ISSN: 0974-5823 Vol. 7 No. 1 January, 2022

International Journal of Mechanical Engineering

A Survey onData Mining Techniques used inNetwork Security Particularly in Intrusion Detection Mechanism

A.Kaveri
Research Scholar , Bharathiar University , Coimbatore. kaverimca.a@gmail.com,
Dr. M. Punithavalli,.
Bharathiar University, Coimbatore. punithavalli@buc.edu.in

Abstract

In modern communication and networking world, network security is an important role for securing our data and resources. This paper discussed about the intrusion detection system and the various data mining algorithms used for effective intrusion detection system. Standard conventional intrusion detection system is compared with new efficient IDS approaches using data mining techniques. To get accurate evaluation results, the data to be supplied for the detection must be error free and pre processed. So that, in this paper data pre processing techniques and feature selection techniques discussed. In Intrusion Detection System, when we including data mining techniques, after data pre processed, the core process is classification. Classification is preferred when there is enough training data is available. When there is no training data, clustering method is preferable. For intrusion detection system mostly used data mining algorithms for classification and clustering is discussed. Particularly K means clustering and its improved version discussed. one more method statistical model also discussed and its performance results were discussed.

Index Terms: Data Mining , Classification Intrusion Detection , Network Security

Introduction

In Networking, more devices are interconnected for the purpose of making communication and sharing the resources. The devices on a network connected via cables, satellites, radio waves and infrared light beams etc. Networking refers to the task of creating and using collection of devices in order to hardware, software, protocols may include wired technology or wireless. Computer networking uses for the terms such as the file sharing in which the users are exchange their files, hardware sharing where the devices like printer , scanner , hard drives can be shared , application sharing , communication like mail , video conferencing , text , newsgroups etc. Without networking the above said sharing the information and resources not possible. So networking occupies main role in our communication world.

Network security is important factor to make a network can be a secured one and to avoid unauthorized involvements in communication to safeguard data. Network security is a very broad area which covers a large number of technologies, processes and devices. Here the work done for the purpose of making confidentiality, integrity and accessibility of network data and resources using both software and hardware technologies by specifying set of rules or protocols.Everywhere regardless of size, the organization requires a high degree of network security to secure it from the ever-growing landscape of cyber threats in the wild today. High priority should be given to network security for any organization where works with networked data and systems takes place. Addition to protecting data and resources, the integrity of network data from external exploits, network security mechanism can also efficiently manage network traffic, enhance the network performance and ensures that the secure data only sharing between the authorized users and data sources.

Network security can be implemented as three different types of controls: physical type, technical type and administrative security control.Physical security controls are focused to protect unauthorized persons from getting physical access to network components like cabling cupboards, routers etc. This type of mechanisms

works with locks, biometric authentication and other electronic devices. This type of physical security control is highly needed for any organization. Technical security method protects data which is placed on the network or that is in transit across, into or out of the particular network. This method is twofold type that protect the data and system devices from unauthorized personnel and also to protect against malicious activities from the users.Administrative security level controls have security policies and processes which control network user's behaviour consists how users are authentication mechanism , user's level of access and also changes to the infrastructure.

There are many different ways to secure the network such as antivirus and antimalware software, network access control, virtual private networks, firewall protection , network segmentation, behavioral analytics, data loss prevention, intrusion prevention systems etc. Antivirus and antimalware software protect a network from a range of malicious software such as ransom ware, viruses, worms and trojans. The best network security software not only scans the data files upon entry to the network but again and again continuously scans and tracks the data files. Virtual private networks build a connection to the network from another side endpoint or site. For this the following thing can be an exact example that is workers working from home can typically get the connection with their company's network over a VPN. The Network sharing Data between the two points can be encrypted and the user needs to authenticate to accept the communication between their system devices and the network. Network access control is most granular level in the concept of Network security.Not every user has the access rights to enter and work within a network. To move out potential attackers, there is a need to recognize each user and each device connected in the network. So that we have to enforce security policies and can block noncompliant endpoint network devices or give the permission only for the limited access. This type of prevention can be called as network access control. Firewall act as a barrier between the side of untrusted external type networks and another sideof trusted internal network. Administrators usually configure a set protocols or rules which blocks or permits entry onto the network.

Within a network there is a need to ensure that network users do not send sensitive data outside the particular network. Data loss prevention mechanism can stop users from forwarding, uploadingor even printing confidential information in an unsafe manner. For finding abnormal network behaviour,behavioural analytics tools find activities that differentiate from the normal. Better identify indicators of compromise that pose big issues and immediately remediate the network threats. Software defined network segmentation places network traffic into different types and makeseasier of security policies. The different classifications are based on identity of endpoints, not based on IP addresses. Here the network security assign access rights based on role, location, and some other things so that the correct level of secured access is given to the right users and suspicious user devices are contained and immediately remediated.An intrusion prevention mechanism scans the network traffic to actively block attacks and track the progression of suspect files and malware across the network to control the spread of outbreaks and reinjection in the network. This can be done by correlating huge global threat intelligence system.

Intrusions in Networking

An intrusion detection system is a software application or a networking device that monitors a network if any malicious activities or network policy violations. Suppose is there any malicious activity or violation is reported thatreceived centrally using a particular security information and event management system. Some intrusion detection mechanisms are able to responding to find intrusion upon discovery. These are specifically classified as intrusion prevention system which is shortly known as IPS.Followings are the various type of intrusion attacks such as Multi-Routing, Living Off the Land type attacks, Buffer Overwriting, Covert CGI Scripts, Protocol-Specific Attacks, Traffic Flooding, Trojan Horse Malware, Worms etc. The followings are the three different practices which are mostly used to circumvent cyber security and network intrusion detection systems are encryption on departing data, deleting logs andinstalling rootkits. Generally there are 2 types of intrusion detection mechanism that is network based and host based. Network intrusion detection systems analyze incoming network traffic. A host based intrusion detection system monitors potential operating system files.

