International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Building an Intrusion Detection System Using a Filter Based Fast Co-Relation Algorithm

[1]**Mrs Y.Swapna (**Assistant Professor**)**-Yenugulaswapna@gmail.com

[2]**J.Mounika** -mounikareddyjakka1@gmail.com

[3]**P.Revanth Reddy**-Revanthreddy194@gmail.com

[4]**P.Sanhith Kumar**-saisanhith09@gmail.com

[1, 2, 3,4]Dept.of Computer Science and Engineering, Vignan Institute of Technology and Science, Deshmukhi, Hyderabad. 508284

***Abstract**:*

Redundant and irrelevant features in data have caused a long-term problem in network traffic classification. These features not only slow down the process of classification but also prevent a classifier from making accurate decisions, especially when coping with big data. In this paper, we propose a filter based fast Correlation-based feature selection (cfs): cfs searches feature subsets according to the degree of redundancy among the features. The evaluator aims to find the subsets of features that are individually highly correlated with the class but have low inter-correlation. The subset evaluators use a numeric measure, such as conditional entropy, to guide the search iteratively and add features that have the highest correlation with the class...

Keywords

Network, Spam, Social Media, behavioral, user-linguistic...

I. Introduction

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase re-venue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

- How Data Mining Works?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- Associations: Data can be mined to identify associations. The beer -diaper example is an example of associative mining.

- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

1. Extract, transform, and load transaction data onto the data warehouse system.

2. Store and manage the data in a multidimensional database system.

3. Provide data access to business analysts and information technology professionals.

4. Analyze the data by application software.

5. Present the data in a useful format, such as a graph or table.

6. Different levels of analysis are available:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k=1). Sometimes called the k-nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Data visualization: The visual interpretation of complex relationships in multidimensional data.

Graphics tools are used to illustrate data relationships.

Characteristics of Data Mining:

1. Large quantities of data: The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.

2. Noisy, incomplete data: Imprecise data is the characteristic of all data collection.

3. Complex data structure: conventional statistical analysis not possible

4. Heterogeneous data stored in legacy systems

Benefits of Data Mining:

1. It’s one of the most effective services that are available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them

2. An analytical CRM model and strategic business related decisions can be made with the help of data mining as it helps in providing a complete synopsis of customers

3. An endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns)

4. Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors

5. Fast paced and prompt access to data along with economic processing techniques have made data mining one of the most suitable services that a company seek

Advantages of Data Mining:

A. Marketing / Retail:

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers. Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

B. Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer’s data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card’s owner.

C. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi -conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

D. Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

E. Law enforcement:

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

F. Researchers:

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects.

II. Literature review

- Traffic-aware design of a high speed fpga network intrusion detection system, Computers Security of today's networks heavily relies on network intrusion detection systems (NIDSs). The ability to promptly update the supported rule sets and detect new emerging attacks makes field - programmable gate arrays (FPGAs) a very appealing technology. An important issue is how to scale FPGA-based NIDS implementations to ever faster network links. Whereas a trivial approach is to balance traffic over multiple, but functionally equivalent, hardware blocks, each implementing the whole rule set (several thousand rules), the obvious cons is the linear increase in the resource occupation. In this work, we promote a different, traffic-aware, modular approach in the design of FPGA-based NIDS. Instead of purely splitting traffic across equivalent modules, we classify and group homogeneous traffic, and dispatch it to differently capable hardware blocks, each supporting a (smaller) rule set tailored to the specific traffic category. We implement and validate our approach using the rule set of the well-known Snort NIDS, and we experimentally investigate the emerging trade-offs and advantages, showing resource savings up to 80 percent based on real-world traffic statistics gathered from an operator's backbone.
- Network-based intrusion detection with support vector machines.

This paper proposes a method of applying Support Vector Machines to network-based Intrusion Detection System (SVM IDS). Support vector machines (SVM) is a learning technique which has been successfully applied in many application areas. Intrusion detection can be considered as two-class classification problem or multi-class classification problem. We used dataset from 1999 KDD intrusion detection contest. SVM IDS was learned with triaing set and tested with test sets to evaluate the performance of SVM IDS to the novel attacks. And we also evaluate the importance of each feature to improve the overall performance of IDS. The results of experiments demonstrate that applying SVM in Intrusion Detection System can be an effective and efficient way for detecting intrusions.

