doi: 10.14355/ijcsa.2014.0304.01

Efficient Path Selection and Data Transmission Using Queue in Open Shortest Path First

J. Kavitha[1], Dr. S. Palaniammal[2]

[1]Department of Mathematics, Chettinad College of Engineering and Technology, Puliyur, Karur, Tamilnadu, India.

[2]Professor and Head, Department of Science and Humanities, Sri Krishna College of Technology, Kovaipudur, Coimbatore-42, Taminadu, India.

kavitha.paveesh@gmail.com

Abstract

In network, information apportioning occurs between one communication points to another using TCP protocol. The collected data are shared among many users via Single Router or Access Point (AP). During routing, traffic is occurred in the network and that is known as Random Early Detection (RED). In order to overcome this, an alternative path must be preferred by using Alternative Path Selection method which utilizes the Queue Process (FIFO). In case the queue is full, it chooses an alternative queue by using channel sensing mode of path selection. Here, we use OSPF (Open Shortest Path First) which selects the shortest path in short period of time and also it delivers the data to the particular destination. However, delay is reduced in OSPF. Hence, efficiency and overall network lifetime is increased. Finally, the network performance is evaluated.

Keywords

TCP; RED; FIFO; Channel Sensing; OSPF

Introduction

Generally, we employ queuing system in our day to day commercial lives which are provided by the commercial organizations like Checkout counters, Super markets, Banking sectors and Fast food restaurants etc. The queue process is also used in networking that has a queue buffer at router which collects data packets from the source and transmits to the specific destination. In routing, congestion may be encounter in outgoing bandwidth and the packets may drop or packet delay will be increased during congestion because of the overflow in queue. Therefore, the congestion reduces the overall performance of the whole network by reducing the throughput. In order to reduce the congestion, Transmission Control Protocol (TCP) is required which precludes the network from crashing. Moreover, the congestion control plays a crucial role in TCP which adjust the data rate in order to avoid congestion. However, the regular TCP cannot fully operate the limited resources in whole network due its unawareness and it can be separate the packet loss from random/congestion loss.

The packet loss is avoided by channel sensing that utilizes the standard link state protocol of Open Shortest Path First (OSPF) which senses the shortest paths in the network for packet routing. It has Link-State Advertisements (LSAs) to indicate the directly connected paths/links. This protocol also maintains the neighbor relationship with their adjacent Access Point/Router. When there is a change like path addition/disconnection is immediately updated by the OSPF. Once in every 30 minutes the LSAs have to be refreshed. The parameters which are used in this paper are briefly explained below.

Queue

FIG. 1 describes the Queuing process which has become a part of our daily lives and Queuing theory is used by the Queue which constructs a simple model that affords in mathematical analysis. It produces enough details about the process that speculate the doings of the real systems. The Queuing theory is splitted into two types namely, highly abstract and highly practical. Basically it is purely a mathematical approach and that are utilized in the waiting line analysis. It employs various models to constitute various queuing systems. However, the formula for every queue model must explain the performance of related queue under various conditions and these queuing models are the most powerful tools to provide the instructions to handle the system in an efficient way.

The queuing theory is known as Random System Theory which has the solutions for statistical interference and problem of behavior and optimization in queuing system. Fair Queuing is a famous scheduling prototype in both wired and wireless networks and most of the wireless networks ad hoc have the drawbacks of incomplete scheduling information, location dependent channel contention and spatial channel reuse. So, this algorithm cannot enforce directly on wireless ad hoc networks. Thus, the packet scheduling cannot follow the local decisions of the sender which controvert to that in the wired or wireless networks.

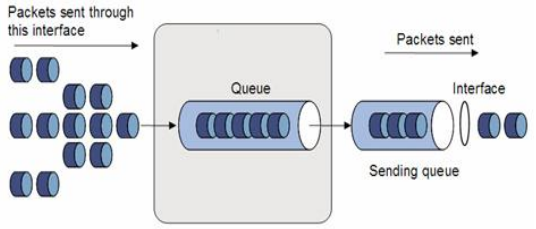

FIG. 1 QUEUE PROCESS IN PACKET SENDING

TCP

In FIG. 2, Transmission Control Protocol (TCP) is a reliable transport protocol which provides authentic data delivery services and end-to-end congestion control. The objective of the congestion control is immediately to transmit packets from source to destination without inducing congestion among the intermediate Access Points/routers. TCP uses the congestion window to prescribe the unacknowledged packets that endure between source and destination. In order to accomplish the reliable services, TCP discover the packet loss and retransmit such packets.

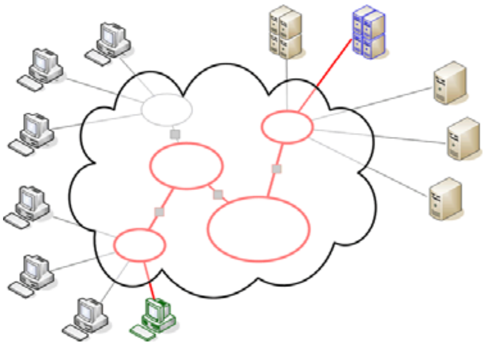

FIG. 2 TCP OPERATION AT ROUTER

Retransmission timer is used to retransmit the dropped packets and it starts during the onset of each packet which wills timeout if the acknowledgement is not obtained before the expiration of time. Then, the TCP retransmits these packets and activate the congestion control mechanism. TCP cannot use in broadcast and multicast networks and the regular TCP cannot fully control the limited resources in an efficient manner because of its unawareness.

OSPF

FIG. 3 describes Open Shortest Path First (OSPF) is the standard Link-State routing protocol that is used to affirm more number of networks in an efficient manner. Here, OSPF is used to sense the shortest path in the network. These are designed to support several areas and OSPF builds relationship with adjacent Access Points in same area and it has Link State Advertisements (LSAs) which are used to mention the position directly connected links/paths. Additionally, they are used to indicate the path connection or disconnection in the network. OSPF employs Dijkstra Shortest Path First algorithm to find out shortest paths in the network.