In networking, intrusion refers an unauthorized activity happened in a network. Network intrusions often involve for the purpose of stealing valuable network software or hardware resources and always jeopardize the protection of networks and data.In our modern networking world intrusion happened frequently. Intruders

ideally use the different version of commands and thereby removing their footprints in the process of network auditing and log files. An Effective Intrusion Detection mechanism identifies intrusive and nonintrusive entries. In 1980 James Anderson introduced the intrusion detection mechanism. In conventional systems have breaches which makes easily vulnerable and it could not be solved. Substantial research is doing in the area of intrusion detection technology that is considered not a suitable tool and against intrusion. So intrusion detection is the most challenging task in the concept of network security. (G.V. Nadiammai, M. Hemalatha 2014)

Intrusion detection Technique starts with instrumentation of a network for data collection. Pattern-based software consist sensors monitor the particular network traffic and raise notifications when the traffic matches with a saved pattern. Security analysts confirm these alarms notify an event serious enough to warrant a response. The response might be to shut down a particular portion of the network, simply make note of unusual network traffic and the internet service provider with suspicious traffic. If the particular network size is small and signatures are kept up to date, the analyst solution to intrusion detection works properly. But when network size is large, complex network the analysts become overwhelmed by the number of notification arms they need to be reviewed. This type of Commercial intrusion detection software collections tend to be signature oriented with little or no state data maintained. These restrictions led us to find the application of data mining to solve this problem. (Eric Bloedorn, Alan D. Christiansen)

Intrusion Detection before Data Mining

The intrusion detection on earlier days not focused on the data mining concept. The intrusion detection starts with following aspects. Those are the performance of sensor, data quantity, data to be displayed, kind of data and basis for highlighting the data. Next, as the supplied data came in, sensor tuning work, incident investigation task and system performance commanded takes place. The analyst crew grew to handle the data load, and training and team coordination were the problems of the day. But the level of analyse and attack on the internet was constantly changing, along with the size of data we were collecting and involving in front of our analysts. We began to suspect that our system which is used for intrusion detection was inadequate for finding the most dangerous network attacks. Those performed by adversaries using the network attacks which are stealthy, new or both type. So we considered the new approach data mining for the purpose of intrusion detection mechanism. (Eric Bloedorn, Alan D. Christiansen)

Data Mining

The term Data mining is a process of extracting data and finding patterns in huge data sets involving methods at the intersection of statistics, machine learning, and database systems. It is an interdisciplinary field of computer science and statistics with an overall aims to extract data from a large data set and converts the information into a suitable structure for the purpose of further use. It is the analysis step of the term KDD which is expanded as "Knowledge Discovery in Databases" process. Data mining consisting database management concepts, data pre-processing techniques, inference considerations model, complexity considerations, interestingness metrics, post-processing of discovered structures, visualization techniques.The process of extracting data to find patterns, trends, data that used to take the data driven decision from large sets of data which is called Data Mining.

Data mining which is known as KnowledgeDiscovery is the taskof automatically searching hugevolumes of data for patterns with association rules. It is a recent concept in computer science it utilizes many conventional computational methods from statisticsmachine learning, information retrieval and pattern recognition. The followings are the few specific things that in data miningwhich contributes to an intrusion detection scheme. 1. The process of removing normal activity from the alarm data to allow security analysts to aims on real attacks. 2. Find false alarm typecreations and wrong sensor signatures. 3. Identify different anomalous activity which uncovers a real network attacks. 4. Find long or ongoing patterns.

To do the said tasks, data mining analysts have one or more of the following methods. 1. Data summarization with the concept of statistics also consist step to finding outliers. 2. Visualization that providing a graphical view of the data supplied. 3. Clustering the data into various categories. 4. Association rule mining which defining the normal activity and also enabling the discovery of anomalies in the given data set. 5.

Classification is a work of predicting the category to which a specific data belongs. (Ankit Naik, S.W. Ahmad, 2015)

Intrusion Detection using Data Mining Techniques

Intrusion Detection System using Data mining can efficiently find the data of user interest and predicts the results which can be referred in the future. Data mining tasks or knowledge discovery in real time databases achieved great deal of attention in computer industry as well as in the society. Data mining techniques has been involved to extract the useful information from huge volumes of data that may noisy, fuzzy and dynamic.

Intrusion detection system has been occupied centrally to receive all the incoming data packets which are transmitted over the network. Data are collected first and then send for data pre-processing for eliminating the noise, irrelevant data and missing values are replaced. Then feature selection is made to concentrate only on potential data. After that the pre processed data can be analyzed and classified based to their severity measures. If the data is normal, then there is no need for any more change or else it send for intrusion report generation to raise alarms. Based on the characteristic of the data, alarms are raised to notify to the administrator for handle the situation in advance manner. The type of network attack is modelled so as to enable the classification task of network data. When all these process happening means as soon as the network transmission starts.(G.V. Nadiammai, M. Hemalatha, 2014)

The supervised learning basednetwork intrusion detection systemuses the advantages of supervised learning and also prediction techniques. Also the step by step procedure of building a predictive model using data mining techniques as follows;

First Step:

Collecting network historical data and log data.

Second Step:

a. Data Preprocessing

b. Feature Selection

c. Implementing supervised learning algorithms forhistorical data or log data for making the predictive model

Third Step: Data Analysis using Mining Tools

Fourth Step: Using Network Security Techniques, Mechanisms and Protocols

Fifth Step: Intrusion Detection System

This method of intrusion detection explored and analyzed the different challenges of threats and network attacks in computer networking in this modern era, different type of network sniffing, snooping tools available for collecting the network data and log data for the purpose of analysis and learning, different data mining tools for learning and generating the predictive models and various supervised learning method classification algorithms for learning the network data and find the behaviour of the network attackers and hackers. (D.Asir Antony Gnana Singh, E.JebamalarLeavline, 2013)

Data Mining Techniques

Data mining techniques are becoming most important component in the process of intrusion detection system. Various data mining tasks like data pre processing , feature selection clustering, classification etc are frequently used to analyze supplied data to achieve intrusion related knowledge. This following will elaborate on different data mining methods and will explain how those techniques are used in the task of intrusion detection.(Luan J. 2002)

Data pre processing

Data pre processing in data mining refers to the technique that is used to convert raw data into efficient format. It is a crucial step in the data mining process. It increases data efficiency. It involves the extraction of various

patterns and models from a huge dataset. Data has to be processed before being mined which makes it an integral step in the data mining process. Data pre-processing involves several processes like data cleaning, integration, transformation and reduction.

