Intrusion Detection based on K-Means Clustering and Ant Colony Optimization: A Survey

Chetan Gupta
Department Of CSE
TIT Bhopal

Amit Sinhal
Department Of CSE
TIT Bhopal

Rachana Kamble
Department Of CSE
TIT Bhopal

ABSTRACT
Identifying intrusions is the process called intrusion detection. In simple manner the act of comprising a system is called intrusion. An intrusion detection system (IDS) inspects all inbound and outbound activity and identifies suspicious patterns that may indicate a system attack from someone attempting to compromise a system. If we think of the current scenario then several new intrusion that cannot be prevented by the previous algorithm, IDS is introduced to detect possible violations of a security policy by monitoring system activities and response in all times for betterment. If we uncover the counterfeit marque in a circumspect bulletin climate, an affirmation seat is initiated to prophesy or lessen the damage to the system. As a result it is a keen intrigue. In this dissertation we survey several aspects with the traditional techniques of intrusion detection we elaborate our proposed work. We also come with some future suggestions, which can provide a better way in this direction. For the above survey we also discuss K-Means and Ant Colony optimization (ACO).

Keywords
IDS, K-Means, ACO, Suspicious Patterns

1. INTRODUCTION
In recent years, many researchers are focusing to use data mining concepts for Intrusion Detection [1]. This is a process to extract the implicit information and knowledge.

Intrusion detection is the process of malicious attack in the system and network when we are in the process of communication or extracting data in the real time environment [2][3]. Since its invention, intrusion detection has been one of the key elements in achieving information security. It acts as the second-line defense which supplements the access controls. n a jiffy the tie defeated, the tumult revelation systems necessity be competent to determine it real-time and guide the anchor officers to take prompt and appropriate actions [3][4].

Intrusion detection system deal with supervising the incidents happening in computer system or network environments and examining them for signs of possible events, which are infringement or imminent threats to computer security, or standard security practices Intrusion detection systems (IDS) have emerged to detect actions which endanger the integrity, confidentiality or availability of are source as an effort to provide a solution to existing security issues [5].

So in the above directions we survey several aspects in the subsequent sections. We also discuss about association, K-Means Clustering and Ant Colony optimization techniques, because it can be used in forming the framework which produces better detection system.

The remaining of this paper is organized as follows. In Section 2 we discuss Literature Survey. In section 3 we discuss about the problem domain and analysis. In section 4 we discuss about data mining and optimization. In section 5 we discuss about the proposed work. The conclusions and future directions are given in Section 6. Finally references are given.

2. LITERATURE SURVEY
In 2010, G. Schaffrath et al. [11] 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 2011, Zhengjie Li et al. [12] propose a K-means clustering algorithm based on particle swarm optimization (PSO-KM). The proposed algorithm has overcome falling into local minima and has relatively good overall convergence. Experiments on data sets KDD CUP 99 has shown the effectiveness of the proposed method and also shows the method has higher detection rate and lower false detection rate.

In 2012, LI Yin–huan [13] 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. [14] 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 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 2011, LI Han [15] focuses on intrusion detection based on clustering analysis. The aim is to improve the detection rate and decrease the false alarm rate. A modified dynamic K-means algorithm called MDKM to detect anomaly activities is proposed and corresponding simulation experiments are presented. Firstly, the MDKM algorithm filters the noise and isolated points on the data set. Secondly by calculating the distances between all sample data points, they obtain the high-density parameters and cluster-partition parameters, using dynamic iterative process we get the k clustering center accurately, then an anomaly detection model is presented. They used KDD CUP 1999 data set to test the performance of the model. Their results show the system has a higher detection rate and a lower false alarm rate, it achieves expectant aim.

In 2011, Z. Muda et al. [16] discuss about the problem of current anomaly detection that it unable to detect all types of attacks correctly. To overcome this problem, they propose a hybrid learning approach through combination of K-Means clustering and Naïve Bayes classification. The proposed approach will be clustering all data into the corresponding group before applying a classifier for classification purpose. An experiment is carried out to evaluate the performance of the proposed approach using KDD Cup '99 dataset. Result show that the proposed approach performed better in term of accuracy, detection rate with reasonable false alarm rate.