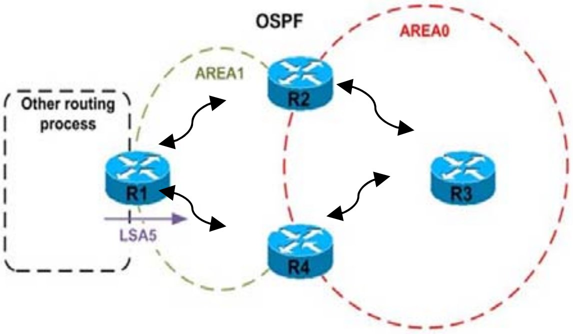

FIG. 3 OSPF ROUTING

Related Work

Active Queue Management (ACM) is a mechanism that is used to discover the congestion within the network. It is necessary to advocate the activity of AQM in Access Points/routers as a measure in order to maintain and improve the performance of Wide Area Network (WAN). These AQM algorithms are run on the AP/router and it continuously monitors the average size of the queue to observe the nascent congestions. In case the average sizes of the queue exceed a particular threshold value which are still less than the total size of the queue, the AQM algorithm deduce congestion in the paths/links and it advises the end systems by dropping some packets that are getting in the router.

It is essential to manage the router queue by setting each queue length as maximum and it accepts the packets till the maximum length is achieved. When the queue is full, it refuses or drops the entering packets till the queue length is reduced. But, in order to defeat the trouble in congestion the AQM algorithms can only adjust the packet dropping rates and these algorithms do not have the property of adaptability. In order to satisfy the router link capacity and various traffic loads, the AQM algorithms have to be prepared/set a group of parameters. It is a challenging task for bandwidth efficiency and video coding techniques to overcome the congestions which improves the performance.

Lock-out and Full Queue are some of the drawbacks in Drop tail (FIFO) method used by the AQM. Sometimes the drop tail grants a single connection and it rejects other connections. This situation is called Lock-out. When the queue is full, next arrival the packets may drop/refuse and it is necessary to concentrate on the steady state queue size. In order to reduce such drawbacks, Alternate Path Selection using OSPF algorithm is proposed.

Data Flow Diagram

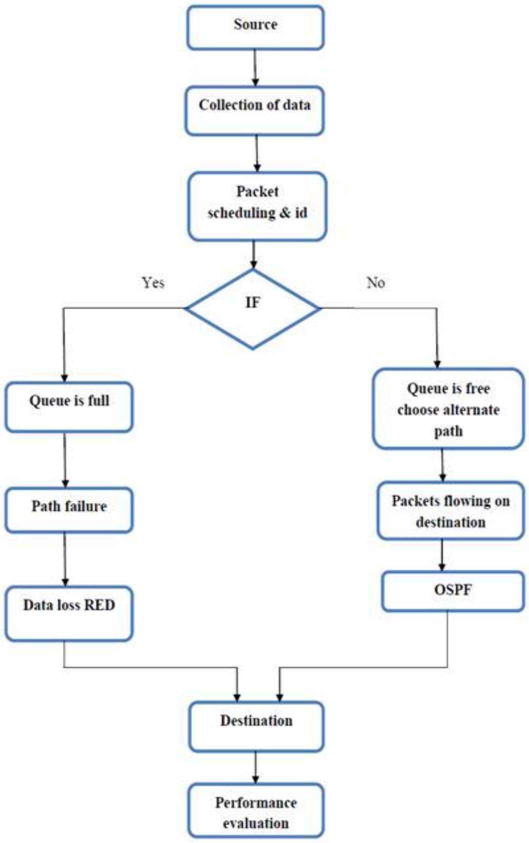

FIG. 4 DATA FLOW DIAGRAM

Algorithm for Alternate Path Selection using Ospf

1. Initialize The S_{ix}
2. Initialize The Routing Process R_i
3. Path Selection Process P_{sel} Using Queue Process
4. Q_{ix} (Fifo)
    a. If The Path Is Free, Packet Transmission Was Start.
    b. If The Path Is Full, Packet Transmission Was Drop.
5. Begin,
6. $Q_{ix<-->}\ P_{sel}$
7. (P_{sel} =0; P_{sel} <15; P_{sel} ++)
8. If
9. P_{sel} = Full
10. $P_{sch<-}$ Red
11. Else
12. $P_{sel<-}$ Null
13. $P_{id<-}$ Transmitting
14. Ospf (Open Shortest Path First)
15. O_{si}
16. Ospf (Open Shortest Path First)
17. $D_{ix<-}$ Ospf
18. End

Description

In FIG. 4, initially, set of nodes are deployed in the network in which one node acts as a source and any one of its nodes will act as a router. Data packets transmit through several nodes and various paths and if a path is failure, it chooses the alternate path at particular time interval. In order to choose the paths, channel sensing using Open Shortest Path First (OSPF) is used by the router at receiver side and a queue buffer is used at the router that adopt the FIFO process in order that to transmits the packets to correct destination. When the queue is full, the next arrival packets are refused or sometimes packets may be dropped until the queue becomes free or the length of the queue is reduced. If any one of the path is short and also efficient, then it gives the first preference to that path. When the packets are collected from that path, it chooses the next shortest efficient path. This process is repeated until the entire transmission is finished. Compare to Active Queue Management efficient transmission is obtained. At last, the overall network throughput is increased.

Experimental Results and Calculations

In FIG. 5, we have to send 1000 packets over 5 paths using queue process. In each path some amount of packets are dropped due to channel noise and in first path 40 packets are dropped likewise different amount of packets are dropped in remaining path. Efficiency of each path is calculated by the way we can calculate the overall network efficiency.

P_{total} = 1000 Pckts; P_{th} = 5 path;

$$P_s = \frac{P_{total}}{P_{th}} = 200 \text{ Pckts}$$

Where, P_{total} = Total number of packets

P_{th} = Total number of paths

P_s = Number of packets transmitted through single path

Ist Path Iteration:

P_{l1} = 40 Pckts

$$\text{Efficiency } (\eta_1) = \frac{P_{l1}}{P_s} \times 100$$

$$= \frac{40}{200} \times 100$$

$$= 20\%.$$

Where, P_{l1} = Number of dropped packets in path 1.

