Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-7, 2016
ISSN: 2454-1362, http://www.onlinejournal.in

Analysis of KDD CUP Dataset using Big Data Approach for Anomaly Detection Over Network Traffic

Namrata Verma & Dr Nitin Mishra
[1]Computer Science And Engineering, Rungta College Of Engg And Technology Bhilai, Chhattisgarh
[2]Assistant Professor, Cse Department, Rungta College Of Engg And Technology Bhilai, Chhattisgarh

Abstract*— In this paper, Intrusion recognition is to find assaults (Intrusions) contrary to a tablet machine. Inside the fabulously organized present day universal, customary procedures of group security which incorporates cryptography, customer verification and interruption counteractive action strategies like firewalls are insufficient to discover new strikes. In this paper, we perform tests at the kddcup99 realities set. The KDD Cup 99 dataset has been the purpose of fascination for some specialists in the field of interruption recognition from the most recent decade. Numerous analysts have contributed their endeavours to break down the dataset by various methods. Investigation can be utilized as a part of an industry that produces and expends information, obviously that incorporates security. We do dimensionality diminishment of the truths set the utilization of PCA (most imperative thing assessment) and clear distinction among ordinary and irregular measurements is resolved with the guide of the utilization of managed insights mining procedures. Extensively explores different avenues regarding kddcup99 system insights demonstrate that the directed procedures comprehensive of Guileless Bayesian, C4.5 can practically identify bizarre attacks and procure a low false gigantic rate. In this proposal streamlining system together with Random woods has completed to enhance the effectiveness of identification cost and accomplish a low fake fine rate. This component can adequately endure interruption.*

Key words: Big Data, intrusion detection, PCA, KDD cup, Network traffic etc.

I. INTRODUCTION (*HEADING 1*)

As people group essentially based pc frameworks have basic parts in cutting edge society, they have end up being the objectives of interlopers. Consequently, we need to manufacture the immense practical rules to shield our structures. The security of a portable workstation framework is inclined when an interruption takes region. An interruption might be characterized as any movement completed that damages the trustworthiness, secrecy or accessibility of the machine. There are a couple interruption counteractive action methodologies which might be utilized to spare you PC structures as a first line of safeguard. A firewall is likewise one in all it. In any case, handiest interruption counteractive action is not adequate. As frameworks develop to be more prominent complex, there are always exploitable shortcomings in the frameworks as a result of outline and programming botches, or various entrance systems. In this manner Intrusion identification is required as each other degree to shield our pc frameworks from such kind of vulnerabilities.

In 1999, the tcp empty archives have been assembled which comprises of the 41 qualities at MIT Lincoln Laboratory for 1998 DARPA Intrusion Detection assessment programming and it has been utilized as KDD cup'99 records.KDD cup'ninety nine records is turned out to be proper referencing actualities for security concentrates on group and for records mining examines region as well.As it's miles a major dataset, it has some flaws in it regardless of the way that numerous specialist have utilized this dataset for demonstrating their studies included works Log reports are realities containing posting of events and exercises that emerge inside the framework. They're produced when exercises happen inside the framework. In tremendous scale structures, which incorporates administered structures, groups and matrix structures, enormous measure of log insights are produced in real time. Their humungous size is one in every one of the vulnerabilities of the gadget because of the reality standalone log document analyzers, likewise called interruption location frameworks, can't look at all of them. In spite of the way that standalone log analyzers can look at all log records, the results won't not be exact.

Despite the fact that, standalone log analyzers can't break down enormous measure of log records, we will by the by make utilization of those log archives in a follow examination. Follow returned is an examination now not a counteractive action technique. Accordingly, assessment of log archives might be additional advantageous than insight back. On this paper, we watch Hadoop, that is a structure for the dispensed gadget for huge scale log investigation. The log records we utilized are KDD'99 [1] data set. KDD'ninety nine is a data set inside the subject of interruption identification gadget. KDD'ninety nine records set is an expansive scale log documents, it has roughly 5 million sections. Also, KDD'ninety nine is by and large utilized as a part of numerous interruption discovery thinks about.

2. RELATED WORK

Denning was amongst the essential individuals to assume inside the range of utility of insights mining to network assurance. He has given a model of a genuine –time interruption discovery master gadget [1]. The thought behind the model is that abuse of a device's vulnerabilities incorporates strange use of gadget and this variation from the norm might be identified by method for looking out the odd examples in the review records. The model proposed can distinguishing harm ins, infiltrations, and different assortments of PC anomaly.In this paper we're the utilization of two systems of irregularity identification SVM (bolster Vector contraption) and C4.Five this is drawn out adaptation of classification calculation ID3. Each the strategies are managed set of standards. We're performing evaluation on the fundamental of location charge and false caution rate.

