ARPN Journal of Engineering and Applied Sciences
©2006-2015 Asian Research Publishing Network (ARPN). All rights reserved.

www.arpnjournals.com

AN INTELLIGENT WEIGHTED OUTLIER DETECTION METHOD FOR INTRUSION DETECTION USING MST AND K-NN

R. Selvi[1] and S. Saravan Kumar[2] and A. Suresh[3]
[1]St.Peter's University, Chennai, India
[2]Panimalar Institute of Technology, Chennai, India
[3]S. A. Engineering College, Chennai, India
E-Mail: ss12.selvi@gmail.com

ABSTRACT

Intrusion Detection System (IDS) is a potential part in the area of network security system. An effective intrusion detection system is necessary for providing effective communications in the past world. The major challenging task in this system is the classification of users such as normal user and attacker. For that purpose so many classification algorithms have been proposed in the past to detect and report about user's behaviour (normal or abnormal) in networks. The sufficient detection accuracy is not yet met due to the lacking of suitable methodology introduction in this field. Outlier detection is the effective process to improve the classification performance. In the past, many outlier detection methods with the combination of different clustering methods have been proposed. These all are have limitations in terms of accuracy and speed. In this paper, we propose a new outlier detection model called Intelligent Weighted MST and k-NN Based Outlier Detection (IWMKOD) to detect the intruders in all kinds of network environment. This model is the combination of the proposed Intelligent MST and kNN based Outlier Detection. [Xiao Chun Wang *et al*., 2015] and Weighted Distance based Outlier Detection **[**S. Ganapathy *et al*., 2011**]**. The experimental results show that the proposed algorithm improves the detection accuracy and reduces the false alarm rate.

Keywords: weighted MST, K-NN, outlier detection, intrusion detection system.

INTRODUCTION

Intrusions refer to the actions against the valuable resource and try to negotiate the integrity, confidentiality of system resource [Xiao Chun Wang *et al*., 2015]. Intruders always act as a person and attempting to misuse the system in violation of an established policy. Major goal of intruders are being denied the important services, system failing to respond or delaying, data being lost or stolen. Intrusion detection means detecting unauthorized use of a system or attacks on a system or network. An Intrusion Detection System (IDS) monitors and restricts user access to the computer system by applying certain rules. IDSs are classified into misuse and anomaly IDS. Misuse detection is a set of events that match with predefined pattern of a known attack. The effectiveness of misuse IDS is largely based on the validity and expressiveness of their database of known attacks and misuse, and the efficiency of the matching engine that is used. The disadvantage of misuse IDS is that it requires frequent updates to keep up with the new stream of vulnerabilities discovered and it cannot detect unknown attacks. Anomaly detection is an intrusion in which the behaviour differs from the normal user behaviour. The biggest challenge of IDS is the detection accuracy and false alarm rate [Xiao Chun Wang *et al*., 2015].

Outlier detection is a major task in many critical applications and environments such as social network and all kind of networks. The outlier is representing the abnormal behaviour nodes or users in any network environments. This running condition is from which major performance lack may result. An outlier denotes an anomaly user in a network and anomaly data or record in a distributed databases and mobile databases. An outlier specifically points an intruder inside a system with malicious objectives quickly. Outlier detection achieves this by analysing and comparing the time series of usage statistics. Another set of functions for which outlier is used are fraud detection, intrusion detection, activity monitoring, network performance measurement, time series monitoring, detecting unexpected entries and detecting mislabelled entries [S. Ganapathy *et al*., 2011].

State-of-the-art K-nearest neighbours (k-NN) based outlier detection algorithms, such as various distance and density based methods and also have demonstrated different ways of filter the normal and abnormal data. The implementation of this method is very simple. Firstly, outlier methods are usually return only top n outliers with two values. Among them, one is the outlier factor or score and another one is the ranking of the points according to the factors or scores. Therefore, many types of methods have various kinds of rankings for top outliers. Secondly, it has been observed that k-nearest neighbours based outlier detection methods are sensitive to the parameter k and a small change in k can lead to changes in the scores and, correspondingly, the ranking. As a result, except for very strong outliers where the scores are distinct, the ranking is sensitive to k as well [S. Ganapathy *et al*., 2011].