- An effective technique for intrusion detection using neuro-fuzzy and radial svm classifier

Intrusion detection is not yet a perfect technology. This has given data mining the opport unity to make several important contributions to the field of intrusion detection. In this paper, we have proposed a new technique by utilizing data mining techniques such as neuro-fuzzy and radial basis support vector machine (SVM) for the intrusion detection system. The proposed technique has four major steps in which, first step is to perform the Fuzzy C-means clustering (FCM). Then, neuro-fuzzy is trained, such that each of the data point is trained with the corresponding neuro-fuzzy classifier associated with the cluster. Subsequently, a vector for SVM classification is formed and in the fourth step, classification using radial SVM is performed to detect intrusion has happened or not. Data set used is the KDD cup 99 dataset and we have used sensitivity, specificity and accuracy as the evaluation metrics parameters. Our technique could achieve better accuracy for all types of intrusions. It achieved about 98.94 % accuracy in case of DOS attack and reached heights of 97.11 % accuracy in case of PROBE attac k. In case of R2L and U2R attacks it has attained 97.78 and 97.80 % accuracy respectively. We compared the proposed technique with the other existing state of art techniques. These comparisons proved the effectiveness of our technique.

- Intrusion detection using an ensemble of intelligent paradigms.

Soft computing techniques are increasingly being used for problem solving. This paper addresses using an ensemble approach of different soft computing and hard computing techniques for intrusion detection. Due to increasing incidents of cyber-attacks, building effective intrusion detection systems are essential for protecting information systems security, and yet it remains an elusive goal and a great challenge. We studied the performance of Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and Multivariate Adaptive Regression Splines (MARS). We show that an ensemble of ANNs, SVMs and MARS is superior to individual approaches for intrusion detection in terms of classification accuracy.

- A new approach to intrusion detection based on an evolutionary soft computing model using neuro fuzzy classifiers

An intrusion detection system's main goal is to classify activities of a system into two major categories: normal and suspicious (intrusive) activities. Intrusion detection systems usually specify the type of attack or classify activities in some specific groups. The objective of this paper is to incorporate several soft computing techniques into the classifying system to detect and classify intrusions from normal behaviors based on the attack type in a computer network. Among the several soft computing paradigms, neuro-fuzzy networks, fuzzy inference approach and genetic algorithms are investigated in this work. A set of parallel neuro-fuzzy classifiers are used to do an initial classification. The fuzzy inference system would then be based on the outputs of neuro-fuzzy classifiers, making final decision of whether the current activity is normal or intrusive. Finally, in order to attain the best result, genetic algorithm optimizes the structure of our fuzzy decision engine. The experiments and evaluations of the proposed method were performed with the KDD Cup 99 intrusion detection dataset.

III. System Design and Architecture

A. Existing System

A significant amount of research has been conducted to develop intelligent intrusion detection techniques, which help achieve better network security. Bagged boosting-based on C5 decision trees and Kernel Miner are two of the earliest attempts to build intrusion detection schemes.

Mukkamala et al. investigated the possibility of assembling various learning methods, including Artificial Neural Networks (ANN), SVMs and Multivariate Adaptive Regression Splines (MARS) to detect intrusions.

Disadvantages Of Existing System:

Existing solutions remain incapable of fully protecting internet applications and computer networks against the threats from ever-advancing cyber-attack techniques such as DoS attack and computer malware.

Current network traffic data, which are often huge in size, present a major challenge to IDSs. These "big data" slow down the entire detection process and may lead to unsatisfactory classification accuracy due to the computational difficulties in handling such data.

i. Classifying a huge amount of data usually causes many mathematical difficulties which then lead to higher computational complexity.

ii. Large-scale datasets usually contain noisy, redundant, or uninformative features which present critical challenges to knowledge discovery and data modeling.