Data cleaning is the process that used to organize inaccurate and poorly formatted data. It involves filling missing values, altering noisy data and resolving inconsistencies.Data integrationis a technique that involves combining data from multiple diversified data sources into a legible data format. It brings data from various sources together to provide the user with a unified view. Data transformation is the process used to convert data from one format to another. The conversion generally happens from the format of a source system into the required format of the destination system. Data reduction this process is used to provide a condensed description of the original data. The reduced data is expected to show same results as the original data. This significantly reduces the quantity of data while maintaining the quality of data.Data pre-processing focuses on the following two major issues:

1. The proper organisation of data for the efficient execution of data mining algorithms.
2. Data sets used must result in better quality and performance of models generated by performing data mining operations.

Feature Selection:

Features selection is the process used to reduce the number of inputs for processing and analysis. It helps to find the most meaningful inputs. The irrelevant attributes present in the data to be mined have to be removed for better execution of mining algorithms.The goal of data feature selection is to select the features that are rich in discriminatory information while focusing on the classification problem.Data contains many features but all its features cannot be used at once. This makes feature selection very essential. The feature selection techniques can be categorized into three types: Filter methods, Wrapper methods, and embedded methods. Every feature selection algorithm uses any one of these three feature selection techniques.

Classification in Data Mining

Classification algorithms in data mining can be used for misuse detection and anomaly detections. In the work of misuse detection, the network traffic data can be collected and labelled as follows normal or intrusion. After this labelled dataset is involved as a training data to know classifiers of various types that can be used to find known defect intrusions. In the task of anomaly detection, the normal behaviour is taken from the training dataset that are called as normal using learning algorithms. Classification algorithms in data mining can be created using various algorithms. Classification classifies the supplied data records in a predefined set of classes used as property to a label each data record; differentiatingdata elements depends to normal or abnormal class. This method has been very famous to detect network attacks but has to be implemented with complementary fine-tuning methods to minimize its high false positives data rate.

Clustering

Clustering tasks in data mining is the process of labelling the supplied data and arranged it into groups. Clustering algorithms can group given data instances into a similar type of groups. These clustering groups can be used to maximize the performance of existing classifiers. High quality clusters can also helps user expert with labelling. (NeelamadhabPadhy, Dr.Pragnyaban Mishra and RasmitaPanigrahi, 2012)

The main aim of association analysis in data mining is to find association relationships between data values features in huge datasets. This helps to find hidden patterns and has a wide range of applications in research and also business type concepts. Association rule mining can helps for discriminating data attributes which are highly useful in the intrusion detection mechanism. Network Intrusions and malicious network attacks are naturally dynamic. Addition to that data streams helps to find intrusions in the sense that an event may be normal. So it is important to perform network intrusion detection in data stream, the environment of real-time.

Stream data analysis helps to identify sequences of data which are continuously encountered, find sequential data patterns and detect outliers in the data set. Network Intruders may from several different physical locations and attack many different physical destinations. Distributed data mining methods can be utilized to analyze network data from several network locations, this type of mechanism helps to find distributed network

attacks and avoid attackers in different places from harming our data and resources. (Ankit Naik, S.W. Ahmad , 2015)

Data Pre processing in predictive Data Mining: Data preprocessing is essential for the performance of generalized machine learning algorithm. Many algorithms such as learning algorithm provide efficient performance only with categorical instances. The data preprocessing steps utilised in this paper is as follows. Initially, the information is collected from a source. Then, the preprocessing steps are applied to the sample to make it legible. Now, this data is given as an input to the learning algorithm which is then applied to solve a IDS problem The algorithms that come under learning algorithms are linear models, neutral networks, lazy learners, decision trees, and Bayesian, SVM and rule learners. The reader may encounter diversifying which will have to do with algorithms that only manage datasets consisting of numbers, such as SVM or neutral network. In this case, the data has to be normalized instead of being discretized. But in other cases, any preprocessing step can be used for any algorithm to get desired results.

Data Preprocessing - A Preliminary Step for Data Mining: Big Data requires large computational infrastructure with high performance processing capabilities. Preparing data for mining and its analysis is a challenging task and demands data to be pre processed to improve data quality. Data preprocessing is primary data mining practice in which raw data is transformed into a format suitable for other processing procedure. Data preprocessing enhances the data quality by cleaning, normalizing, transforming and extracting relevant features from raw data. It significantly enhances the performance of machine learning algorithms which in turn leads to accuracy in data mining. Knowledge discovery from noisy, irrelevant and redundant data is a difficult task, hence, precision in identification of extreme values and outlier and filling up missing values poses challenges. Its about discusses various big data pre processing techniques which help to prepare it for mining and analysis tasks.Any data analysis algorithm would not be able to discover hidden pattern or trend from data if the dataset under observation is inadequate, irrelevant or incomplete. Thus data preprocessing is a central process in any data analysis process. The preprocessing of data solves various kinds of problems such as noisy data, redundancy, missing values, etc. High quality of results can only be achieved with high quality of data which in turn also reduces the cost for data mining. The foundation of decision making system in any organization comprises of the three C's properties of data i.e. Completeness, Consistency and Correctness. Deprived quality of data quality effects decision making process which eventually decreases customer's satisfaction. Furthermore larger dataset affects the performance of any machine learning algorithm, therefore instance selection reduces data and is an efficient approach to make machine learning algorithm work effectively.