In this paper, a new weighted outlier detection algorithm has been developed and implemented by using Minimum Spanning Tree (MST), K-Nearest Neighbour (kNN) and Intelligent Agents. Here, we have calculated three values as scores such as global and local. The various experiments have been conducted with standard datasets. The rest of the paper is organized as follows. Section 2 gives the brief literature survey about the outlier detection related intrusion detection systems which were developed in the past by various researchers. Section 3 explains the overall system architecture. Section 4 describes the proposed work. Section 5 discusses the experimental results and reason for the achievements. Section 6 gives the conclusion and future enhancements.

LITERATURE SURVEY

There are many works in the literature that discuss about outlier detection related intrusion detections. Among them, [Fabrizio Angiulli *et al*., 2006] proposed a distance based outlier detection method, which is to find the top outliers in an unlabelled data set and to provide a subset of it, called the outlier detection by solving agent. This solving agent investigates the accuracy effectively based on various outliers. [Xiao Chun Wang *et al*., 2015] proposed a new global outlier factor and a new local outlier factor and an efficient outlier detection algorithm developed upon them that is easy to implement and can provide competing performances with existing solutions. [T. Luo *et al*., 2010] proposed a Multi-prototype clustering algorithm based on minimum spanning tree for improving the classification accuracy. They achieved better detection accuracy with the help of this clustering process. An intelligent Intrusion Detection System has been proposed and implemented by Ganapathy *et al* [S. Ganapathy *et al.*, 2011] by using the proposed Weighted Distance Based Outlier Detection algorithm for effective intrusion detection. They achieved better detection accuracy for DoS, Probe and other attacks are 99.62%, 99.42% and 99.52%. They improved the detection accuracy than existing works and also reduced the false alarm rates. Huang *et al* [H. Huang *et al.*, 2013] proposed a rank based outlier detection for effective data classification. They introduced a new rank assignment for nodes or records of the standard data set.

A novel intrusion detection method by combining two anomaly methods namely conformal predictor K-Nearest Neighbour (KNN) and distance based outlier detection (CPDOD) algorithm is used to detect anomalies with low false alarm rate and high detection rate [Farhan Abdel-Fattah *et al.*, 2012]. [Parneeta Dhaliwal *et al*., 2010] introduced a new clustering approach, which divides the stream in chunks and clusters each chunk using k-median into variable number of clusters. The weighted medians found in each phases of their work were tested for outlierness and assign a number to all phases, it is either declared as a real outlier or an inlier. Their proposed method is theoretically proved that is better than the k-means as it does not fix the number of clusters to k. It provides a more stable and better solution which runs in poly-logarithmic space.[Jo-Anne Ting *et al*., 2013] introduced a Bayesian weighted regression algorithm that is able to detect and eliminate outliers automatically in real-time without help of any users interference, parameter tuning, sampling or model assumptions about the underlying data structure. They compared their proposed algorithm to standard approaches for outlier detection, such as thresholding using Mahalanobis distance, mixture models, robust least squares with bi-square weights and an alternate variational Bayesian approach to robust regression.

[Jeen Shing Wang *et al*., 2008] proposed a cluster validity measure with outlier detection and cluster merging algorithms for the Support Vector Clustering (SVC) algorithm. This algorithm is capable of identifying the ideal cluster number with compact and smooth arbitrary shaped cluster contours increases robustness to outliers and noises. [Ganapathy S. *et al*., 2012] proposed a new intelligent agent-based intrusion detection model for mobile ad hoc networks using a combination of attribute selection, outlier detection, and enhanced multiclass SVM classification methods. In their model, an effective pre-processing technique is proposed that improves the detection accuracy and reduces the processing time. In addition, two new algorithms, namely, an Intelligent Agent Weighted Distance Outlier Detection algorithm and an Intelligent Agent-based Enhanced Multiclass Support Vector Machine algorithm are proposed for detecting the intruders in a distributed database environment that uses intelligent agents for trust management and coordination in transaction processing.

