. J. Stolfo, K. W. Mok, "A data mining framework for building intrusion detection models", Proc. of the 1999 IEEE Symp.on Security and Privacy, pp. 120--132. Oakland, CA, 1999.
[8] M. Panda, M. Patra, "Ensemble rule based classifiers for detecting network intrusions", pp 19-22, 2009.
[9] Z. Yu, J. Chen, T. Q. Zhu, "A novel adaptive intrusion detection system based on data mining", pp.2390-2395, 2005.
[10] Kddcup 1999 data [Online]. Available: kdd.ics.uci.edu/databases/ kddcu p99/kddcup99.html.
[11] Darpa Intrusion Detection datasets [Online]. Available:www.ll.mit.edu/missioncommunications/ist/corpora/ideval/data/index.html.
[12] Zhou Mingqiang; Huang Hui; Wang Qian, "A graph-based clustering algorithm for anomaly intrusion detection," Computer Science & Education (ICCSE), 2012 7th International Conference on , vol., no., pp.1311,1314, 14-17 July 2012.
[13] M.Jianliang, S.Hai kun, B.Ling. The Application on Intrusion Detection Based on K-means Cluster Algorithm. International Forum onInformation Technology and Application.2009.
[14] Yu Guan, Ali A. Ghorbani, and Nabil Belacel. Y-means: a clustering method for intrusion detection. In Canadian Conference on Electrical and Computer Engineering, pages 14, Montral, Qubec, Canada, May 2003.
[15] Wei Jiang, Min Yao, Jun Yan. Intrusion detection based on improved fuzzy c-means algorithm. Information science and engineering, 2008.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

[16] Amuthan Prabakar Muniyandia, R. Rajeswarib, R. Rajaramc. Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm. International Conference on Communication Technology and System Design. Procedia Engineering 30 (2012):174-182.

[17] Dubey, Ashutosh Kumar, V. Agarwal, and Y. Khandagre. "Knowledge discovery with a subset-superset approach for Mining Heterogeneous Data with dynamic support." Software Engineering (CONSEG), 2012 CSI Sixth International Conference on. IEEE, 2012.
[18] Dubey, Ashutosh K., and Shishir K. Shandilya. "A novel J2ME service for mining incremental patterns in mobile computing." Information and Communication Technologies. Springer Berlin Heidelberg, 2010.
[19] Karthikeyan .K.R and A. Indra," Intrusion Detection Tools and Techniques A Survey", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010.
[20] "Top 5 Intrusion Detection Systems", http://sectools.org/ids.html.
[21] G.Schaffrath,R. Sadre,C. Morariu,A.Pras and B.Stiller, "An Overview of IP Flow-Based Intrusion Detection", Communications Surveys & Tutorials, IEEE 2010.
[22] R.Venkatesan, R. Ganesan and A. Arul Lawrence Selvakumar, "A Comprehensive Study in Data Mining Frameworks for Intrusion Detection", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-4, Issue-7, December-2012.
[23] Sneha Kumari, Maneesh Shrivastava, "A Study Paper on IDS Attack Classification Using Various Data Mining Techniques", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5, September-2012.
[24] Vineet Richariya,Uday Pratap Singh, Renu Mishra,"Distributed Approach of Intrusion Detection System: Survey", International Journal of Advanced Computer Research (ISSN (IJACR) ,Volume-2,Number-4,Issue-6 December-2012.
[25] Deepak Rathore, Anurag Jain, "Design Hybrid method for intrusion detection using Ensemble cluster classification and SOM network", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5, September-2012.
[26] LI Yin–huan , "Design of Intrusion Detection Model Based on Data Mining Technology", International Conference on Industrial Control and Electronics Engineering, 2012.
[27] P. Prasenna,R. KrishnaKumar,A.V.T RaghavRamana and A. Devanbu "Network Programming And Mining Classifier For Intrusion Detection Using Probability Classification", Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012.
[28] Manish Somani, Roshni Dubey," Design of Intrusion Detection Model Based on FP-Growth and Dynamic Rule Generation with Clustering", International Journal of Advanced Computer Research (IJACR) Volume-3 Number-2 Issue-10 June-2013.