3. PROBLEM DOMAIN

After discussing several research works we can come with some problem area in the traditional approaches which are following:

1. Need of Hybrid Intrusion Detection System which is better at detecting R2L and U2R attacks [16].
2. The IDS approach can be enhanced by providing more security to mobile agents [13].
3. Step Propagation is missing.
4. Neuro-Fuzzy Combination can be used as the distributed classifier.
5. All type of attacks is not well detected.
6. Maintain long log file for detection.
7. Triangle Area Nearest Neighbor (TANN) and K-Means with K-Nearest Neighbor (KMKNN) approach for better intrusion detection. This approaches showed a reasonable detection rate compare to our approach. Unfortunately, a potential drawback of this technique is the rate of false alarms [16][17].
8. In [18] Evolutionary Soft Computing based Intrusion Detection System (ESC-IDS) which focuses to detect and classify intrusion has proposed. This approach has serious shortcomings in its low accuracy rate as well as the tendency to produce high false alarm compare to [16].
9. Probability of less detection in U2R and R2L Detection technique so there is the need of a detection technique which improves in the hybridization of above two.

Analysis

After studying and observing several research works we compare the result discussions by their techniques, so that we identify the weak attack detection area.

Table 1: Analysis

S.no	Approach	Accuracy (%)	Precision (%)	Recall (%)
1	NB Training [16]	DOS 94.3	U2R 80	R2L 65.6
2	KM + NB Training [16]	DOS 99.5	U2R 40	R2L 61.6
3	NB Testing [16]	DOS 82.5	U2R 80	R2L 90.3
4	Km + NB Testing [16]	DOS 99.6	U2R 80	R2L 83.2
5	Rule based [15]	92.14	87.24	84.43
6	FSVM [15]	97.14	96.11	94.10
7	Rule based [15]	89.90	84.32	83.23
8	FSVM [15]	95.23	92.14	91.21
9	Rule based [15]	91.34	86.14	85.11
10	FSVM [15]	95.12	93.21	91.13
11	Rule based [15]	92.22	88.21	87.66
12	FSVM [15]	97.13	94.52	93.67

4. OPTIMIZED ASSOCIATION

Association rules mining are useful for identifying patterns which are needed, in our analysis it can be useful for finding patterns from the data set like KDD cup 99 dataset. It finds the search which is applicable for finding those global patterns which can be impact with search findings. There uses are analyzed in severa sectors like medical mining, rule optimization, market analysis etc.[19][20][21].

Association rule (AR) is commonly understood as an implication $X \rightarrow Y$ in a transaction database $D = \{t1 \ldots\ldots\ldots tm\}$[22].

Each transaction $ti \in D$ contains a subset of items $I = \{i1.....in\}$. X and Y are disjoint itemsets, it holds $X; Y \subseteq I$ and $X \cap Y = \Phi$. The cuff allot side of this implication is called antecedent; the right hand side is referred to as consequent [23]. The pact database D in addition be looked on as a boolean dataset hoop the boolean rationalism of presentation in accounts express occurrence of items in transactions [23].

Association rule mining problem poses the question of efficiency. The number of potential rules $X \rightarrow Y$ defined by $X \subseteq Ix \in \{ix1;\ldots\ldots ; ixn\}$, $Y \subseteq Iy \in \{iy1;\ldots. ; iym\}$[23], where Ix and Iy are disjoint, is equal to $2(m+n)$[22]. In a jiffy commonplace datasets are steady; the association rule mining problem is known to be NP-complete [24]. In restricted cases, for example in sparse boolean datasets (where it holds all ti

∈ D; |ti| <= O(log|I|)) lower complexity bounds have been proved to hold. It also improves the efficiency [25]. It can icrease subsequent number of associated values. [6] and [7].

For categorization several traditional technique used 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.

The Ant Colony Optimization algorithm is mainly inspired by the experiments run by Goss et al. [8] which using a grouping of real ants in the real environment. This ant behavior was first formulated and arranged as Ant System (AS) by Dorigo et al. [9][10].They study and observe the behavior of those real ants and suggest that the real ants were able to select the shortest path between their nest and food resource, in the existence of alternate paths between the two. When ants are travelling for the food resources, ants deposit a chemical substance, called pheromone, on the ground. When they arrive at a destination point, ants make a probability based choice, biased by the intensity of pheromone they smell. This behavior has an autocatalytic effect because of the very fact that an ant choosing a path will increase the probability that the corresponding path will be chosen again by other ants in the next move. By using the above behavior we can increase the detection probability by improving the pheromone substances.