In [2], an accumulation of scientists proposed blended interruption recognition gadget which incorporates abuse and inconsistency identification. k-way calculation is actualized for the abnormality identification module. In their investigations, they produce four examination insights sets. Each test insights set has two thousand and one hundred passages. Every realities set comprises of two thousand passages of typical records. The unwinding is interruption certainties. Their outcomes demonstrate that alright means set of standards has high recognition expense for an unmarried interruption and has low identification cost for numerous interruptions. In correlation, KD set of principles, which is their propelled k-implies set of tenets has high discovery expense for both unmarried interruption or different interruptions. Be that as it may, their test insights set is best thousand and one hundred sections. Rupali Datti et. Al. [3] proposed Linear Discriminant assessment (LDA) to decrease highlights on the NSL-KDD dataset to four capacities best this offers ninety seven% markdown inside the enter records and around ninety four% rebate inside the preparation time. Shilpa Lakhina et. Al. [4] proposed vital thing examination (PCA) as a diminishment gadget, it decreases the elements to 8 highlights this offers eighty.4% records rebate also, around 35%-40% lessening inside the training time and 75%-eighty% markdown in the testing time. Neveen I. Ghali [5] proposed harsh set hypothesis (RST) to pick abilities of the KDD99 dataset. Just 7 elements are chosen bringing about 83% lessening inside the enter records and eighty five%-90% time markdown and roughly 90% diminishment in mean squared botches in distinguishing new attacks.

3. PROPOSED METHODOLOGY

3.1 METHODOLOGY

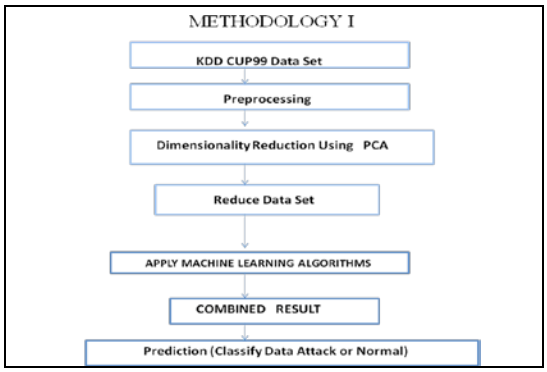

Figure 1 Methodology -I

3.2 METHODOLOGY II

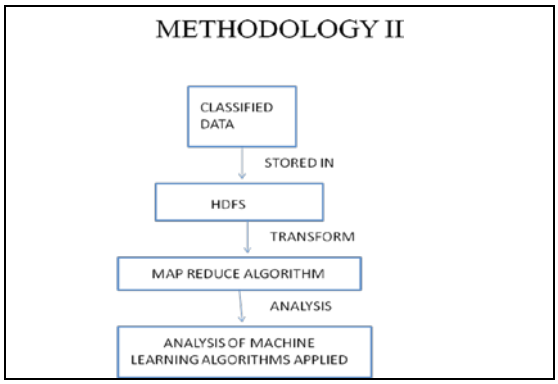

Figure 2 Methodology -II

4. CLASSIFIER FOR ATTACK DETECTION

4.1 Principal Component Analysis(PCA):

PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension. The entire subject of statistics is based on around the idea that you have this big set of data, and you want

to analyze that set terms of the relationships between the individual points in that set [3-4].

Algorithm:

Suppose x1, x2,…….xm are N x 1 vectors

Step 1: find mean of input matrix

Step 2: Normalize data by subtracting mean

Step 3: find covariance matrix C=cov(y)

Step 4: compute the eigenvectors (V) or eigenvalues (D) of the covariance matrix [V,D]=eig(C)

Step5: sort eigenvalues in descending order by first diagonalising eigen value matrix, idx stores order to use when ordering eigenvectors.

Step6: put eigenvectors in order corresponding with eigen values.

Step 7: (dimensionality reduction step) keep only the terms corresponding to the k largest eigenvalues where k=no of dimensions after reduction of data set

How to choose the principal components:- To choose k, use the following criterion.