B. Proposed System:

We have proposed a Filter Based Fast Correlation algorithm. This algorithm consists of two phases.

i. The upper phase conducts a preliminary search to eliminate irrelevant and redundancy features from the original data. This helps the wrapper method (the lower phase) to decrease the searching range from the entire original feature space to the pre-selected features (the output of the upper phase). The key contributions of this paper are listed as follows.

ii. This work proposes a new filter-based feature selection method, in which theoretical analysis of mutual information is introduced to evaluate the dependence between features and output classes.

iii. The most relevant features are retained and used to construct classifiers for respective classes. As an enhancement of Mutual Information Feature Selection (MIFS) and Modified Mutual Information based Feature Selection (MMIFS), the proposed feature selection method does not have any free parameter, such as in MIFS and MMIFS. Therefore, its performance is free from being influenced by any inappropriate assignment of value to a free parameter and can be guaranteed. Moreover, the proposed method is feasible to work in various domains, and more efficient in comparison with HFSA, where the computationally expensive wrapper-based feature selection mechanism is used.

iv. We conduct complete experiments on two well-known IDS datasets in addition to the dataset used. This is very important in evaluating the performance of IDS since KDD dataset is outdated and does not contain most novel attack patterns in it. In addition, these datasets are frequently used in the literature to evaluate the performance of IDS. Moreover, these datasets have various sample sizes and different numbers of features, so they provide a lot more challenges for comprehensively testing feature selection algorithms.

v. Different from the detection framework proposed that designs only for binary classification, we design our proposed framework to consider multiclass classification problems. This is to show the effectiveness and the feasibility of the proposed method.

Advantages Of Proposed System:

- FMIFS is an improvement over MIFS and MMIFS.
- FMIFS suggests a modification to Battiti's algorithm to reduce the redundancy among features.
- FMIFS eliminates the redundancy parameter required in MIFS and MMIFS.

IV. Results

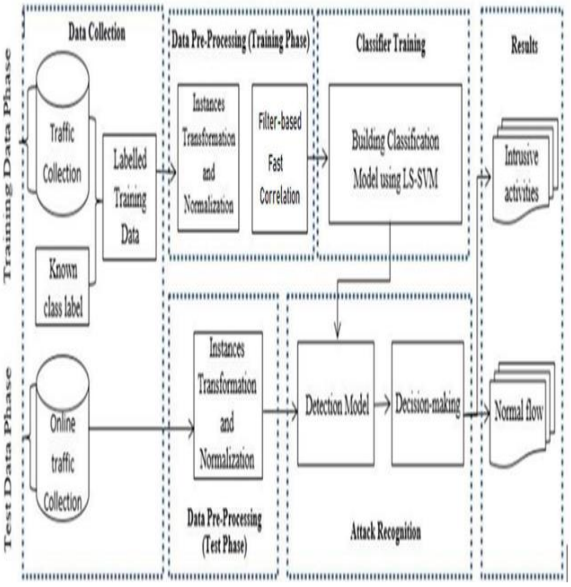
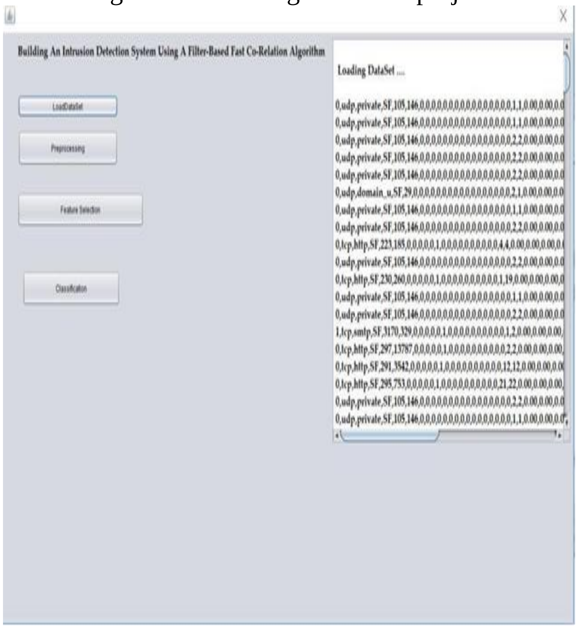

Data Collection
Data Pre-Processing (Training Phase)
Classifier Training
Results
Training Data Phase
Traffic Collection
Labelled Training Data
Known class label
Instances Transformation and Normalization
Filter-based Fast Correlation
Building Classification Model using LS-SVM
Intrusive activities
Test Data Phase
Online traffic Collection
Instances Transformation and Normalization
Data Pre-Processing (Test Phase)
Detection Model
Decision-making
Attack Recognition
Normal flow