η_1 = Efficiency of 1st path

IInd Path Iteration:

$$\text{Efficiency } (\eta_2) = \frac{P_{l2}}{P_s} \times 100$$

$$\eta_2 = 18.5\%.$$

IIIrd Path Iteration:

$$\text{Efficiency } (\eta_3) = \frac{P_{l3}}{P_s} \times 100$$

$$\eta_3 = 17.5\%.$$

IVth Path Iteration:

$$\text{Efficiency } (\eta_4) = \frac{P_{l4}}{P_s} \times 100$$

$$\eta_4 = 16\%.$$

V th Path Iteration:

$$\text{Efficiency } (\eta_5) = \frac{P_{l5}}{P_s} \times 100$$

$$\eta_5 = 15\%.$$

Overall efficiency:

$$\text{Efficiency } (\eta_i) = \frac{\sum_{i=1}^{5} \eta_i}{5}$$

$$\eta_i = 17.4\%$$

Path Iterations	Path 1	Path 2	Path 3	Path 4	Path 5
Efficiency (%)	20	18.5	17.5	16	15

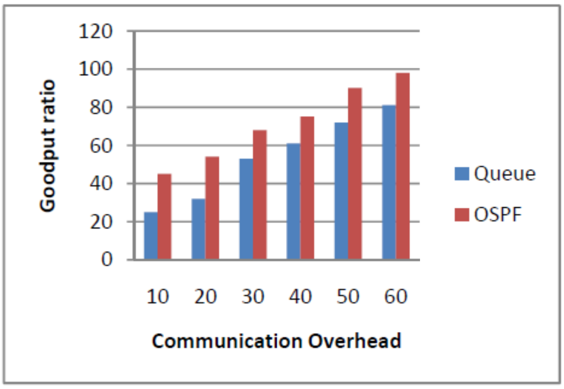

FIG. 5 COMPARE THE PERFORMANCE BETWEEN GOOD PUT RATIO AND COMMUNICATION OVERHEAD

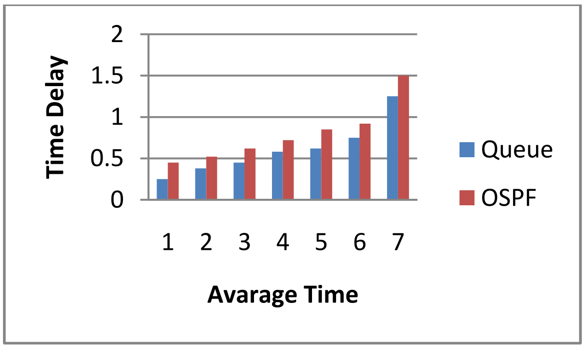

FIG. 6 COMPARE THE PERFORMANCE BETWEEN TIME DELAY AND AVERAGE TIME

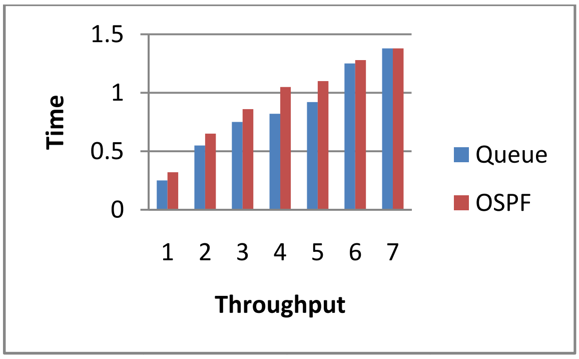

FIG. 7 COMPARE THE PERFORMANCE BETWEEN TIME AND THROUGHPUT

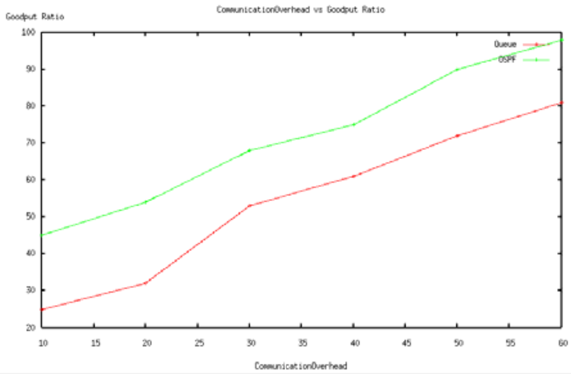

FIG. 8 COMPARE THE PERFORMANCE BETWEEN GOOD PUT RATIO AND COMMUNICATION OVERHEAD USING QUEUE PROCESS NETWORK MODEL

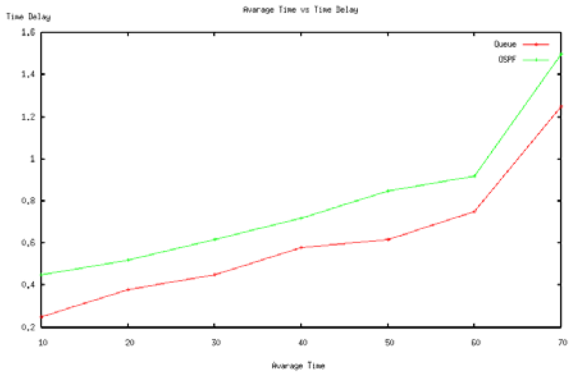

FIG. 9 COMPARE THE PERFORMANCE BETWEEN TIME DELAY AND AVERAGE TIME USING OSPF PROCESS IN QUEUE MANAGEMENT

Conclusion

In previous method of data sharing Active Queue Management (ACM) is used which increases the packet loss and packet drop. To reduce the above drawbacks, Alternative Path Selection method that has OSPF (Open Shortest Path First) is proposed. In OSPF, an alternative shortest path is detected in short duration using channel sensing. When compared to Active Queue Management process, packet efficiency and throughput are also increased. Hence, the overall network performance is evaluated.

REFERENCES

A. Shaikh, M. Goyal, A. Greenberg, R. Rajan, and K. Ramakrishnan. An OSPF Topology Server: Design and Evaluation. IEEE Journal on Selected Areas in Communications (J-SAC), 20(4), May 2002.

C. Chen, H. Wang, XinWang, M. Li, A.O. Lim, "A Novel Receiver-aided Scheme for Improving TCP Performance in Multihop Wireless Networks", International Conference on Communications and Mobile Computing, pp. 272 – 277, 2009.

D. Goldsman, A simulation course for high school students, in Proceedings of the 2007 Winter Simulation Conference, S.Henderson, B. Biller. M. Hsieh, J. Shortle, J. Tew, and R. Barton, eds., J.W. Marriott Hotel, Washington, D.C. 9-12 December, 2007, pp. 2353-2356.