SYSTEM ARCHITECTURE

The proposed intrusion detection system consists of two major modules namely, user interface module and weighted outlier detection module. The architecture of the system is shown in Figure-1.

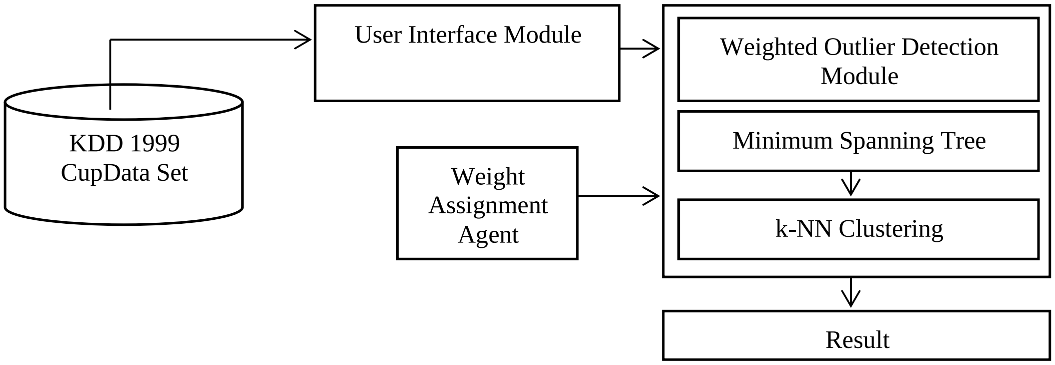

Figure-1. System architecture.

User interface module collects the necessary networks data from KDD'99 Cup dataset and send to the weighted outlier detection module for further processing. The weighted outlier detection module consists of two sub modules namely Minimum Spanning Tree and k-NN clustering. These sub module helps to do the pre-processing pre-process and classify the input data. In addition, a weight assignment also helps to assign the necessary weightage for all the input training data. This weight assignment agent plays a major role in this system to detect the intruders. This resulted data can be stored in result Table.

PROPOSED WORK

In this paper, a new Intelligent Weighted Outlier Detection Model is proposed by using Minimum Spanning Tree (MST), K-Nearest Neighbour (k-NN) and Intelligent Agents according to [Xiao Chun Wang *et al*., 2015] [S. Ganapathy *et al*., 2011]. The proposed method is the combination of Weighted Distance Based Outlier Detection (WDBOD) [S. Ganapathy *et al*., 2011] and Fast MST and k-NN based Outlier Detection [Xiao Chun Wang *et al*., 2015].

Outlier detectors

This proposed method is used three outlier detectors namely mini MST-based global outlier indicator, MST-inspired k-NN-based global outlier factor and MST-inspired k-NN-based local outlier factor. In addition to that intelligent agents are used for assigning weights to the necessary features of the dataset and for making effective decision in the intrusion detection process of the proposed model. First, mini MST-based global outlier indicator is to be constructed based on a data point and its k-nearest neighbors, dist [i] denotes the ith edge weight of such miniMST starting at the point ,miniMST-based global outlier indicator. Second, MST-inspired k-NN based global outlier factor is to be constructed upon a data point and its k-nearest neighbours, dist[i] denote the ith edge weight of such miniMST and cut-thred be a user provided threshold for SOM$_{MST}$ which measures the possibility of outlier existence. Third, MST-inspired kNN-based local outlier factor is to be constructed upon a data point and its k-nearest neighbour and dist[i] denote the ith edge weight of such miniMST. The local outlier detection can be assumed to be over when the iodrops below a threshold (say3). Finally, the outlier scores are used to assign the return edd at a point a degree of being outlying.