Fig 1: Content Diagram of the project

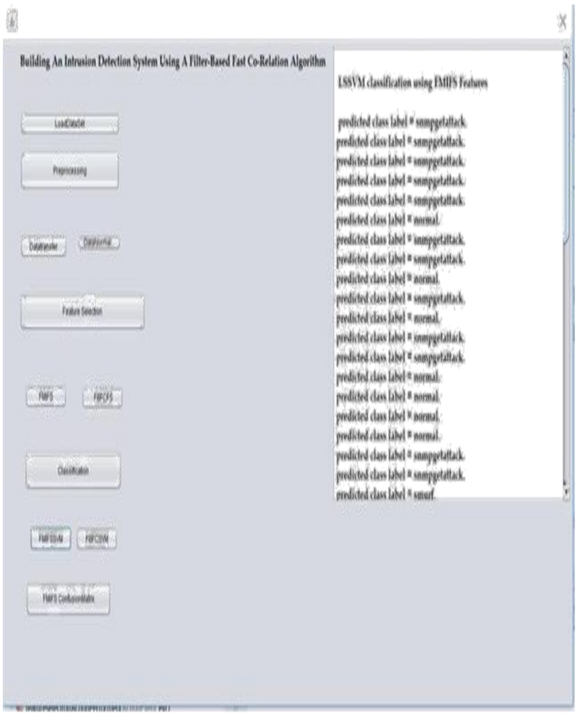


Building An Intrusion Detection System Using A Filter-Based Fast Co-Relation Algorithm

```
Loading DataSet ....

0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.0
0,udp,domain_u,SF,29,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,1,0.00,0.00,0.0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.0
0,tcp,http,SF,223,185,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,4,4,0.00,0.00,0.00,0.0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.0
0,tcp,http,SF,230,260,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,19,0.00,0.00,0.00,0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.0
1,tcp,smtp,SF,3170,329,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,2,0.00,0.00,0.00,
0,tcp,http,SF,297,13787,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,
0,tcp,http,SF,291,3542,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,12,12,0.00,0.00,0.0
0,tcp,http,SF,295,753,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,21,22,0.00,0.00,0.00,
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.0
0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.0
```

Fig.2 Loading dataset

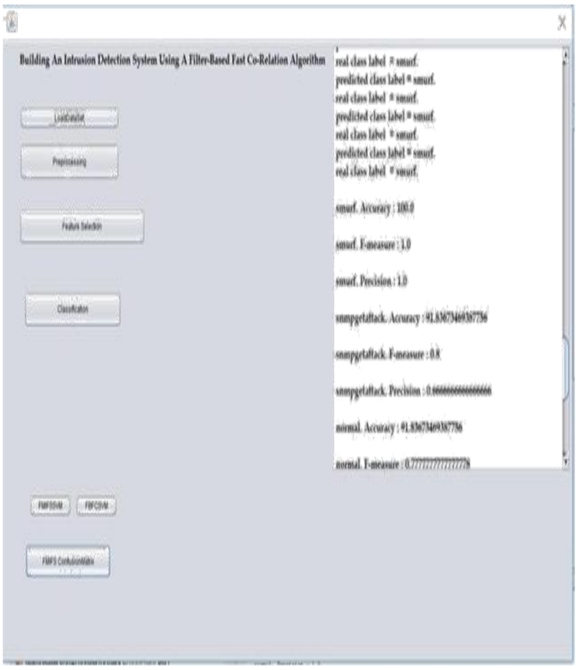


Building An Intrusion Detection System Using A Filter-Based Fast Co-Relation Algorithm

```
LSSVM classification using FMIFS Features

predicted class label = snmpgetattack.
predicted class label = snmpgetattack.
predicted class label = snmpgetattack.
predicted class label = snmpgetattack.
predicted class label = snmpgetattack.
predicted class label = normal.
predicted class label = snmpgetattack.
predicted class label = snmpgetattack.
predicted class label = normal.
predicted class label = snmpgetattack.
predicted class label = normal.
predicted class label = snmpgetattack.
predicted class label = snmpgetattack.
predicted class label = normal.
predicted class label = normal.
predicted class label = normal.
predicted class label = normal.
predicted class label = snmpgetattack.
predicted class label = snmpgetattack.
predicted class label = smurf.
```