4.2 Naive Bayesian Classifier(NBC)

Bayesian classifiers are factual classifiers. They can foresee class enrollment probabilities, for example, the likelihood that a given specimen has a place with a specific class. Bayesian characterization depends on Bayes hypothesis. Considers contrasting order calculations have found a straightforward Bayesian classifier known as the innocent Bayesian classifier to be equivalent in execution with choice tree and neural system classifiers. Bayesian classifiers have additionally displayed high exactness and velocity when connected to expansive databases.[5-6].

Innocent Bayesian classifiers expect that the impact of a quality estimation of a given class is autonomous of the estimations of alternate characteristics. This presumption is called class contingent freedom. It is made to disentangle the calculations included and, in this sense, is viewed as "guileless".

4.3 C4.5 (Decision Tree)

C4.5 is a suite of calculations for arrangement issues in machine learning and information mining. It is focused at managed learning: Given a property estimation information set where cases are portrayed by accumulations of ascribes and have a place with one of an arrangement of totally unrelated classes, C4.5 takes in a mapping from credit qualities to classes that can be connected to characterize new, concealed instances.[21] All tree prompting techniques start with a root hub that speaks to the whole, given information set and recursively split the information into littler subsets by testing for a given characteristic of every hub. The sub trees indicate the segments of the first dataset that fulfill indicated quality worth tests. This procedure commonly proceeds until the subsets are ""immaculate"" that is, all examples in the subset fall in the same class, at which time the tree developing is terminated[18]. This calculation can be utilized to create a choice tree that can be utilized to characterize information occurrences in various classes which helps in further investigation recognizing legitimate results. This calculation made various enhancements on ID3 calculation. These are: It can deal with both ceaseless and discrete qualities. For nonstop qualities, it makes a limit and afterward parts the list into those whose quality worth is over the edge and those that are not exactly or equivalent to it[7-8].

Missing quality qualities are just not utilized as a part of increase and entropy estimations. It can deal with traits with various expenses. Proposed Algorithm steps are as follows:

Input: an attribute-valued dataset D (after apply Dimensionality reduction method PCA)

1: Tree = {}

2: if D is “pure” OR other stopping criteria met then

3: terminate

4: end if

5: for all attribute a € D do

6: Compute information-theoretic criteria if we split on a

7: end for

8: abest = Best attribute according to above computed criteria

9: Tree = Create a decision node that tests abest in the root

10: Dv = Induced sub-datasets from D based on abest

11: for all Dv do

12: Treev = C4.5(Dv)

13: Attach Treev to the corresponding branch of Tree

14: end for

15: return Tree

4.4 BAGGING & BOOSTING

Bootstrap aggregating (bagging) and boosting are useful techniques to improve the predictive performance of tree models. Boosting may also be

useful in connection with many other models, e.g. for additive models with high-dimensional predictors; whereas bagging is most prominent for improving tree algorithms[9-10].

5. EXPERIMENTS

On this paper our work is tested using the 1999 KDD cup community anomaly facts set [19]. It originated from the 1998 DARPA Intrusion Detection assessment software controlled by means of MIT Lincoln Labs.

The first level is pre-processing. Data on this segment is decreased to lower dimensionality (18 attributes)then partition into training and checking out. Inside the next step, we implemented C4.Five and NBC at the training dataset with the intention to build and train the fashions. Eventually trained models are evaluated at the checking out dataset to calculate the efficiency of the fashions. The education statistics set consists of seven weeks of visitors with round 5 million connections and the trying out data consists of two weeks of site visitors with round 300,000 connections. The records includes 4 primary classes of attacks:

1 Denial-of-service (Dos) such as smurf, apache2, pod, etc.

2. Remote-to-local (R2L) like imap, worm, phf, etc..

3. User to root (U2R) such as perl, rootkit and so on.

4. PROBING such as nmap, portsweep, etc.

Mining algorithms can lead to better results if data under analysis have been normalized [6]. Detection of attack can be measured by following metrics:

1 False positive (FP): Or false alarm, Corresponds to the number of detected attacks but it is in fact normal.

2. False negative (FN): Corresponds to the number of detected normal instances but it is actually attack, in other words these attacks are the target of intrusion detection systems.

3. True positive (TP): Corresponds to the number of detected attacks and it is in fact attack.

4. True negative (TN): Corresponds to the number of detected normal instances and it is actually normal.

The accuracy of an intrusion detection system is measured regarding to detection rate and false alarm rate. In this work, we use 1999 KDD cup Dataset which consist of (3500 records). Table 1 given below shows the percentage of the data. Then, 15% of the data is extracted by sampling. 70% of this new set belonged to training set, and 40% dedicated to test data.