Nearest neighbour search structure

The data points are normal and the goal of finding top n outliers can be achieved by first quickly finding a good estimate of the outlying score for each data item and then focusing on top mZn ones. By removing all the inliers among them, the required top n outliers show up. To meet this goal, we are particularly interested in an approximate nearest neighbours search facility, called the divisive hierarchical clustering algorithm (DHCA) [X. Wang *et al*., 2009]. Essentially, to start the DHCA, k centres at the top level are randomly selected from the whole dataset. Next, each data point is assigned to its closest centre, creating K partitions. At each successive level in the iteration, for each of these K partitions, K random centres are recursively selected within each partition and the clustering process continues to format most Kn partitions at the nth stage. The procedure continues until the number of elements in a partition is below K+2, at which time a nearest neighbour search among all the data items in that partition is conducted .Such a strategy ensures that points that are close to each other in space are likely to be collocated in the same partition, and multiple runs of DHCA greatly enhance such possibilities. A more detailed demonstration and proof of the effectiveness of DHCA on approximate nearest neighbours' search have been given in [X. Wang *et al*., 2009] and will not be repeated here. After several iterations, exact kNNs and the correspondingly scores are computed fort opoutliers. The number of top outliers is small, thus the computation time is fast. With these observations in mind, a simple outlier detection method is developed in the following.

Weighted distance based outlier detection algorithm

This paper used an existing algorithm called Weighted Distance based Outlier Detection algorithm [S. Ganapathy *et al*., 2011] for primary outlier detection on input dataset. The resulted dataset will be given as input data source to the WMKOD for further process.

Intelligent MST and k-NN based outlier detection algorithm

We combine the above three factors to create the proposed MST-inspired k-NN-based weighted outlier detection algorithm. Moreover, intelligent agents are used for assign suitable weights to the variables effectively and also used for making effective decision.

a) 1.Agent set k to bethelargestoutlyingclustersizeplus1;

b) 2. Agent reads the data sequentially and initializes k neighbours of each data item from its immediate predecessors or successors on the fly as per Minimum Spanning Tree process (Sequential Initialization (SI);

c) 3. Agent execute the DHCA process multiple times, and, for each iteration, calculate an one dimensional array of the average distance of each data item to its k-NN and then the array mean; stop this step when the percentage differential of the mean between two consecutively iterations is below 10^{-6};

d) 4. Agent constructs the mini MST over each data item and its k nearest neighbours;

e) 5. Computesthree1-dimensionalarraysoftheestimated outlying scores for iNN (where i=2:k), and sort them in a non-increasing order by the decision agent.

f) 6.For data items with top global outlying scores, find their true iNN and calculate the irtrue outlying scores, check its SOM_{MST}, return it if the SOM_{MST} is larger than a threshold(say1.5),then if i=k, the global outlier detection is completed, otherwise, i=i+1 and goto5;

g) 7. For data items with top local outlying scores, find their true iNN and calculate their true outlying scores, return it if the score is larger than a threshold(say3),then if i=k, the local outlier detection is completed, otherwise, i=i+1 andgoto5;

h) 8. Repeat steps 5, 6 and7 until all the outliers whose score is above the threshold of global or local are mined.

The number of outliers is expected to be relatively small; the number of distance computations consumed is expected to be relatively small as well. The proposed algorithm consumed the physical resource and it includes the space to hold the full data set in memory, space to store their k neighbours, and some temporary space for miniMSTs. The numerical parameters of the proposed algorithm and it needs from the user include the number (k) of nearest neighbors, the global and local outlier thresholds, while the outputs include a set of ranked outliers from the dataset. The proposed intelligent outlier algorithm improves the readability.

Fast MST-inspired outlier detection algorithm

In this paper, we have used the existing algorithm developed by Wang *et al* [Xiao Chun Wang *et al*., 2015] which is the combining the Minimum Spanning Tree (MST) [Xiao Chun Wang *et al.*, 2015] and DHCA [Xiao Chun Wang *et al*., 2015] for effective outlier detection and removing the inliers.

Step 1: Initialize a set of nodes N, number of New Nodes (NN), number of clusters in each step, threshold for local and global outlier detection.

Step 2: Set cluster maximum size

Step 3: Perform a sequential initialization (SI)

Step 4: Call the procedure DHCA [Xiao Chun Wang *et al*., 2015] repeatedly until the percentage difference between two consecutively update.