D. Murray, T. Koziniec, M. Dixon, "D-Proxy - Reliability in Wireless Networks", 16th Asia Pacific Conference on Communications, pp. 129 – 134, 2010.

E. W. Dijkstra. A Note on Two Problems in Connexion with Graphs.NumerischeMathematik 1, pages 269–271, 1959.

HaiyunLuo, Jerry Cheng and Songwu Lu: "Self-coordinating Localized Fair Queuing in Wireless Ad Hoc Networks"2004 IEEE.

J. T. Moy. OSPF Version 2. RFC2328, April 1998. [4] R. Rastogi, Y. Breitbart, M. Garofalakis, and A. Kumar.Optimal Conguration of OSPF Aggregates.In Proc. IEEE INFOCOM, June 2002.

J. T. Moy. OSPF: Anatomy of an Internet Routing Protocol. Addison-Wesley, January 1998.

Janos Sztrik: "Queuing theory and its applications a personal view" 2010 International conference on applied informatics Vol.1

L. Devroye, Non-Uniform Random Variate Generation, Springer-Verlag, New York, 1986.

Lachlan L.H. Andrew, Stephen V. Hanly and Rami G. Mukhtar: "Active Queue Management for Fair resource Allocation in Wireless Networks"2008 IEEE

N. Itoh , M. Yamamoto , "Proxy-based TCP with Adaptive Rate Control and Intentional Flow Control in Ad Hoc Networks", IEEE Global Telecommunications Conference, 2008.

N.SaravanaSelvam and Dr.Radhakrishnan: "Processor Based Active Queue Management for providing QoS in Multimedia Application"2010 International Journal of computer science and information Security.

Peter Marbach and Yiyi Lu: "Active Queue Management and Scheduling for Wireless Networks:The Single-Cell Case"2006 Informatics sciences and systems conference at Princeton, NJ.

Qiuyan Xia, Xing Jin and Mounir Hamdi: "Dual Queue Management for Improving TCP Performance in Multi-rate Infrastructure WLANs"2008 IEEE.

S. Prasanthi, S. Chung, "An Efficient Algorithm for the Performance of TCP over Multi-hop Wireless Mesh Networks", Seventh International Conference on Information Technology, pp. 816 – 821, 2010.

S. Prasanthi, S. Chung, C. Ahn, "An Enhanced TCP Mechanism for Detecting and Differentiating the Loss of Retransmissions over Wireless Networks", International Conference on Advanced Information Networking and Applications, pp. 54 – 61, 2011.

Snehal Shegaonkar and Seema Wasnik: "Traffic Sensitive Active Queue Management for Improving QoS of the

System”2013 International Journal of Software and Web Sciences (IJSWS).

T. Shikama, "Mitigation of Bursty Packets by a TCP Proxy improving TCP Performance in a Wired and Wireless Network", IEEE Globecom Workshop on Complex Communication Networks, pp. 425 – 429, 2010.

T. Bhaskar Reddy and Ali Ahammed: “Performance Comparison of Active Queue Management Techniques”2008 journal of computer science.

Toshiba Sheikh, Sanjay Kumar singh, Anil kumarkashyap: “Application of queuing theory for the improvement of Bank service” 2008 International journal of advanced computational Engineering and networking.

V. Sinthu Janita Prakash, Dr. D. I George Amalarethinam, Dr. E. George Dharma Prakash Raj: “Extended Queue Management Congestion Control Algorithms for TCP Bulk Transfers in Wireless Environment” 2012 International Journal of Scientific & Engineering Research, Volume 3, Issue 5.

doi: 10.14355/ijcsa.2014.0304.02

Intrusion Detection System Using Feature Selection and Classification Technique

Senthilnayaki Balakrishnan*1, Venkatalakshmi K2, Kannan A3

Department of Information Technology, Department of Electronics and Communication Engineering, Department of Information Science and Technology, Anna University Villupuram, Anna University Tindivanam, Anna University Chennai, Tamilnadu, India

*1nayakiphd@gmail.com; 2venkata_krish@yahoo.com; 3kannan@annauniv.edu

Abstract

With the growth of Internet, there has been a tremendous increases in the number of attacks and therefore Intrusion Detection Systems (IDS's) has become a main stream of information security. The purpose of IDS is to help the computer systems to deal with attacks. This anomaly detection system creates a database of normal behaviour and deviations from the normal behaviour to trigger during the occurrence of intrusions. Based on the source of data, IDS is classified into Host based IDS and Network based IDS. In network based IDS, the individual packets flowing through the network are analyzed where as in host based IDS the activities on the single computer or host are analyzed. The feature selection used in IDS helps to reduce the classification time. In this paper, the IDS for detecting the attacks effectively has been proposed and implemented. For this purpose, a new feature selection algorithm called Optimal Feature Selection algorithm based on Information Gain Ratio has been proposed and implemented. This feature selection algorithm selects optimal number of features from KDD Cup dataset. In addition, two classification techniques namely Support Vector Machine and Rule Based Classification have been used for effective classification of the data set. This system is very efficient in detecting DoS attacks and effectively reduces the false alarm rate. The proposed feature selection and classification algorithms enhance the performance of the IDS in detecting the attacks.

Keywords

Intrusion Detection; Information Gain; Support Vector Machine; Feature Selection Technique and Classification

Introduction

Computers have been networked together with very large user source and so security has been a vital concern in many areas. With the rapid growth of internet communication and availability of tools to intrude the network, security for network has become indispensable. Current security policies do not sufficiently guard the data stored in the databases. Many other technologies like firewalls, encryption and authorization mechanisms can offer security, but they are still sensitive for attacks from hackers who takes advantage of the system flaws .To protect these systems from being attacked by intruders, a new Intrusion Detection System has been proposed and implemented in this project work, which combines a simple feature selection algorithm and SVM technique to detect attacks. Using KDD cup data set and Data Mining extract the hidden predictive information from large Databases. It is a powerful new technology with great potential that helps companies focus on the most important information in their data warehouses. Data mining can be applied to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Recently, internet has become a part of daily life. The current internets based on information processing systems are prone to different kind of threats which lead to various types of damages resulting in significant losses. Therefore, the importance of information security is evolving quickly. The most basic goal of network security is to develop defensive networking systems which are secure from unauthorized access, using disclosure, disruption, modification, or destruction. Moreover, network security minimizes the risks related to the main security goals like confidentiality, integrity and availability.