BIG DATA

Enormous information investigation is the procedure of inspecting substantial information sets containing an assortment of information sorts - i.e., huge information - to reveal concealed examples, obscure relationships, market patterns, client inclinations and other helpful business data. The investigative discoveries can prompt more powerful advertising, new income opportunities, better client administration, enhanced operational productivity, upper hands over opponent associations and different business advantages[11-12].

The essential objective of huge information investigation is to help organizations settle on more educated business choices by empowering information researchers, prescient modelers and different examination experts to break down huge volumes of exchange information, and additionally different types of information that might be undiscovered by ordinary business knowledge (BI) programs. That could incorporate Web server logs and Internet clickstream information, online networking substance and interpersonal organization movement reports, content from client messages and overview reactions, cellular telephone call point of interest records and machine information caught by sensors associated with the Internet of Things. Some individuals only partner huge information with semi-organized and unstructured information of that sort, however counseling firms like Gartner Inc. what's more, Forrester Research Inc. additionally consider exchanges and other organized information to be legitimate parts of enormous information examination applications[13-14].

Huge information can be dissected with the product devices ordinarily utilized as a feature of cutting edge examination trains, for example, prescient investigation, information mining, content investigation and measurable examination. Standard BI programming and information perception devices can likewise assume a part in the investigation procedure. Be that as it may, the semi-organized and unstructured information may not fit well in conventional information stockrooms taking into account social databases. Moreover, information stockrooms will be unable to handle the preparing requests postured by sets of huge information that should be overhauled as often as possible or even ceaselessly - for instance, constant information on the execution of versatile applications or of oil and gas pipelines. Accordingly, numerous associations hoping to gather, prepare and investigate huge information have swung to a more up to date class of advancements that incorporates Hadoop and related instruments, for example, YARN, MapReduce, Spark, Hive and Pig and additionally NoSQL

databases. Those advancements shape the center of an open source programming structure that backings the handling of vast and assorted information sets crosswise over bunched frameworks[15].

BIG DATA Experimental Setup

For our research we're going to be used check pattern statistics as mashable on-line news records to be had in mashable.Com. It is freely to be had for check and studies. For writing java program we're using notepad++ v6.Nine., Java improvement package version is JDK 1.7 for java environment and hadoop 2.3 for windows, and windows eight.1 operating system. Right here in this web website "http://www.Codeproject.Com/Articles/757934/Apache-Hadoop-for-home windows-Platform" the installation manner is given, observe the ones steps for setting up hadoop environment16.19.21].

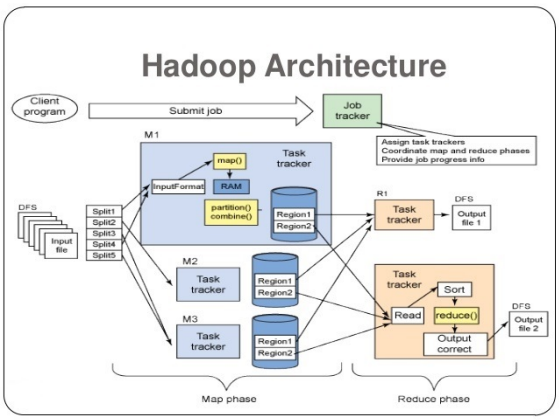

Figure 3 Hadoop Architecture

Open cmd prompt in admin mode and start hadoop demon using F:\hadoop\Hadoop-2.3-master\Hadoop-2.3-master\sbin\start-yarn command snap shot is follows in figure 4.After this start dfs by using F:\hadoop\Hadoop-2.3-master\Hadoop-2.3-master\sbin\start-dfs command the snapshot in figure 5.