Step 5: Find the miniMST for each data point

Step 6: For each cluster

Compute three one dimensional arrays of the estimated outlying scores for NN namely MST_MAX, MST_MAX_MIN and SOM and sort the first two in non-increasing orders are remembered in MST-MAX-INDEX and MST-MAX-MIN-INDEX

Step 7: Find true iNN for MST-MAX-INDEX[j]

Step 8: Re compute the global score and SOM

Step 9: if SOM [MST-MAX-INDEX[j]] > SOM-TH)
    GO_T = [GO_T MST-MAX-INDEX[j]];
  End

Step 10: GO=[GO GO_T]

Step 11: LO_T=[]

Step 12: for j = 1 to N
    Find iNN for MST-MAX-MIN-INDEX[j]
    Recomputed the local score
    If (MST-MAX-MIN[j] > MAX-MIN-TH)
     LO_T = [LO_T MST-MAX-MIN-INDEX[j]]
    End

Step 13: LO=[LO LO_T]

Step 14: End

The combination of WDBOD, Fast MST and k-NN based Outlier Detection and intelligent agent is called the proposed outlier detection model. The proposed model is like a hybrid approach for effective intrusion detection by using the removal of outliers. First, remove the outlier (intruder) based on the weighted distance. The resulted data set is send to the second phase for further process. In second phase, detect the outliers (intruders) effectively by using local and global scores which are calculated based on the Minimum Spanning Tree (MST) concept. Here, the threshold is set for making the decision for the various sized clusters.

RESULT AND DISCUSSIONS

Training and test data

The KDD Cup 1999 dataset is used for carried out experiments in this work. It was taken from the International Knowledge Discovery and Data Mining Tools Competition. All the records of this dataset have 41 attributes. These attributes are the combination of both continuous-type and discrete type variables with statistical distributions varying drastically from each other.

Experimental results

The proposed weighted MST and k-NN based outlier detection algorithm has achieved the highest detection accuracy for known attacks such as DoS, Probe and R2L. Similarly, it is also provides better detection accuracy on unknown attacks which are comes under the types of DoS and Others. Moreover, the disadvantage of this method is that it did not perform so well for detecting R2L attacks. This is the trade-off for the high detection rate of other attack types. Furthermore the less number of instances for R2L connections in the training and testing data makes the detection rate of R2L negligible compared to other attacks. Table-1 shows the performance analysis of WMKOD in comparison with various classification algorithms such as ID3, C4.5, SVM, EMSVM and WDBOD. From the Table, we observed that WMKOD algorithm is better than the previous classification algorithms in classifying DoS, Probe and other attacks.

Table-1. Performance analysis of WDBOD.

Algorithm	**Accuracy** (%)		
	DoS	**Probe**	**Others**
ID3	95.67	95.45	95.62
C 4.5	96.32	96.06	96.21
SVM	97.52	97.34	97.42
EMSVM	99.12	99.00	99.19
WDBOD	99.52	99.12	99.41
WMKOD	99.71	99.53	99.59

Table-2 shows the performance analysis of the proposed WMKOD algorithm. We generated 100 attackers in the system and captured 98 intruders by using the WMKOD algorithm in the first experiment with a false positive of 2 only. Next, we generated 200 attacks in which we captured 196 intruders, having a false positive of 9 only. Similar to this way, we captured 294, 392 and 490intruders out of 300, 400 and 500generated intruders by using this proposed algorithm having false positives of 11, 13 and 15 respectively.

Table-2. The performances of classification algorithms.

No. of intrusions generated	Number of intruders captured	True Negative	False positive
100	98	2	2
200	196	7	9
300	294	9	11
400	392	11	13
500	490	13	15

Table-1 shows the comparative analysis of three kinds attack types such as DoS, Probe and others when they are detected using ID3, C4.5, MSVM, EMSVM and WMKOD. It shows that this proposed weighted outlier detection algorithm is very efficient in detecting intrusions with an extremely low false negative rate compared with other algorithms namely ID3, C4.5, SVM, MSVM and EMSVM.