Fig.4 FMIFS classification



Building An Intrusion Detection System Using A Filter-Based Fast Co-Relation Algorithm

```
real class label = smurf.
predicted class label = smurf.
real class label = smurf.
predicted class label = smurf.
real class label = smurf.
predicted class label = smurf.
real class label = smurf.

smurf. Accuracy : 100.0

smurf. F-measure : 1.0

smurf. Precision : 1.0

snmpgetattack. Accuracy : 91.83673469387756

snmpgetattack. F-measure : 0.8

snmpgetattack. Precision : 0.6666666666666666

normal. Accuracy : 91.83673469387756

normal. F-measure : 0.7777777777777778
```

Fig.5 Confusion matrix

V. Conclusion

Recent studies have shown that two main components are essential to build an IDS. They are a robust classification method and an efficient feature selection algorithm. In this paper, a supervised filter-based feature selection algorithm has been proposed, namely Flexible Mutual Information Feature Selection (FMIFS). FMIFS is an improvement over

MIFS and MMIFS. FMIFS suggests a modification to Battiti's algorithm to reduce the redundancy among features. FMIFS eliminates the redundancy parameter _ required in MIFS and MMIFS. This is desirable in practice since there is no specific procedure or guideline to select the best value for this parameter.

FMIFS is then combined with the LSSVM method to build an IDS. LSSVM is a least square version of SVM that works with equality constraints instead of inequality constraints in the formulation designed to solve a set of linear equations for classification problems rather than a quadratic programming problem. The proposed LSSVMIDS + FMIFS has been evaluated using three well known intrusion detection datasets: KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets. The performance of LSSVM-IDS + FMIFS on KDD Cup test data, KDDTest+ and the data, collected on 1, 2 and 3 November 2007, from Kyoto dataset has exhibited better classification performance in terms of classification accuracy, detection rate, false positive rate and F-measure than some of the existing detection approaches. In addition, the proposed LSSVM-IDS + FMIFS has shown comparable results with other state-of the-art approaches when using the Corrected Labels sub-dateset of the KDD Cup 99 dataset and tested on Normal, DoS, and Probe classes; it outperforms other detection models when tested on U2R and R2L classes. Furthermore, for the experiments on the KDDTest 21 dataset, LSSVM-IDS + FMIFS produces the best classification accuracy compared with other detection systems tested on the same dataset. Finally, based on the experimental results achieved on all datasets, it can be concluded that the proposed detection system has achieved promising performance in detecting intrusions over computer networks. Overall, LSSVM-IDS + FMIFS has performed the best when compared with the other state-of the- art models.

Future Work:

Although the proposed feature selection algorithm FMIFS has shown encouraging performance, it could be further enhanced by optimizing the search strategy. In addition, the impact of the unbalanced sample distribution on an IDS needs to be given a

.

VI. References

- S. Pontarelli, G. Bianchi, S. Teofili, Traffic-aware design of a high speed fpga network intrusion detection system, Computers, IEEE Transactions on 62 (11) (2013) 2322–2334.

- B. Pfahringer,Winning the kdd99 classification cup: Bagged boosting, SIGKDD Explorations 1 (2) (2000) 65–66.

- Levin, Kdd-99 classifier learning contest: Llsoft's results overview, SIGKDD explorations 1 (2) (2000) 67

- D. S. Kim, J. S. Park, Network-based intrusion detection with support vector machines, in: Information Networking, Vol. 2662, Springer, 2003, pp. 747–756.

- Chandrasekhar, K. Raghuveer, An effective technique for intrusion detection using neuro-fuzzy and radial svm classifier, in: Computer Networks & Communications (NetCom), Vol. 131, Springer, 2013, pp. 499–507.

- S. Mukkamala, A. H. Sung, A. Abraham, Intrusion detection using an ensemble of intelligent paradigms, Journal of network and computer applications 28 (2) (2005) 167–182.

- N. Toosi, M. Kahani, A new approach to intrusion detection based on an evolutionary soft computing model using neurofuzzy classifiers, Computer communications 30 (10) (2007) 2201–2212.

- Z. Tan, A. Jamdagni, X. He, P. Nanda, L. R. Ping Ren, J. Hu, Detection of denial -of-service attacks based on computer vision techniques, IEEE Transactions on Computers 64 (9) (2015) 2519–2533