Related Work

In recent times, network security has been the subject of many research works with advent of internet. There are many works in the literature that discuss about Intrusion Detection System**.** IDSs are used to detect the attacks made by intruders. Sindhu et al proposed a genetic based feature selection algorithm for minimizing the computational complexity of the classifier. Lee et al proposed an adaptive data mining approach for intrusion detection in which association rules and frequent episodes derived from audit data are used as the basis for the feature selection process. Xiang and Lim proposed a misuse IDS using multiple-level hybrid classifier.

Moradi and Zulkernine presented an IDS that uses ANN for effective intrusion detection. One of the limitations of their approach is that it increases the training time. Sarasamma et al proposed a novel multilevel hierarchical Kohonen networks to detect intrusions in networks. In their work, they randomly selected data points from KDD Cup 99 to train and test the classifier. Jianping Li et al proposed a new method based on Continuous Random Function for selecting appropriate feature sets to perform network intrusion detection.

There are many classification algorithms based on SVM that are found in the literature for IDS. For example, an algorithm called Tree Structured Multiclass SVM had been proposed by Snehal A. Mulay et al for classifying data effectively. There are many works in the literature that discuss about pre-processing. Most of the real life problems definitely need an optimal and acceptable solution rather than calculating them precisely at the cost of degraded performance, time and space.

The feature selection search started with null set where features were added one by one or it was started with a full set of features where features were eliminated one by one. Li et al proposed a wrapper based feature selection algorithm in order to develop an IDS.

The feature selection algorithm proposed by Geetha Raman ideals with the statistical method for analyzing the voluminous KDD Cup dataset. There are many works in the literature that discuss about classification techniques and tools. Support Vector Machines (SVM) were the classifiers which were originally designed for binary classification.

Debar et al developed a Neural Network model for IDS. Du Hongle et al proposed an improved v-FSVM through introduction membership to each data point. Dewan Md. Farid proposed a new learning approach for network intrusion detection using naïve Bayesian classifier and ID3 algorithm is presented, which identifies effective attributes from the training dataset, calculates the conditional probabilities for the best attribute values, and then correctly classifies all the examples of training and testing dataset. The SVM-based intrusion detection system combines a hierarchical clustering algorithm, a simple feature selection procedure, and the SVM technique.

Data Preparation Subsystem

Data Collector

The Data collection agent collects the records from the KDD’99 cup data set. This data is sent to the data pre-processing module for pre-processing the data. The records collected from the KDD cup dataset may be a normal data or an attacked data**.**

Pre-processing Module

Pre-processing techniques are necessary for data reduction since it is quiet complex to process huge amount of network traffic data with all features to detect intruders in real time and to provide prevention methods.

Classification Subsystem

Rule Based Classifier

In this system, decisions on anomaly intrusion detection and prevention are improved by the application of rules fired using the rule system invoked by the intelligent agents. The main advantage of using rules with knowledge base is that it helps to perform effective decision making on intrusions.

Support Vector Machine

SVM is the learning machine that can perform binary classification and regression estimation tasks. They are becoming increasingly popular as a new paradigm of classification and learning because of two important factors. First, unlike the other classification techniques, SVM minimizes the expected error rather than minimizing the classification error. Second, SVM employs the duality theory of mathematical programming to get a dual problem that admits efficient computational methods.

Proposed Algorithm for Optimal Feature Selection

This algorithm has been developed by calculating Information Gain Ratio for attribute selection. In order to achieve this, the data set D is divided into n number of classes C_i. The attributes F_i having maximum number of non-zero values are chosen by the agent and the Information Gain Ratio (IGR) is computed using equations:

$$Info\,(D) = -\sum_{j=1}^{m} \left[\frac{freq(C_j,D)}{|D|}\right] log_2\left[\frac{freq(C_j,D)}{|D|}\right] \quad (1)$$

$$Info\ (F) = \sum_{i=1}^{n} \left[\frac{|F_i|}{|\ F\ |}\right] * info\ (F_i) \qquad (2)$$

$$IGR\ (Ai) = \left[\frac{Info(D)-Info(F)}{Info(D)+Info(F)}\right] * 100 \qquad (3)$$

The steps of the optimal feature selection algorithm are as follows.

Algorithm: Intelligent Agent based Attribute Selection Algorithm

Input: Set of 41 features from KDD'99 Cup data set

***Output*:** Reduced set of features R

Step 1: Select the attributes which have variation in their values.

Step 2: Calculate the Info (D) values for the selected attributes using the equation 1.

Step 3: Select the attributes which have maximum number of non-zero values.

Step 4: Calculate the Info(F) value for the attributes selected in step 3 using the equation 2.

Step 5: Calculate the IGR value using the equation 3.

Step 6: Depending on the IGR value, select the attributes.

The OFS algorithm has selected 10 important features for effectively detecting the attacks and to reduce the computation time.

The pseudo code for optimal feature selection is given below.

```
Input the data set
for each column in the data set
Select non-varying columns
end for
for each non-varying columns
calculate frequency of each value in the data set
calculate info(d)
end for
for each column with maximum no. of non-zero values
calculate frequency of each value
calculate info(f)
end for
for each column
calculate IGR value
end for
```

In this chapter, the algorithm to implement the OFS is presented. The next chapter discusses the implementation and analysis the performance results of the project.

Implementation

Optimal Feature Selection

The normal feature selection algorithms take large computation time for calculating IGR values. Hence in this work, a new feature selection algorithm called Optimal Feature Selection algorithm that reduces the time taken for computation is proposed and implemented. This algorithm calculates the Information Gain Ratio (IGR) value for the varying attributes in the data set. It performs column reduction based on the IGR value. OFS increases the accuracy in detection and reduces the false alarm rates.

The simulated attacks fall in one of the following four categories namely, Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L) and Probe attack.