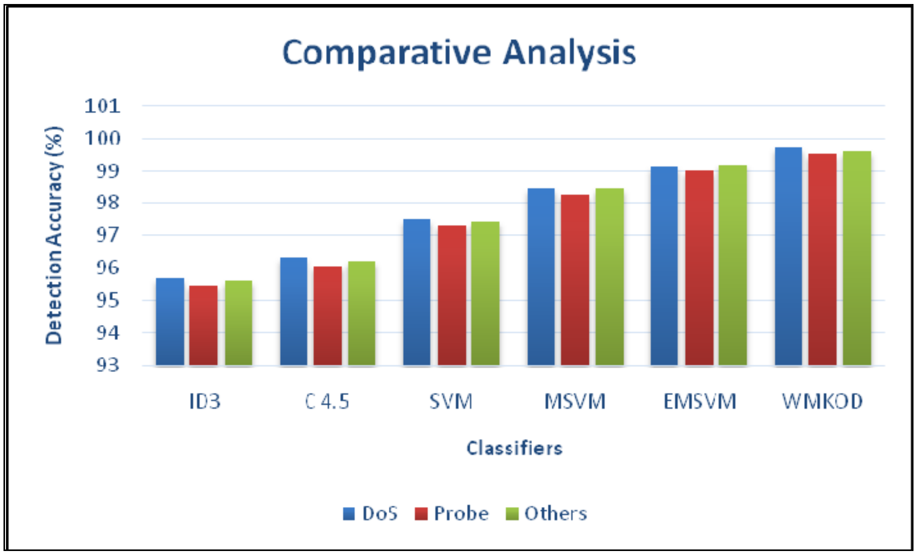

Figure-2. The results comparison of MSVM, EMSVM and WDBOD.

Figure-2 shows the performance of WMKOD in comparison with the WDBOD algorithm. From this graph, it can be observed that the false alarm rate decreases with respect to increase in number of packets when WMKOD is used. On the other hand, the existing WDBOD provides only a marginal decrease in false alarm rate with respect to the increase in number of packets sent.

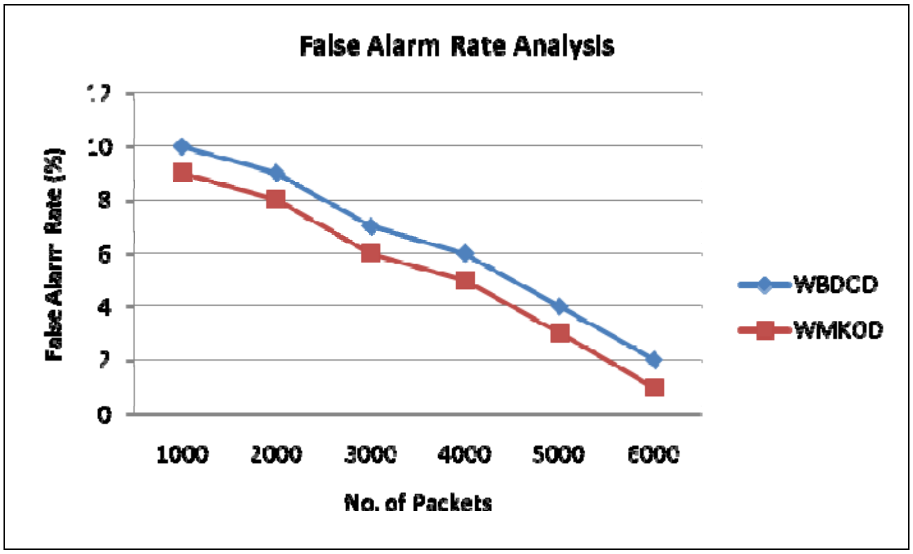

Figure-3. Performance of reducing false alarm rate.

The proposed method provides better detection accuracy than all other classifiers due to the uses of minimum Spanning Tree (MST), K-Nearest Neighbour Clustering (k-NN) and Weight assignment agent. The weight assignment agent assigns suitable weight for all the training records which are present in the training dataset effectively depends on the attacks types.