Preprocessing Methods and Pipelines of Data Mining: Data mining is about fetching new knowledge from existing datasets. However, this raw data can be scattered, noisy, and even incomplete. Although lots of effort are put to develop or fine-tune the data mining models to make them more robust to the noise of the input data, their qualities still strongly depend on the quality of data. The article provides an overview of the data mining pipeline, where the procedures in a data mining task are briefly introduced. Then it provides an overview of the data preprocessing techniques which are categorized as the data cleaning, data transformation and data preprocessing. Detailed preprocessing methods and their influence on the data mining models are covered in this article.The success of a data mining model completely depends on the proper data preprocessing work. The raw data before pre processed can be of irrelevant format for model input, causing instability for the optimization algorithm of the model, having a great impact on its performance because of its noise and outliers, and causing performance problems on the training process. With careful selection of preprocessing algorithms, these problems can be reduced or avoided.

A Data Preprocessing Techniques towards Efficient and Reliable Knowledge Discovery From Building Operational Data: A comprehensive review of data preprocessing techniques for analysing massive building operational data. A wide range of data preprocessing techniques are summarised in terms of their applications in missing value imputation, outlier detection, data reduction, data scaling, data transformation, and data partitioning. Added to his, three state-of-the-art data science techniques are proposed to solve practical data challenges in the building field, i.e., data augmentation, transfer learning, and semi-supervised learning. Data cannot be fully automated due to important variations in building operating characteristics and data quality. At present, it is more like a trail and error process which is heavily dependent on domain expertise and practical tasks at hand. More research efforts have to be made towards the automation of building operational data preprocessing tasks to improve efficiency in data analysis. Meanwhile, semi-supervised learning can be used

to fully determine the hidden value in massive amounts of raw data. It is especially helpful in developing classification models for building systems, as it could be very expensive and labour-intensive to determine labels for building operational data, e.g., whether a data sample corresponds to normal or faulty operations.

Intelligent Assistance for Data Pre processing Considering all the possible pre-processing operators, a staggeringly large number of alternatives can be found. As a consequence, non-experienced users could be overwhelmed with pre-processing alternatives. In this paper, this problem has been addressed by automating the pre-processing with the support of meta-learning. To this end, a wide range of data pre-processing techniques and a set of classification algorithms have been analysed. For each classification algorithm that has been considered and a given dataset, the transformations could be automatically suggested to improve the quality of the results of the algorithm on the dataset. This approach will help non-expert users to identify the transformations and appropriate to their applications to achieve improved results. This paper includes a tool that draws on a range of classification algorithms in Weka and makes it easy for non-experts to perform data pre-processing. An extensive evaluation on hundreds of datasets confirmed that for the set of algorithms considered, even blindly applying the recommended transformations improves the final result of the algorithms on average. This can be a helpful tool for experienced users as well, because they can differentiate between the recommended transformations and choose the ones that are potentially more suitable to solve the problem at hand.

Data Preprocessing And Feature Selection For Machine Learning Intrusion Detection Systems: Flow-based anomaly detection is an affair that still flourishes in a computer network security environment. Many previous studies have applied data mining as a solution for detecting anomaly in an intrusion detection system. In this paper, data mining has been further applied to classifying those anomaly data. This is based on the facts that there are many data which cannot be used by a classification algorithm. In addition, that algorithm may use all features which are actually irrelevant to the classification target. According to these two problems, two steps have been defined: pre-processing and feature selection, whose results are classified by using k-NN, SVM, and Naive Bayes. The experimental results display that such pre-processing and combination of CFS and PSO are better to apply to SVM which is able to achieve about **99.9291**% of accuracy on KDD Cup99 dataset.

Furthermore, the evaluation is also carried out by differentiating the performance of the proposed method with another existing one, where in general, the proposed method is superior. In the future, this proposed method can be applied to other datasets. This is helpful to measure its capability to work on various characteristics of data. Also, more data reduction could be done to have simpler data. It is targeted to reduce running time and complexity.

This paper explains in details about the data reports preliminary depends on dataset, how the pre-processing data (mainly for data cleaning and reduction process) is applied to a dataset. The dataset that will be used in his paper is number of visitors to Taiwan by their residence and purpose.Accurate forecasts of demand for international tourism are essential to effectively promote tourism and to provide sufficient resources for operations, marketing, investment, and financial planning for the Taiwanese tourism industry. Although forecasting demand is essential to all industrial planning, forecasting is particularly crucial in the tourism industry because tourism products and services are inherently perishable.In this work, the default classification would be used using the Bayes Naïve algorithm. The extraction of the data becomes easy after the pre-processing has been done. Through this algorithm, we can select the best features contribute to this task before applying classification process.

Comparison of Data Preprocessing Approaches: Data preprocessing is an integral part of deep learning projects and takes up a huge part of the whole analytical pipeline. Data segmentation and data transformation are two significant steps of data preprocessing. This study analyzes the effect of segmentation methods on deep learning model performance, and compares four data transformation approaches. An experiment based on acceleration data from multiple wearable devices was implemented. The highest overall accuracy achieved was 97.20% for eight daily activities, based on the data from seven wearable sensors, which outperformed most of the other machine learning techniques.

After preprocessing, the original acceleration data segments are transformed into different types of images, to which the deep learning methods are implemented. In this study, the deep CNN algorithm is used. The results proved that this method can achieve satisfying recognition accuracy. It can help better analyze workers'

activities in a factory environment and help integrate people into the cyber-physical format in an Industry 4.0 context.

Data pre processes to hypothesis generation: Most genetic variations related to human complex traits are located in non-coding genomic regions. Hence, a comprehensive catalogue of functional non-coding genomic elements is required to understand the genotype-to-phenotype axis, most of which are involved in epigenetic regulation of gene expression. Before generating biological hypotheses with downstream analysis, scATAC-seq data has to undergo preprocessing steps for accurate interpretation. Preprocessing of scATAC-seq data begins from demultiplexing of sequence files and removing low-quality cells Harmony is a fast and scalable algorithm of single-cell data integration which is based on iterative adjustment of data-specific clusters. Recently, more approaches for data integration have been reported, including the maximum mean discrepancy manifold alignment (MMD-MA) algorithm and DeConvolution and Coupled-Clustering (DC3). Moreover, computational algorithms and software tools for scATACseq data analysis have been developed. However, algorithmic approaches and parameters for each step of the data analysis pipeline must be carefully selected for efficient translation of chromatin accessibility information into novel biological hypotheses.