TABLE 1 THE 41 FEATURES IN KDD'99 DATASET

S.NO	FEATURE NAME	S.NO	FEATURE NAME
1	Duration	22	Is_guest_login
2	Protocol type	23	Count
3	Service	24	Serror_rate
4	Src_byte	25	Rerror_rate
5	Dst_byte	26	Same_srv_rate
6	Flag	27	Diff_srv_rate
7	Land	28	Srv_count
8	Wrong_fragment	29	Srv_serror_rate
9	Urgent	30	Srv_rerror_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_count
14	Root_shell	35	Dst_host_diff_srv_count
15	Su_attempted	36	Dst_host_same_src_port_rate
16	Num_root	37	Dst_host_srv_diff_host_rate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_shells	40	Dst_host_rerror_rate
20	Num_outbound_cmds	41	Dst_host_srv_rerror_rate
21	Is_hot_login		

Calculation of info (D)

The information gain criterion is derived from information theory. The essential idea of information theory is that the information conveyed by a message depends on the probability and can be measured in bits as minus the logarithm of base 2 of that probability. Suppose we have a dataset D with q classes $C_1,\ldots C_n$. Suppose further that we have a possible test x with m outcomes that partitions D into m subsets $D_1,\ldots,D_m$. For a numeric attribute, m=2, since we only perform binary split. The probability that is selected one record from the set D of data records and announce that if belongs to some class C_j is given by ,

$$\sum_{j=1}^{m} \left[\frac{freq(C_j,D)}{|\,D\,|}\right] \qquad (4)$$

Where freq (Cj, D) represents the number of data records(points) of the class Cj in D, while |D| is the total number of data records in D. So the information that is convey is

$$-\,log_2\left[\frac{freq(C_j,D)}{|\,D\,|}\right] bits \qquad (5)$$

To find the expected information needed to identify the class of a data record in D before partitioning occurs, summation is performed over the classes in proportion to their frequencies in D, giving

$$Info\ (D) = -\sum_{j=1}^{m} \left[\frac{freq(C_j,D)}{|\,D\,|}\right] log_2\left[\frac{freq(C_j,D)}{|\,D\,|}\right] \qquad (1)$$

Now, suppose that the dataset D has been partitioned in accordance with the m outcomes of the test x. The expected amount of information needed to identify the class of a data record in D after the partitioning has occurred, can be found as the weighted sum over the subsets, as:

$$Info\ (F) = \sum_{i=1}^{n} \left[\frac{|F_i|}{|\,F\,|}\right] * info\ (F_i) \qquad (2)$$

where $|F_i|$ represents the number of data records in the subset D_i after the partitioning had occurred.

The information gained due to the partition is:

$$Gain(Ai) = Info(D) - \ Info(F) \qquad (6)$$

Clearly, it is necessary to maximize the gain. The gain criterion is to elect the test or cut the maximizes the gain to partition the current data.

$$IGR\ (Ai) = \left[\frac{Info(D)-Info(F)}{Info(D)+Info(F)}\right] * 100 \qquad (3)$$

Result

The performance and time analysis for the different types of attacks are analyzed and tabulated below. Table's shows the detection accuracy and computation time obtained using the features of the KDD'99 Cup data set by applying the feature selection techniques of existing and proposed work.

Rule Based Classification

The Rule based classification is the first step in the classification of different types of attacks. The performance analysis in terms of accuracy and time taken for classifying the attacks using Rule Based Classifier is tabulated in the TABLE 2. This table shows the detection accuracy and time taken for 5000 records.

TABLE 2 PERFORMANCE ANALYSIS FOR RULE BASED CLASSIFICATION

Attacks	Detection Accuracy (%)	Time in seconds
Dos	76.2	228
Probe	87.3	218
U2R	47.17	227
R2L	64.68	221

The classification time taken for rule based classification is pictorially represented in FIG. 1.

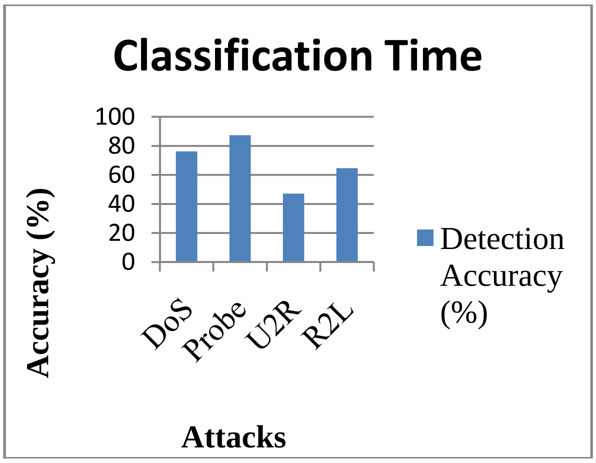

FIG. 1 CLASSIFICATION TIME ACCURACY OF RULE BASED CLASSIFICATION

The Classification accuracy for rule based classification is shown in the FIG. 2.

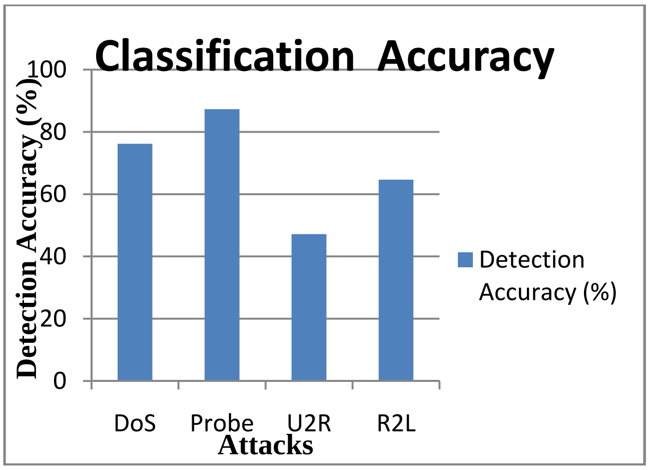

FIG. 2 CLASSIFICATION DETECTION ACCURACY OF RULE BASED CLASSIFICATION

Classification Using SVM

Further classification of the records obtained from Rule Based Classification is carried out using SVM. The comparison of detection accuracy for classification, done using all the 41 features of the KDD cup data set and the features selected using OFS are tabulated in the TABLE 3.

TABLE 3 CLASSIFICATION ACCURACY FOR SVM

Attacks	**No. of records**	**Detection Accuracy (%) for selected features**	**Detection Accuracy (%) for total features**
DoS	1581	99.11	99.11
Probe	1902	92.03	96.31
U2R	1745	91.51	96.15
R2L	1745	91.51	96.15

The computation time taken for classifying the 1581 records that are obtained from rule based classification as DoS attack is tabulated in TABLE 4.