5. PROPOSED WORK

The below picture shows the working phenomena of our approach. In our working process first we consider a database like KDD Cup 1999 Data Set [26]. This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99[27]. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

We apply K-Means clustering as a pre classifier component. It can make clusters on the basis of Denial of service (DoS) attack; gather information on a potential target which maps the details of the system about the network which is probe. Then remote 2local attack (R2L) and user to root (U2R). But some behavior an intrusion instances are similar to normal and other intrusion instance as well. So K-Means clustering techniques are unable to correctly distinguish intrusion instances.

So we want to improve the shortcoming in the classifier and we apply the combination of Association Rule classifier with K-Means. K-Means works on k clusters for example if we consider k=5 clusters means it can be like the following:

C1=Normal	C2=Dos	C3=Probe
C4=R2L	C5=U2R	

Basis of the above we can maintain a log file so that proper matching of the shortcomings can be performed.

Then we want to increase the detection probability, so our objective function can be like:
∑d1+d2+……dn

It can be making the agents as an ant *k which* can be assigned a start state s^k_s and more than one termination conditions e^k. Ants start from a start state and move to feasible neighbor states, building the solution in an incremental way. The procedure stops when at least one termination condition e^k for ant *k* is satisfied.

An ant *k* in state $sr =< s_{r-1}; i >$ can move to any node *j* in its feasible neighborhood N^k_i , defined as $N^k_i = \{j \mid (j \in Ni) \wedge (< sr, j > \in S)\}$ $sr \in S$, with *S* is a set of all states. When moving from node *i* to neighbor node *j*, the ant can update the pheromone trails τ_{ij} on the edge (*i, j*).
Once it has built a solution, an ant can retrace the same path backward, update the pheromone trails and die.
The trail intensity is determined by the below formula:

$$T_{ij}(t + n) = p \cdot T_{ij}(t) + \Delta T_{ij}$$

And the next move will be determine

$$\Delta T^k_{ij} = \begin{cases} \dfrac{Q}{L_k} & \text{if the kth ant uses edge}(i, j) \text{ in its tour} \\ & (\text{between time } t \text{ and } t+n) \\ 0 & \text{otherwise} \end{cases}$$

And then we concentrate on the group highest value, so that we increase the prediction rate.