Data Collection Methods and Data Preprocessing Techniques:

Knowledge Discovery in Databases facilitates organizations and researchers by turning their data collection into valuable information. Data collection and pre-processing are the most essential stages to acquire the final data that can be taken as suitable for further data mining tasks. Healthcare organizations that take advantage of KDD are able to lower the healthcare costs while improving healthcare quality by using rapid and better clinical decision making.Three popular classification techniques are used to construct the training models for comparison namely, CART decision tree and K-nearest neighbor (K-NN) and SVM. Real world healthcare data contains inconsistency and noisy data which leads to inaccurate decision and treatment. After data collection, various data preprocessing methods can be applied to clean the data. For handling the missing value, imputation method is best for the healthcare data, since every attribute plays an important role for decision making. Integrating the medical data needs a very strong knowledge based system and transformation requires mapping of data present in each format. Finally, data reduction is essential to cut the cost of data management and predict the accurate values from large scale medical data.

The feature selection techniques can be categorized into three types: Filter methods, Wrapper methods, and embedded methods. Every feature selection algorithm uses any one of these three feature selection techniques.

Filter Methods: in this method, ranking techniques are the principle criteria. The basic filter selection method are- chi-square test, Euclidean distance, correlation criteria, information gain, Mutual information, Correlation based Feature Selection (CFS) and Fast Correlation Based Feature Selection.

Wrapper Methods: these methods are better in defining optimal features than simply relevant features. This can be achieved by using heuristics of learning algorithm and training set. It uses backward elimination to remove inefficient features from the subset. Wrapper method can be broadly divided into Sequential Selection Algorithms and Heuristic Search Algorithms.

Embedded Methods: This method is incorporated into learning algorithm and optimized for it. It is also called hybrid model since it is the combination of filter and wrapper method.

A survey on intelligence approaches to feature selection in data mining:

To cope with high dimensional datasets, feature selection is utilised to reduce the number of features via casting off irrelevant and redundant features. Feature selection has been used to improve many machine learning tasks inclusive of regression, clustering and classification. However, most studies follow feature selection to classification problems.This paper presented a comprehensive survey on the works applying swarm intelligence to attain feature selection in classification, with a focus at the representation and search mechanisms.

This survey showed that there were many research attempting to not only apply SI to feature selection but also enhance the selection performance. Both parameters of the wrapped classifiers and the feature subset can be optimized. The performance can also be improved by modifying the search mechanism which is different for different algorithms due to their characteristics.

Feature Selection Methods: Supervised learning is one of the most operational fields in machine learning. It involves training a predictive model with a set of samples that includes the target outputs so that once the model is trained; it can produce the output for samples that have not been observed yet.Feature selection has been successfully used for problem solving in different fields and text classification. Efficient feature selection prior to classification has been developed in this paper. For classifiers, four popular algorithms belonging to different families (SVM, C4.5, Naive Bayes, and k-NN) are selected, which are available in Weka. For feature selection methods, six state-of-the-art algorithms: CFS, INTERACT, InfoGain, CFS, ReliefF, and SVM-RFE; have been used. The suitability of applying feature selection in two real world ophthalmology problems is also demonstrated. In one case, feature selection surpassed previous classification results; in the second case, feature selection lessened the computation time required to extract the data features that had previously prevented the real time data.

Swarm Intelligence Algorithms for Feature Selection: Feature Selection is tackled more and more with Swarm Intelligence algorithms, because Swarm Intelligence has been proved as a technique which can solve NP-hard computational problems and finding an optimal feature subset is that- kind of a problem. In recent years, SI algorithms have gained immense popularity, and, nowadays, we have by far surpassed its division into only the Particle Swarm Optimization (PSO) and Ant Colony Optimization (ASO), being the two best known SI approaches.SI algorithms are a natural choice to be used for utilizing the feature subset selection process within a wrapper model approach. Wrapper models utilize a predefined Machine Learning algorithm (i.e., a classifier) to process the quality of features and representational biases of the algorithm are avoided by the FS process.

An Improved Feature Selection Algorithm Based on Ant Colony Optimization:To improve the classification performance of the classifier, an improved feature selection algorithm, FACO is proposed. FACO is a combination of the ant colony optimization algorithm and feature selection. A fitness function has been designed, and the pheromone updating rule is utilized to efficiently reduce redundant features and prevent feature selection from falling into a local optimum. The experimental results show that the classification accuracy of the classifier can be significantly enhanced by selecting the data features using the FACO algorithm, which is of practical importance. The fitness function for the feature selection was designed to enhance the path transfer probability method of the ant colony. It was aimed at the defects in the existing algorithms.Experimental results show that the FACO algorithm can enhance the classification efficiency and accuracy of the classifiers, which is of great practical significance.

Benchmarking reliefbased feature selection methods: Modern data mining requires feature selection methods that can be implemented on large scale feature spaces , function in noisy problems, detect complex patterns of association be flexibly adapted to various problem domains and data types and are computationally manageable. To that end, this paper examines a set of filter-style feature selection algorithms (inspired by the 'Relief' algorithm, i.e. Relief-Based algorithms (RBAs)). The results of this study support the assertion that RBAs are particularly powerful, efficient and flexible feature selection methods that differentiate relevant features having univariate, multivariate, epistatic, or heterogeneous associations, confirm the efficiency of expansions for classification vs. regression, discrete vs. continuous features, missing data, multiple classes, or class imbalance, identify previously unknown limitations of specific RBAs, and suggest that while MultiSURF performs best for explicitly identifying pure 2-way interactions, MultiSURF yields the most authentic feature selection performance across a wide range of problem.The present study focuses on the family of Relief-based feature selection methods referred to here as Relief-Based Algorithms (RBAs) that can be characterized as individual evaluation filter methods.