TABLE 4 TIME ANALYSIS FOR DOS ATTACK IN SVM

Exp No.	Accuracy (%)	
	Selected features using OFS (10)	Total features (41)
1	2.22	2.31
2	2.14	2.26
3	2.03	2.23
4	2.01	2.18
5	2.0	2.17
Avg	2.08	2.23

It is observed that the time taken for classifying the DoS attack in SVM using features selected in OFS is less, as compared to that of the classification done using the 41 features of the KDD cup data set.. The computation time for classifying the DoS attacks in SVM is pictorially represented in the FIG. 3.

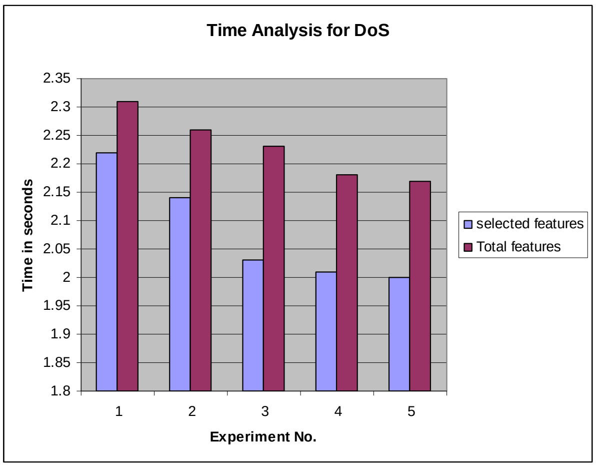

FIG. 3 COMPUTATION TIME FOR CLASSIFICATION OF DoS ATTACK

The computation time taken for classifying the records thatare obtained from rule based classification as probe attack is tabulated in TABLE 5.

TABLE 5 TIME ANALYSIS FOR PROBE ATTACK IN SVM

Exp No.	**Accuracy (%)**	
	Selected features using OFS (10)	**Total features (41)**
1	4.37	7.24
2	4.01	6.91
3	3.99	6.87
4	3.75	6.78
5	3.52	6.67
Avg	**3.93**	**6.89**

It is observed that the time taken for classifying the probe attack in SVM using the features selected in OFS is less, as compared to that of the classification done using the 41 features of the KDD cup data set**.**

The computation time taken for classifying the probe attack in WEKA is pictorially represented in FIG. 4.

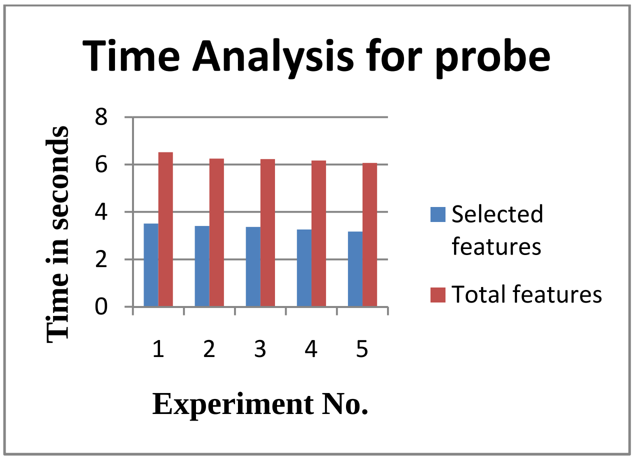

FIG. 4 COMPUTATION TIME FOR CLASSIFICATION OF PROBE ATTACK

The computation time taken for classifying the records that are obtained from rule based classification as U2Rattack is tabulated in TABLE 6.

TABLE 6 TIME ANALYSIS FOR U2R ATTACK IN SVM

Exp No.	Accuracy (%)	
	Selected features using OFS (10)	Total features (41)
1	3.51	6.52
2	3.41	6.25
3	3.37	6.23
4	3.26	6.17
5	3.17	6.07
Avg	3.34	6.24

It is observed that the time taken for classifying the U2R attack in WEKA using the features selected in OFS is less, as compared to that of the classification done using the 41 features of the KDD cup data set. The table 5.8 shows the time analysis for 1745 records.

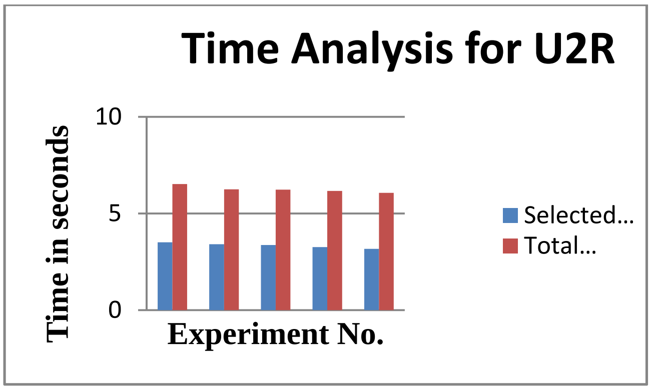

FIG. 5 COMPUTATION TIME FOR CLASSIFICATION OF U2R ATTACK

The computation time taken for classifying the U2R attack in WEKA is pictorially represented in FIG. 5.

The computation time taken for classifying the 1745 records that are obtained from rule based classification as R2Lattack is tabulated in TABLE 7.

TABLE 7 TIME ANALYSIS FOR R2L ATTACK IN SVM

Exp No.	Accuracy (%)	
	Selected features using OFS (10)	Total features (41)
1	3.56	6.91
2	3.48	6.48
3	3.46	6.42
4	3.37	6.23
5	3.07	5.97
Avg	**3.38**	**6.40**

It is observed that the time taken for classifying the R2L attack in WEKA, using the features selected in OFS is less, when compared to that of the classification, using the 41 features of the KDD cup data set.

The computation time taken for classifying the U2R attack in SVM is pictorially represented in FIG. 6.