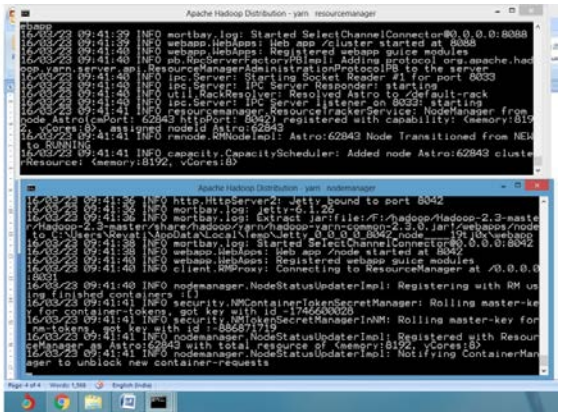

Figure 4 Yarn Hadoop

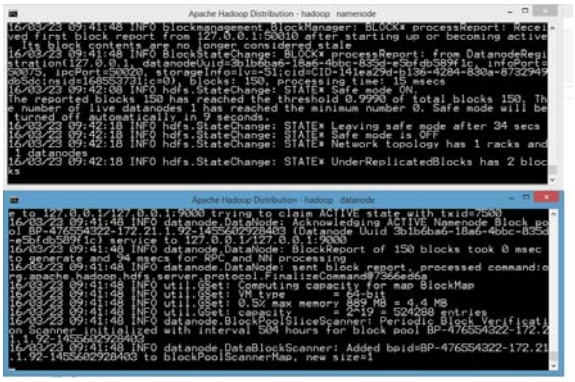

Figure 5 Hadoop HDFS

After doing this we are now going to concentrate the actual work that is our KDD cup analysis. Using weka and principal component analysis we are reduce the dimension of our main data set refer below figure 6.

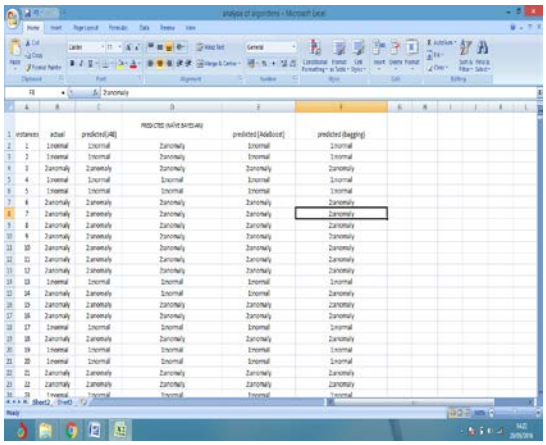

Figure 6 Sample data set for our work

Now the final data set is ready for our final analysis. We are analyzing theses data with the ig data hadoop. Here we have 4 set of program that will compare all four column that is 3,4,5,6 with 2nd column and will give result of FP, FN, TP,TN against the actual column values.

CONFUSION MATRIX

1. J48

NORMAL/ANOMALY	NORMAL	ANOMALY
NORMAL	37	3
ANOMALY	3	46

2. NAÏVE BAYESIAN

NORMAL/ANOMALY	NORMAL	ANOMALY
NORMAL	32	8
ANOMALY	7	42

3. ADABOOSTING

NORMAL/ANOMALY	NORMAL	ANOMALY
NORMAL	35	5
ANOMALY	1	48

4. BAGGING

NORMAL/ANOMALY	NORMAL	ANOMALY
NORMAL	40	0
ANOMALY	5	44

5. Proposed method

NORMAL/ANOMALY	NORMAL	ANOMALY
NORMAL	37	6
ANOMALY	5	45

COMPARISION OF THE ALGORITHMS

-	ACCURACY	TPR	FPR	DETECTION RATE
J48	93%	92.5%	0.061	44%
NAÏVE BAYESIAN	83%	82%	0.16	43%
ADABOOST	93%	97%	0.094	42%
BAGGING	92%	88%	0.043	47%
Combination(j48+nb+adaboost +bagging)	93.2%	86%	0.025	48%

CONCLUSION

In the event that any one attempt to utilize the system then identifying assault is an essential need in system frameworks, in this paper information mining strategies specifically NAÏVE BAYESIAN, J48 , BAGGING and BOOSTING are utilized to recognize oddity in the system. Test results appear, packing calculation has better results in both location rate and exactness in our information set. There are some difficulties confronted by the IDS. Like other managed learning calculations, the new sort of assault can't be effectively found by these IDS. In the event that new assault is found in the testing information it is recognized as a typical information nonetheless, clients' practices change every now and then. The static preparing information may get to be obsolete and inadequate for expectation as time passes by.

References

[1] Dorothy E. Denning. “An Intrusion-Detetcion Model” 1986 IEEE Computer Society Symposium on Research in Security and Privacy ,pp 118-31

[2] C. Zhang, G. Zhang, and S. Sun. A mixed unsupervised clustering-based intrusion detection model. In Genetic and Evolutionary Computing, 2009. WGEC ’09. 3rd International Conference on, pages 426 –428,oct. 2009.