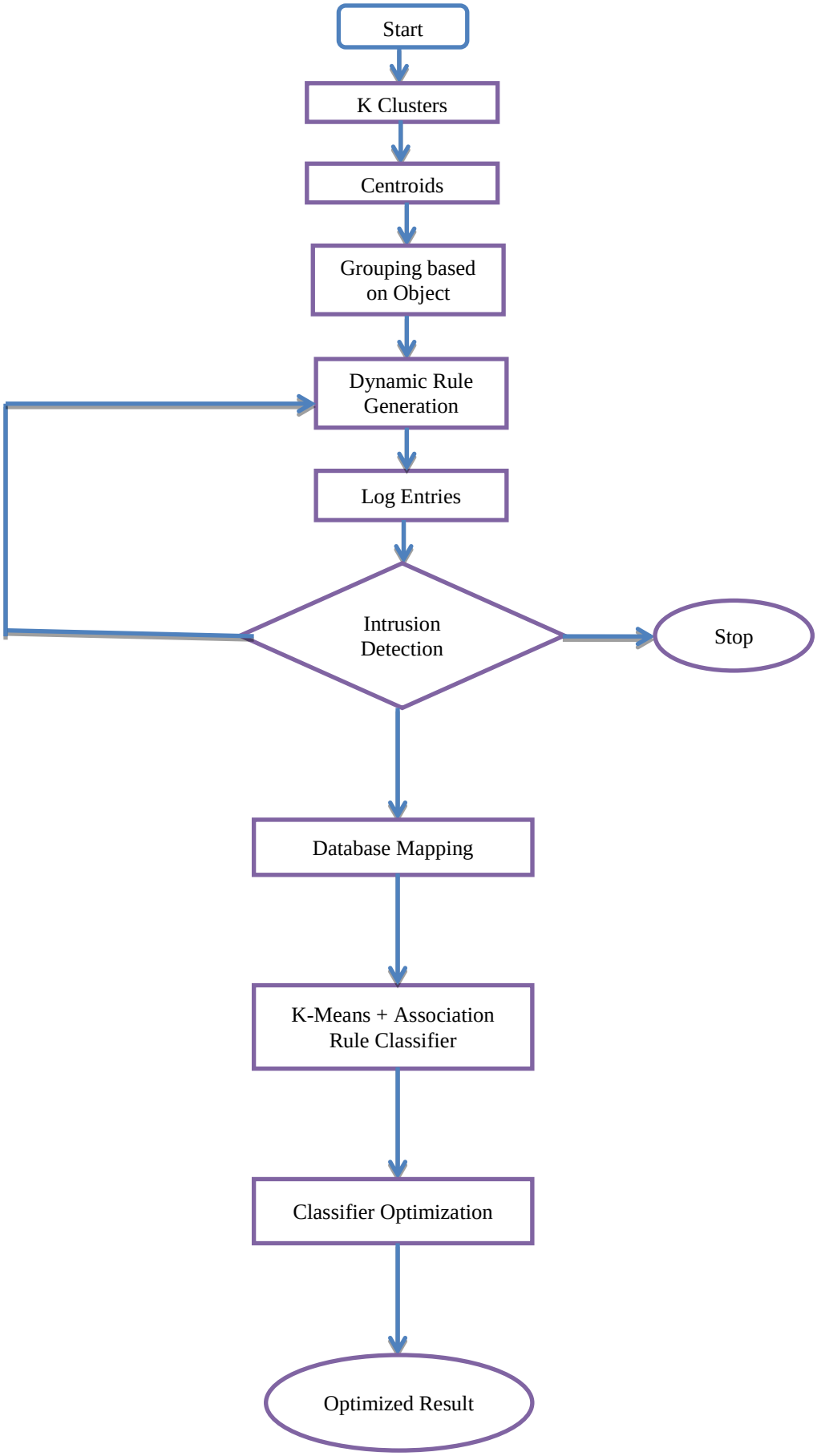

Figure 1: Working Flowchart

6. CONCLUSION AND FUTURE WORK

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems. Intrusion detection is an area growing in relevance as more and more sensitive data are stored and processed in networked system. After analyzing several research works in this direction, we come with several advantages and disadvantages. We analyze that there are several techniques which provide good detection rate in the case of Denial of Service (DOS) attack. But fail to achieve good detection rate in the case of U2R and R2L attacks. Based on the above study we provide the following future directions which can be helpful in better detection:

1. A data Mining technique which dynamic rule association can be fruitful.
2. Combination K-Means with Optimization can increase the pattern recognition.
3. Ant Colony Optimization can help in better pattern detection.
4. The misuse detection approach better at detecting R2L and U2R attacks more efficiently as well as anomaly detection approach, which is better at detecting attacks at the absence of match signatures as provided in the misuse rule files[16].
5. Hybridization of Association, K-Means and Optimization can provide better detection.

7. REFERENCES

[1] Meng Jianliang,Shang Haikun,Bian Ling," The Application on Intrusion Detection Based on K-means Cluster Algorithm", International Forum on Information Technology and Applications, 2009.

[2] Lundin, E. and Jonsson, E. "Survey of research in the intrusion detection area", Technical Report, Department of Computer Engineering, Chalmers University of Technology, Göteborg, Sweden. January 2002.

[3] Li Tian,Wang Jianwen," Research on Network Intrusion Detection System Based on Improved K-means Clustering Algorithm", International Forum on Computer Science-Technology and Applications, 2009.

[4] S.Devaraju, S.Ramakrishnan:,"Analysis of Intrusion Detection System Using Various Neural Network classifiers, IEEE 2011.

[5] Moriteru Ishida, Hiroki Takakura and Yasuo Okabe," High-Performance Intrusion Detection Using OptiGrid Clustering and Grid-based Labelling", IEEE/IPSJ International Symposium on Applications and the Internet, 2011.