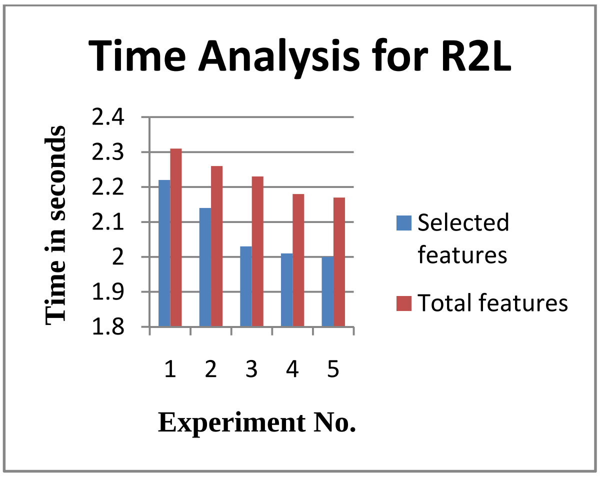

FIG. 6 COMPUTATION TIME FOR CLASSIFICATION OF R2L ATTACK

The computation time for SVM for different types of records is shown in the table8. The accuracy for detecting the records with selected features using OFS steadily increases, as the number of records increases and it becomes approximately equal to the accuracy of detecting the records with all features using SVM.

TABLE 8 ACCURACY ANALYSES FOR ATTACKS IN SVM

Exp No.	No. of records	DoS	Probe	U2R	R2L
1	5000	99.11	92.03	91.51	91.51
2	9000	99.15	94.17	93.83	93.83
3	13000	99.17	95.24	94.99	94.99
4	17000	99.20	95.77	95.57	95.57
5	21000	99.23	96.03	95.86	95.86
6	25000	99.25	96.16	96.00	96.00

In this chapter, the implementation and results of the proposed system are analyzed. The next chapter will provide a conclusion on this work and the future work to be done.

Conclusion

In this work, a new IDS has been proposed and implemented by combining an Optimal Feature Selection (OFS) algorithm and two classification techniques for securing the system. The computation time taken for detecting and classifying the records using all the forty one features of the KDD'99 cup data set is observed to be large. The proposed feature selection algorithm selects only the important features that help in reducing the time taken for detecting and classifying the records. Further the rule based classifier and SVM help achieve a greater accuracy. The main advantage of the proposed IDS is that it reduces the false positive rates and also reduces the computation time.

REFERENCES

Daramola O. Abosede, Adetunmbi A. Olusola, AdeolaS. Oladele,. "Analysis of KDD'99 Intrusion Detection Dataset for Selection of Relevance Features", Proceedings of the World Congress on Engineering and Computer Science, Vol. I, October 20-22, 2010.

Devale.P.R,Garje.G.V.,SnehalA.Mulay, 2012. "Intrusion Detection System using Support Vector Machine and Decision Tree ", International Journal of Computer (0975 – 8887), Vol. 3, June 2010.

Debar, H., Becker, M. and Siboni, D. "A Neural Network Component for an Intrusion Detection System",IEEE Symposium on Research in Computer Security and Privacy, pp. 240-250, 1992.

Du Hongle, Teng Shaohua and Zhu Qingfang, "Intrusion detection Based on Fuzzy support vector machines", International Conference on Networks Security, Wireless Communications and Trusted Computing, pp. 639-642, 2009.

Dewan Md. Farid, Jerome Darmont, Nouria Harbi, Nauyen HuuHoa, Mohammad Zahidur Rahman. "Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification", International Conference on Computer Systems Engineering, version 1 - 19, 2010.

Farid D.M, Jerome Dormont, NouriaHarbi, Nguyen HuuHoa and Rahman, M.Z. "Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification", International Conference on Computer Systems Engineering, Version 1, pp. 321-337, 2010.

Geetha Ramani R, Siva Sathya S, Sivaselvi K. "Discriminant Analysis based Feature Selection in KDD Intrusion Dataset", International Journal of Computer Applications (0975 – 8887),Vol. 31, No.11, 2011.

Leng J, Valli C, and Armstrong L. "A Wrapper-based Feature Selection for Analysis Large Data Set", Proceedings of 2010 3rd International Conference onand Electrical Engineering (ICCEE), pp. 167-170, 2010.

Moradi M and Zulkernine M "A Neural Network based System for Intrusion Detection and Classification of Attacks", Proceedings of IEEE International Conference on Advances in Intelligent Systems – Theory and Applications, Luxembourg, Vol. 148, pp. 1-6, 2004.

Sarasamma S., Zhu, Q. and Huff, J. "Hierarchical Kohonen Net for Anomaly Detection in Network Security", IEEE Transactions on System, Man, Cybernetics, Part B, Cybernetics, Vol. 35, No. 2, pp. 302-312, 2005.

Sindhu, .S, Geetha, S. and Kannan, A. "Decision Tree based Light Weight Intrusion Detection using a Wrapper Approach", Expert Systems with Applications, Vol. 39, pp. 129–141, 2012.

Shi-Jinn Hornga. B, Ming-Yang Suc, Yuan-HsinChenb, Tzong-WannKaod, Rong-JianChenb, Jui-Lin Laib, Citra DwiPerkasaa."A Novel Intrusion Detection System based on Hierarchical Clustering and Support Vector Machines". Science Direct, Expert Systems with Applications, Vol.38, pp.306–313, 2011.

Snehal A. Mulay, Devale, P.R. and Garje, G.V. "Intrusion Detection System using Support Vector Machine and Decision Tree", International Journal of Computer Applications, Vol.3, pp.975-987, 2010.

Stuart Russell and Peter Norvig, "Artificial Intelligence", Pearson Education, 2003.

Tan, P. N., Steinback, M. and Kumar, V. "Introduction to Data Mining", Addison Wesley, 2006.

Wang Jianping, Chen Min and Wu Xianwen, "A Novel Network Attack Audit System based on Multi-Agent Technology", Physics Procedia, Elsevier, Vol. 25,pp. 2152 – 2157, 2012.

Wei Lu, Mahbod Tavallaee, Ebrahim Bagheri, Alia A.Ghorbani,."A Detailed Analysis of the KDD CUP 99 Data Set", Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications, Vol. 97,pp.4244-37641, 2009.

Weka software, Machine Learning. "Weka 3–Data Mining with Open Source Machine Learning Software in Java" Machine Learning Group at University of Waikato Website, http://www.cs.waikato.ac.nz/ml/weka/, accessed May 2012.

Wei Wang, Xiangliang Zhang, Sylvain Gombault and Svein J. Knapskog, "Attribute Normalization in Network Intrusion Detection", 10th International Symposium on Pervasive Systems, Algorithms, and Networks, pp. 543-559, 2009.