Different machine learning techniques can be implemented to identify classes of particular applications. In this paper, machine learning algorithms are used to categorize classes of skin disease using ensemble techniques, and then a feature selection method is utilized to compare the results obtained. In this paper, a new method has been presented which applies six different data mining classification techniques to develop an ensemble approach using Bagging, AdaBoost and Gradient Boosting classifier techniques to predict classes. Furthermore, a feature importance method is utilized to choose the most salient 15 features which will play a significant role in prediction. A subset of the original dataset is obtained after selecting the 15 features, to compare the results of six machine learning techniques, and an ensemble approach is implemented to the entire dataset.The subset derived from the feature selection method is compared to the ensemble method. The outcome shows that the dermatological prediction accuracy of the test dataset is increased as compared to the

use of an individual classifier, and enhanced accuracy is obtained as compared with the feature selection subset method.As compared to individual classifier algorithms, the ensemble method and feature selection applied to dermatology datasets yields a better performance. The ensemble method provides a more accurate and effective results.

Application of Data Mining Algorithms for Feature Selection and Prediction: This research is aimed at determining the effect of using feature selection to a particular problem. The Implementation of Multi-layer Perceptron (MLP), k-Nearest Neighbour(kNN), C4.5 decision tree and Support Vector Machines was conducted on available data with and without feature selection. The algorithms were accessed in terms of accuracy and sensitivity. It is observed from the results that, using feature selection on algorithms increases the accuracy as well as the sensitivity of the algorithms considered and it is mostly reflected in the support vector machine algorithm. Making use of feature selection for classification also increases the time taken for the prediction of diabetes retinopathy.

Adaptive Human Machine Interaction Approach For Feature Selection-Extraction Task : Feature Selection task is one of the most intricate tasks in the areas of Data Mining and Human Machine Interaction. Many approaches to itssolving are based on non-mathematical and hypothesis. Newapproach to evaluate medical information quantity, based onoptimized combination of feature selection and feature extraction methods isproposed. This approach is used to produce optimal reduced number offeatures. Hybrid system of featureselection based on Neural Network-Physician interaction isinvestigated. This system is numerically simple and can produce featureselection with any number of factors in online mode using neuralnetwork-physician interaction based on Oja's neurons. The efficiency of proposed approaches in medical data mining area is confirmed by a series of experiments and this allows physicians to have the mostinsightful features without losing their linguistic interpreting. In this paper a hybrid system of feature selection-extraction based on Neural Network-Physicianinteraction is proposed. This system helps toextract the most informative features without losingthe physical sense of reduced feature space inonline-mode, and it can comprehend the HumanMachine Interaction in the area of Medical DataMining.

Detection of anomalies -Feature Selection and Data Mining Techniques: This article presents a innovative approach for fraud detection in automobile insurance claims by applying various data mining techniques. Initially, the most relevant features are chosen from the original dataset by using an evolutionary algorithm based feature selection method. A dataset is then extracted from the selected feature set and the remaining dataset is subjected to the Possiblistic Fuzzy C Means (PFCM) clustering technique for this approach. The 10-fold cross validation technique is then used on the optimized set for training and validation of WELM classifier. By taking different combinations of WELM parameters, a group of trained WELM classification models are built. The best classifier is then chosen from the trained models after being validated by the validation set as input. Finally, the test set is used on the validated model for identifying the fraudulent claims.In future, other feature selection algorithms and under sampling techniques can be applied for further enhancing the performance of the proposed system. Besides, the WELM parameters can be optimized by applying various optimization techniques for improving the classifier performance. In addition, although this work focuses on a specific application of fraud detection, the present model can be effective in fraud detection in other applications and generic databases as well.

Hybrid Harmony Search Algorithm to Solve the Feature Selection for Data Mining Applications: Recently, the increasing size of all sorts of text and data information on websites makes the method of text clustering a lot more complicated. The TC technique is implemented to cluster a gigantic variety of documents into a set of intelligible and connected clusters. Usually, TC is utilized in several domains like text mining, data processing, pattern recognition, image clustering.The harmony search rule generates irregular harmony memory which contains a collection of candidate solutions. The harmony search rule then reinforces the harmony memory to obtain the best answer to reveal the matter. This paper utilized the harmony search algorithmic rule to determine the feature choice downside and k-means for text clump problem.The modern highlight choice strategy is utilized to look for an ideal unused subset of data to make the clustering strategy successful by getting extra adjusted clusters. This technique was presented so as to enhance the execution of the content clustering procedure.

This rare strategy is called the highlight choice strategy misuse concordance look algorithmic program for the content were clustering procedure (FSHSTC), which overcomes the k-means clustering algorithmic program by enhancing the execution of the content clustering algorithmic program. FSHSTC has been surveyed using

numerous benchmark content datasets. The results were that the execution content clustering may be developed and make strides with the predicted highlight choice technique.

The followings are the data mining algorithms used to implement intrusion Detection System: Bayes Classifier, K-Nearest Neighbour, Neural Network, Support Vector Machine. Bayesian classifier is a model that encodes probabilistic relationships among data of interest. This method is commonly used for intrusion detection with statistical methods, an approach that yields many advantages with the capability of encoding interdependencies between data and of predicting methods and also the ability to combine both prior knowledge and data.

K-Nearest Neighbour algorithm is an instance type learning for classifying data objects based on closest training data examples in the feature area. This type of lazy learning is applied where the function is only approximated locally and all calculations deferred until the work classification completed. The k-nearest neighbour algorithm is a simple one when compared with all machine learning algorithms. Here the data objects are classified by a maximum of its neighbours, with the data object being allotted to the class most familiar amongst its k nearest neighbours.

The Decision tree algorithm is a predictive data modelling method most often used for classification in data mining. The Classification algorithm in data mining is learned to build a model from the preclassification of supplied data set. Each record item is declared by values of the data properties. Data Classification techniques may be viewed as mapping from a set of attributes to a particular class.The decision tree algorithm in first level built from a set of pre classified supplied given data set. The main method is to select the data attributes and divides the given data items into their classes. Based to the values of the attributes the data elements are partitioned. This process is again and again applied to each data subset of the data items.