[3] Rupali Datti,Bhupendra Verma,"Feature Reduction for Intrusion Detection Using Linear Discriminant Analysis." (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 1072-1078.

[4] Shilpa Lakhina, Sini Joseph and Bhupendra Verma, “Feature reduction using using Principal Component Analysis for Effective Anomaly Based Intrusion Detection on NSL-KDD”, Int. J.of engineering science and technology, Vol. 2(6),2010, 1790-1799.

[5] Neveen I. Ghali, "Feature selection for effective anomaly based intrusion detection." IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.3, March 2009.

[6] J. Han, and M. Kamber, “Data mining: concepts and techniques’” (2nd ed.). Morgan Kaufmann Publishers, 2006.

[7] M. S. Abadeh and J. Habibi, “Computer Intrusion Detection Using an Iterative Fuzzy Rule Learning Approach,” *Proceedings of the IEEE International Conference on Fuzzy Systems*, London, 2007, pp. 1-6.

[8] B. Shanmugam and N. BashahIdris, “Improved Intrusion Detection System Using Fuzzy Logic for Detecting Anamoly and Misuse Type of Attacks,” *Proceedings of the International Conference of Soft Computing and Pattern Recognition*, 2009, pp. 212-217.

[9] J. T. Yao, S. L. Zhao and L. V. Saxton, “A study on Fuzzy Intrusion Detection,” *Proceedings of Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security*, 2005, pp. 23-30.

[10] Q. Wang and V. Megalooikonomou, “A Clustering Algorithm for Intrusion Detection,” *Proceedings of the Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security*, Vol. 5812, 2005, pp. 31-38.

[11] E. Oja, ”Principal components, minor components, and linear neural networks” Neural Networks, vol. 5, pp. 927-935, 1992.

[12] G.K. Kuchimanchi, V.V. Phoha, K.S. Balagami and S.R. Gaddam, ”Dimension reduction using feature extraction methods for Real-time misuse detection systems” In proceedings of the IEEE Workshop on Information Assurance and Security, West Point, (New York), pp. 195-202, 2004.

[13] K. Labib and V.R. Vemuri, ”Detecting and visualizing denialofservice and network probe attacks using principal component analysis” In Third Conference on Security and Network [14] L. Y. Chuang, S. W. Tsai, C.H. Yang, "Catfish Binary Particle Swarm Optimization for Feature Selection" *International Conference on Machine Learning and Computing, IPCSIT, IACSIT Press, Singapore, 3: 40-44, 2011.*Architectures, La Londe, (France), 2004.

[15] D. P. Rini, S. M. Shamsuddin, S. S. Yuhaniz, " Particle Swarm Optimization: Technique, System and Challenges," international Journal
of Computer Applications, 14(1): 0975 – 8887, 2011.

[16] R. Parimala, R. Nallaswamy, "Feature selection using a novel particle swarm optimization and its variants," *I.J. Information Technology and Computer Science*, 5: 16-24, 2012.

[17] R.Mendes, J.Kennedy, J. Neves," The fully informed particle swarm: Simpler, maybe better," IEEE Transactions on Evolutionary
Computation 8(3) 204 – 210, 2004.

[18] Parsopoulos KE, Plagianakos VP, Magoulas GD and Vrahatis MN,“Streching Technique for Obtaining Gloabal Minimizers ThroughParticle Swarm Optimization,” In:

Proceedings Particle SwarmOptimization Workshop: 22–29, 2001.

[19] Kennedy J, Eberhart RC, "Particle Swarm Optimization," In:Proceedings of the IEEE Int. Conf. Neural Networks: 1942–1948, 1995.

[20] Chen Guolong, Chen Qingliang and Guo Wenzhong, "A PSO-BasedApproach to Rule Learning in Network Intrusion Detection," FuzzyInformation and Engineering (ICFIE), ASC 40, pp. 666–673, 2007.

[21] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57:289–300, 1995.

[22] A. Bronstein, J. Das, M. Duro, R. Friedrich, G. Kleyner, M. Mueller, S. Singhal, and I. Cohen. Bayesian networks for detecting anomalies in internet-based services. In Intl. Symposium on Integrated Network Mgmt., 2001.

[23] K. Das and J. Schneider. Detecting anomalous records in categorical datasets. In Proc. ACM Knowledge Discovery and Data Mining, Aug 2007.G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)