[6] Prakash Ranganathan, Juan Li, Kendall Nygard, "A Multiagent System using Associate Rule Mining (ARM), a collaborative filtering approach", IEEE 2010, pp- v7 574- 578.

[7] Prof Thivakaran.T.K, Rajesh.N, Yamuna.P, Prem Kumar.G, "Probable Sequence Determination Using Incremental Association Rule Mining And Transaction Clustering", IEEE 2009, pp 37-41.

[8] S. Goss, S. Aron, J. L. Deneubourg, and J. M. Pasteels. "Self-organized Shorcuts in the Argentine Ant." Naturwissenschaften, 76:579–581, 1989.

[9] M. Dorigo, Gianni Di Caro, and Luca M. Gambardella. "Ant Algorithms for Discrete Optimization." Technical Report Tech. Rep. IRIDIA/98-10, IRIDIA, Universite Libre de Bruxelles, Brussels, Belgium, 1998.

[10] M. Dorigo and M. Maniezzo and A. Colorni. "The Ant Systems: An Autocatalytic Optimizing Process." Revised 91-016, Dept. of Electronica, Milan Polytechnic, 1991.

[11] G.Schaffrath,R. Sadre,C. Morariu,A.Pras and B.Stiller, "An Overview of IP Flow-Based Intrusion Detection", Communications Surveys & Tutorials, IEEE 2010.

[12] Zhengjie Li, Yongzhong Li , Lei Xu," Anomaly Intrusion Detection Method Based on K-means Clustering Algorithm with Particle Swarm Optimization", International Conference of Information Technology, Computer Engineering and Management Sciences,2011.

[13] LI Yin–huan, "Design of Intrusion Detection Model Based on Data Mining Technology", International Conference on Industrial Control and Electronics Engineering, 2012.

[14] P. Prasenna, R. Krishna Kumar, A.V.T Raghav Ramana and A. Devanbu "Network Programming And Mining Classifier For Intrusion Detection Using Probability Classification", Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012.

[15] LI Han,"Using a Dynamic K-means Algorithm to Detect Anomaly Activities", Seventh International Conference on Computational Intelligence and Security, 2011.

[16] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir," Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification", 7th International Conference on IT in Asia (CITA), 2011.

[17] C. F. Tsai, and C.Y Lin, "A triangle area-based nearest neighbors approach to intrusion detection," Pattern Recognition, 2010, 43(1):222-229.

[18] C. Xiang, P.C. Yong, and L.S. Meng, "Design of multiple level hybrid classifier for intrusion detection system using Bayesian clustering and decision tree," Pattern Recognition Letters, 2008, 29: 918-924.

[19] Pragati Shrivastava, Hitesh Gupta," A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (IJACR), Volume-2 Number-3 Issue-5 September-2012.

[20] Anshuman Singh Sadh, Nitin Shukla," Association Rules Optimization: A Survey", International Journal of Advanced Computer Research (IJACR), Volume-3 Number-1 Issue-9 March-2013.

[21] Mr. Sachin sohra, Mr. Narendra Rathod," An Improved Single and Multiple association Approach for Mining Medical Databases", International Journal of Advanced Computer Research (IJACR) Volume 2 Number 2 June 2012.

[22] 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.

[23] Shantakumar B.Patil, Dr.Y.S.Kumaraswamy," Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, February 2009.

[24] Parag Deoskar, Dr. Divakar Singh, Dr. Anju Singh," Mining Lung Cancer Data and Other Diseases Data Using Data Mining Techniques: A Survey", International Journal of Computer Engineering and Technology (IJCET), Volume 4, Issue 2, March – April (2013).

[25] Anshuman Singh Sadh, Nitin Shukla," Apriori and Ant Colony Optimization of Association Rules", International Journal of Advanced Computer Research (IJACR),Volume-3 Number-2 Issue-10 June-2013.

[26] Alexander O. Tarakanov, Sergei V. Kvachev, Alexander V. Sukhorukov ," A Formal Immune Network and Its Implementation for On-line Intrusion Detection", Lecture Notes in Computer Science Volume 3685, 2005, pp 394-405.

[27] Foster Provost," Machine Learning from Imbalanced Data Sets 101", AAAI Technical Report WS-00-05. Compilation copyright © 2000, AAAI (www.aaai.org).