Neural networks algorithms have been used both detection of anomaly intrusion and detection of misuse intrusion. In anomaly intrusion detection, neural networks were modelled to learn the common characteristics of system users and find statistically significant differentiation from the user's established behaviour. In misuse intrusion detection the neural network will collect data from the network data stream and analyze the data for instances of misuse type.

Support vector machines algorithms have been proposed as a novel method for intrusion detection mechanism. A support vector machines maps input vectors into a higher dimension space via some nonlinear mapping. Support vector machines algorithms are developed on the principle of structural risk minimization.

The standard mechanism which uses the pattern based intrusion detection system , the data mining technique k means algorithm to detect the intrusion process delivers the higher detection rate, the lower rate of false positives, it explicitly shows that the better the performance of the approach data mining algorithm to detect intrusions in networks is recommended.

K Means Clustering Algorithm: For getting improvement in the detection of invasion, minimize the false detection rate, here the one of the system try to incorporate the technique k-means clustering algorithm. In this method results declares the improvement of intrusion detection mechanism comparatively with the standard network intrusion detection system. Here this simulation experimented with the usage of KDD Cup 99data set. The Intrusion Detection is in continuous process in which monitoring the behaviour of a particular computer network and the different malicious behaviour diagnosis and responses in the network and information security technology. The study of IDS systems and algorithm is so important as theory and strong practical type application value.

Clustering analysis is a method to classify the data without any label tag for training and for the learning. In the clustering, same characteristics of the classified data and the clustering methods have the ability of understanding the data , then data grouped as per their similarities and complete the intrusion detection system knowledge under some conditions . The clustering analysis applied for the training data then the amount of data is far greater than anomalies data volume, and is there measured differences in the normal data and abnormal data, clustering forms different groups as per their characteristics. Then the method separates the normal activity and abnormal network activity.

In this research work, the clustering K – Means clustering algorithm is used in the effective intrusion detection system. The K – Means clustering algorithm is widely used algorithm to classify the unlabeled training data. In this algorithm, the algorithm fixed the value k initial clustering center, based on the principle of minimum

distance will be allotted to each sample k in the particular class, after class constantly and change the pattern of categories, then makes the samples to minimize the sum of the center distance. Compare with standard network and other intrusion detection systems, this approach offers a better solution to the identification of intrusion in network system but here some remarkable drawbacks are cannot determine the perfect number of clustering, in advance to clustering quality is sometimes not high. So according to these problems, the basis of this approach is improved.

In this research work, this approach proposes a clustering based on K - means high efficiency. At the first step the density for each sample data set parameters to be calculated . The algorithm is divided into different data set of isolated points and the points not participate in all kinds of sample mean included in the process of clustering. The selection of clustering center to find in the larger than the average density collection for calculating the sample set which is Larger than the average density of a subset of the m sample points the distance between the two values. The next distance in the two sample points as the initial clustering center. In the remaining sample points (m - 2), selection of the center of the front two initial clustering centers and their maximum distance product the sample points as the third initial clustering centers. In the remaining sample (m - 3) on site, have to select the three initial cluster centers to the front of the particular term maximum distance of the sample points as the fourth initial clustering centers. The value of K can be found then the initialize clustering center. Based on the principle of minimum distance to the remaining sample points is assigned to the nearest center of the cluster, until this process is repeated to complete classification of data sets. Efficient K-Means algorithm based distance and maximum distance calculation steps as follows :

Step 1: The determination of K Cluster

Step 2: Clustering validity index function

Step 3: In the improvement of this algorithm, first calculate the value distance between the given sample point and the average distance. Then density values are calculated, then outlier values will be predicted. For all the sample point values, the distance value is going to be calculated and this process will continue until all the sample points evaluated.

TABLE I. SAMPLE DISTRIBUTION TABLE

Type	Training sample	Test sample
Normal	800	700
DoS	1000	200
Probe	600	400
U2L	300	200
U2R	30	20

The approach tested with the Experimental data the KDD99 data set, the selection data set for training about large network traffic connection records each sample data contains a total of more than 40 properties which including 41 as attributes and one for decision attribute. The data set also includes 4 invasion types, such as the Probe, DoS , U2R and R2L.

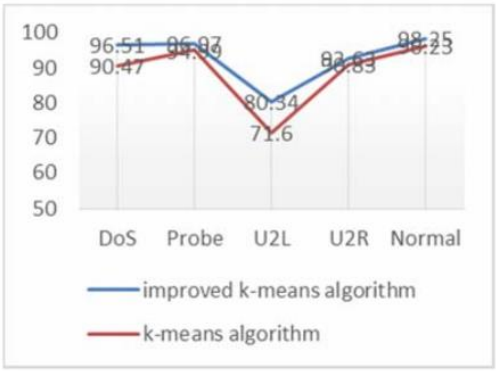


Fig 1. Detection Rate

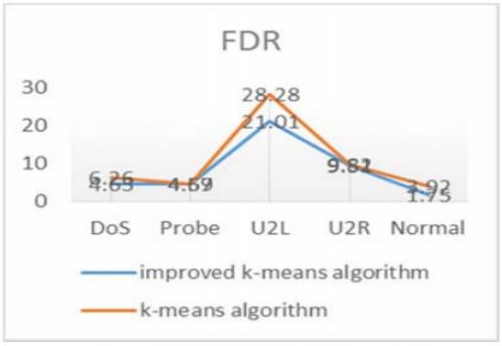


Fig 2. Error Rate

The results about this approach when compare with the previous standard intrusion detection system and some other techniques, the improved K means algorithm provides the better distribution summary and low error rate.

Anomaly detection algorithm based on statistical model: An anomaly detection algorithm usually based on statistical model which is taken to get different clusters for the process of anomaly detection. First, the characteristic factors of networks are identified. Then specific characteristics are going to be extracted such as unit time and threshold value etc. The threshold is defined by statistical method.