CONCLUSION AND FUTURE WORKS

A new outlier detection algorithm called Weighted MST and k-NN Based Outlier Detection (WMKOD) algorithm is proposed and implemented in this paper to detect the intruders in all kinds of network environment. This proposed algorithm considered the weights for each training data to apply the concept of Minimum Spanning Tree (MST) and the K-nearest neighbour (k-NN) clustering method. The proposed system performs well on the detection of attackers in the networks effectively. The experimental results show that the proposed algorithm improves the detection accuracy and reduces the false alarm rate. Future works could be done in this direction is to the introduction of effective clustering methods.

REFERENCES

[1] Xiao Chun Wang, Xia Li Wang, Yong Qiang Ma, D. Mitchell Wilkes. 2015. A fast MST-inspired kNN-based outlier detection method. Information Systems. 48: 89-112.

[2] S. Ganapathy, N. Jaisankar, P. Yogesh, A. Kannan. 2011. An Intelligent System for Intrusion Detection Using Outlier Detection. IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011. pp. 119-123.

[3] Parneeta Dhaliwal, MPS Bhatia and Priti Bansal. 2010. A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median Outlier Miner). Journal of Computing. 2(2): 74-80.

[4] Jo-Anne Ting, Aaron D'Souza, Stefan Schaal. 2013. Automatic Outlier Detection: A Bayesian Approach", H. Huang, K. Mehrotra, C.K. Mohan, "Rank-based outlier detection. J. Stat. Comput. Simul. 83(3): 1-14.

[5] T. Luo, C. Zhong, H. Li, X. Sun. 2010. A Multi-prototype clustering algorithm based on minimum spanning tree. Proceedings of Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). pp. 1602-1607.

[6] Ganapathy S, Yogesh P and Kannan Arputharaj. 2012. Intelligent Agent Based Intrusion Detection System Using Enhanced Multiclass SVM. Computational Intelligence and Neuroscience. 1-10.

[7] Farhan Abdel-Fattah, Zulkhairi Md. Dahalin and Shaidah Jusoh. 2010. Dynamic Intrusion Detection Method for Mobile Ad-hoc Networks Using CPDOD Algorithm. IJCA Special Issues on Mobile Ad-hoc Networks MANETs. 12(5): 22-29.

[8] Fabrizio Angiulli, Stefano Basta and Clara Pizzuti. 2006. Distance based Detection and prediction of Outliers. IEEE Transactions on Knowledge and Data Engineering. 18(2).

[9] S. Ganapathy, K. Kulothungan, S. Muthuraj Kumar, M. Vijaya Lakshmi, P. Yogesh, A. Kannan. 2013. Intelligent Feature Selection and Classification Techniques for Intrusion Detection in Networks: A Survey. EURASIP Journal of Wireless Communications and Networking. 271: 1-16.

[10] Jeen Shing Wang and Jen-Chieh Chiang. 2008. A Cluster Validity Measure with Outlier Detection for Support Vector Machine. IEEE Transactions on Systems, Man, and Cybernatics-Part B Cybernatics. 38(1): 78-89.

[11] S. Ganapathy, N. Jaisankar, P. Yogesh, A.Kannan. 2011. An Intelligent Intrusion Detection System Using Outlier Detection and Multiclass SVM. International Journal of Recent Trends in Engineering and Technology. 05(01): 166-169.

[12] X. Wang, X. L. Wang, D. M. Wilkes. 2009. A divide-and-conquer approach for minimum spanning tree-based clustering, IEEETKDE. 21(7): 945-958.

[13] T. Luo, C. Zhong, H. Li, X. Sun. 2010. A multi-prototype clustering algorithm based on minimum spanning tree. In: Proceedings of 7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD2010). pp. 1602-1607.

[14] S. Ganapathy, P. Yogesh, A. Kannan. 2011. An Intelligent Intrusion Detection System for Mobile Ad hoc Networks Using Classification Techniques. Communications in Computers and Information Systems, Springer. Vol. No. 148, pp. 117-122.