The particular node behaviour is coming within the range of threshold value, then we can say the node is normal; otherwise that particular node is abnormal. Feature selection is one of the key concentrated portions of Intrusion Detection System. There is no specific linear relationship between the terms such as number of the features retrieved and the performance of the detection.

The typical steps of this approach are as follows:

Step 1: Calculate the parameters K. The network is divided into k number of clusters by implementing k-means algorithm.

Step 2: Activate the IDS of cluster heads to monitor nearby nodes in each cluster. The number of packet loss and the number of packet reception are noted for each node for a time period of t.

Step 3: Find the terms the number of packet loss and the number of packet reception in unit time after T time. Calculate the confidence value interval of total based for each cluster.

Step 4: IDS system continues to monitor the network nodes in the cluster.

After t time, we find the number of data packet loss and the number of data packet reception for comparing with the confidence interval. If the data value satisfies the confidence interval, the node is treated as normal, otherwise we can say abnormal.

In this work , the simulation use NS2 as simulation platform. A region value tested 100m*100m and a WSN having 100 nodes. The wireless sensors are randomly deployed in the region. Setting time unit t is 30s and the capacity of node sample m to be 20 and readiness time unit T is set to 10 minutes as follows ;

Parameter	Value
Network size(m2)	100*100
Number of nodes	100
Data packet size(Byte)	56
Rate(kbps)	19.2
Route protocol	DSR
MAC protocol	IEEE802.11

For this implemented values, the k means clustering implemented and the value of K is determined as 8. Then for every cluster confidence interval is created by the values of data packets received and loss. Then the sensors are simulated black hole attack and selective forwarding attacks. When comparing with Bayesian classification method for intrusion detection, traffic prediction, the method which is proposed here is evaluated by based on the false alarm rate.

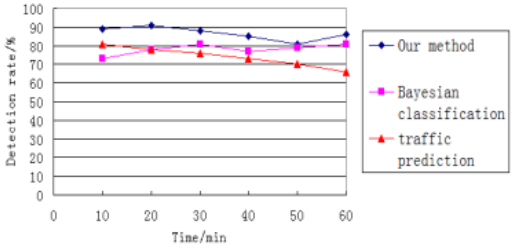
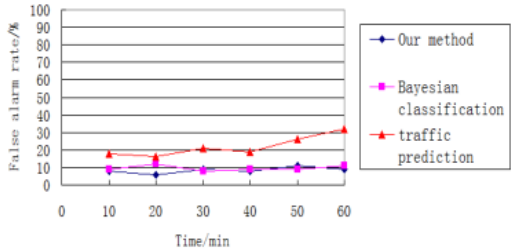


Fig.2: Detection rate



The detection rate = The ratio of (The total number of attacks - The correct number of attacks detected

The false alarm rate = The ratio between the number of a normal measurement identified as anomaly and the number of actual normal measurement. The results we can see how this method can provide higher detection rate and lower false alarm when compare the other two schemes. So that we can conclude K-means algorithm is an efficient algorithm for the node clustering of WSN. The threshold values setting based on statistical methods is universal and consumes much lower computational complexity. Finally we can say the method proposed in this paper can full fill the efficient intrusion detection in data communication system. In this paper, we explained various data mining techniques and their usefulness in the topic of Network Security and intrusion detection system. This paper also gives the description of the current Intrusion Detection Systems with data mining techniques for detecting intrusion.

REFERENCES

1. Jinying Wang1, “Analysis on the Application of Campus Network Firewall And Intrusion Detection System”, International Conference on Intelligent Systems Research and Mechatronics Engineering (ISRME 2015)
2. EvgeniyaNikolova, “A Clustering-Based Unsupervised Approach to Anomaly Intrusion Detection”, 2nd International Symposium on Computer, Communication, Control and Automation (3CA 2013)
3. Zhidong Shen, “A Bayesian Classification Intrusion Detection Method Based on the Fusion of PCA and LDA”, Hindawi Security and Communication Networks Volume 2019, Article ID 6346708
4. Xu Tao1, “A Novel Intrusion Detection System Based on Data Mining”, 4th International Conference on Computer, Mechatronics, Control and Electronic Engineering (ICCMCEE 2015)
5. Xiaomeng Li, “Discussion and Development of Network Attack and Prevention”, Advances in Social Science, Education and Humanities Research, volume 86 International Conference on Economics and Management, Education, Humanities and Social Sciences (EMEHSS 2017)
6. Uma R. Salunkhe1, “Security Enrichment in Intrusion Detection System Using Classifier Ensemble”, Hindawi Journal of Electrical and Computer Engineering Volume 2017, Article ID 1794849
7. Hui Liu, “Design of the Computer Intrusion Detection System”, 3rd International Conference on Management, Education, Information and Control (MEICI 2015)
8. LIU Chun1 A, “Research and Simulation of Network Intrusion Detection Algorithm Based on Fuzzy Classification”, 2014 International Conference on Computer Science and Electronic Technology (ICCSET 2014)
9. I-Hsien Liu1, “IDS Malicious Flow Classification”, Journal of Robotics, Networking and Artificial Life Vol. 7(2)
10. Li Wang1, “Network Intrusion Detection Using Support Vector Machine Based on Particle Swarm Optimization”, International Conference on Applied Science and Engineering Innovation (ASEI 2015)

11. TomášBajtoš, "Network Intrusion Detection with Threat Agent Profiling", Hindawi Security and Communication Networks Volume 2018, Article ID 3614093
12. Jianfeng Pu, "A Detection Method of Network Intrusion Based on SVM and Ant Colony Algorithm", National Conference on Information Technology and Computer Science (CITCS 2012)
13. Guanglei Qi, "Construction and Application of Machine Learning Model in Network Intrusion Detection", Atlantis Highlights in Engineering, volume 5 International Conference on Precision Machining, Non-Traditional Machining and Intelligent Manufacturing (PNTIM 2019)
14. Bo Sun1 A, "Research on Computer Network Intrusion Detection System", Advances in Engineering Research, volume 150 4th International Conference on Machinery, Materials and Computer (MACMC 2017)