Ramping up Data mining algorithms for Intrusion Detection

M.Deepa
Ph.D Research Scholar, Department of Computer Science,
Vivekananda College of Arts and Sciences for Women (Autonomous),
Elayampalayam

Dr.P. Sumitra
Professor, Department of Computer Science,
Vivekananda College of Arts and Sciences for Women (Autonomous),
Elayampalayam

Abstract

At present the importance of ensuring that computer systems are safe from attacks because of modern society 's reliance on those systems becomes increasingly relevant. The intruders' identification with a new data mining technique offers a higher detection rate compared to other conventional systems. The developments in data mining have gained a considerable traction in recent years due to the industry 's international use of this technology. The aspect of data mining in numerous areas, especially in fraud detection, cyber security and biology classification, has shown impressive results. Several data mining algorithms are specifically appropriate for the detection of intruders. There have been several attempts to attain methods for upgrading of existing data mining algorithms. This paper looks at the methods used to tackle the issue of optimization. Instead of concrete implementations we concentrate on general ideas that can be used to provide a general view of existing approaches to upgrading data extraction methods. A methodology is proposed to the algorithms and several examples are given of various tasks.

Keywords: Intrusion Detection System, Data mining, supervised learning, unsupervised learning.

1. Introduction

In order to massive presence and the rapid growth of the internet, networks for corporations, social media and governments have been gradually created. A change in the behavior of the natural network that is clearly visible is known as an interference and assault in the field of computer networks. In fact, the majority of irregularities and deviations can be observed by analyzing and analyzing the moving network traffic[1]. So people want to keep their belongings secure. In modern times, a protection device in a house is extremely common. We have all implemented security mechanisms on our machines now that technology has evolved. For a while now, different approaches have been used as frewalls and antivirus applications to protect the privacy of both users and confidential data.[2][3]. An Intrusion Detection System is one of that kind protection system. A great deal of attention has been directed to the evaluation of intrusion detection systems [1] (IDSs) in the field of information technology. All such algorithms operate cyber-attack detection, using a number of techniques to detect security systems failures and malicious behaviour. An ID generally follows two methods: (a) a signature-based approach or (b) the anomaly-based approach. Signature-based detection requires prior awareness of an attack before it can be detected, however techniques based on the detection of anomalies by obtaining information about patterns that represent "standard" or "attack" data, and then identify new data in keeping with their likeness to those patterns. In a number of fields which include marketing, manufacturing, data processing, fraud prevention or network administration, data mining techniques have indeed been successfully applied. Data mining has been used in that several problems in intruding detection by an increasing number of research projects over the last five years. Presently, intrusion detection technology implementation of data mining has become a hotspot.[4]. This report discusses the conclusions of a document analysis on the use of intrusion detection technologies in the cyber security industry. This study also explains and contrasts Data mining approaches, which have been suggested in the updated literature for the defense of new reflected IDS solutions. The following document is structured. Section 2 outlines the literature analysis of the study submitted. The theoretical outline of IDS and related Data mining concepts is given in the third chapter. Section 4 shows the research methodology , Section 5 contains the literature review findings with an overview of surveys, new system recommendations and other articles focused on DM and IDS focus categorization. The last section of the paper introduces our inference.

2. Literature review

We have opted to concentrate primarily on modern IDS Data mining appliances. Drawing from the analysis carried out, most surveys span a broad scope such as mechanical learning and/or data mining. In accordance with the adaptive data sampling technique, the method was used to speed up the learning process using the RMS prop method[5] . Their system, called Rand Net, which stands for randomized neural network for the identification of outer structures, demonstrated robustness, which prevents overwhelming problems.

It is regarded as a one-class classification as to determine whether a new instance is a class of data used for training the classifier or whether it is an outlier. This means that in the training process the classifier can learn only the data patterns of a class (target class). Other names like news or outliers and idea learning [6] are called for in this area.

In order to reach global optimality, several hybrid optimization algorithms are proposed. In order to slash energy demand and treat a wide range on the network, Barekatain et al .[7] suggested a new mix of k-median and enhanced GA. The proposed approach is a hybrid model from PSO (Simulated Annealing Article Swarm Optimize), SA (Simulated Annealing), SS (Scatter Search), k-means and some other heuristics. A hybrid evolutionary k-means model is being suggested in the Karimov-and Ozbayoglu [8]. The ultimately increases data clustering algorithm K-MCI, which combine K-mean with modified cohort knowledge, has been proposed by Krishnasamy et coll.[9].

A deeper learning methodology was introduced in IDS in Alrawashdeh and Purdy[10] for the identification of irregularities using a restricted Boltzmann system (RBM). Their approach requires an unregulated RBM layer in order to accomplish uncontrolled reduction of functionality. The resulting weights of this RBM were moved to a separate RBM, forming a strong network of beliefs. They reach a 97.9 percent detection score.

Dada[11] indicated that multiple classifications were hybridized to boost IDS precision. The results revealed that the combination of SVM , kNN AND Primal – Dual PSO provided a better rating accuracy than any single classifer in the KDD99 dataset. One of the most recent projects on IDS was the one suggested by AlYaseen et al .[12]. This was a mix of SVM and strengthened algorithm K-means. In comparison, the Bayes network combined with the Wrapper function reduction algorithm in the IDS by Onik and Samad (2017).

3. Intrusion detection & data mining algorithms

3.1 Intrusion Detection System

In reality, protection mechanisms are configured to track, recognize and respond to malicious threats, a data device, a network or information systems generally. These attacks are intended to compromise the completeness of these systems and loot knowledge, thus compromising the systems in certain instances. IDS is either an intrusion prevention system or hardware, tracks network traffic for unusual activity and transmits alerts to an administrator[13]. The Figure 1 shows the working framework for intrusion detection system. Day to day network data is taken by different sensors and it will be stored in general data warehouse. The detection engine get through all network packets and find out whether is it normal or abnormal by using either supervised or unsupervised data mining algorithms. If the detection engine find out any abnormal packets then it initiate the alarm manager and make the alarm.

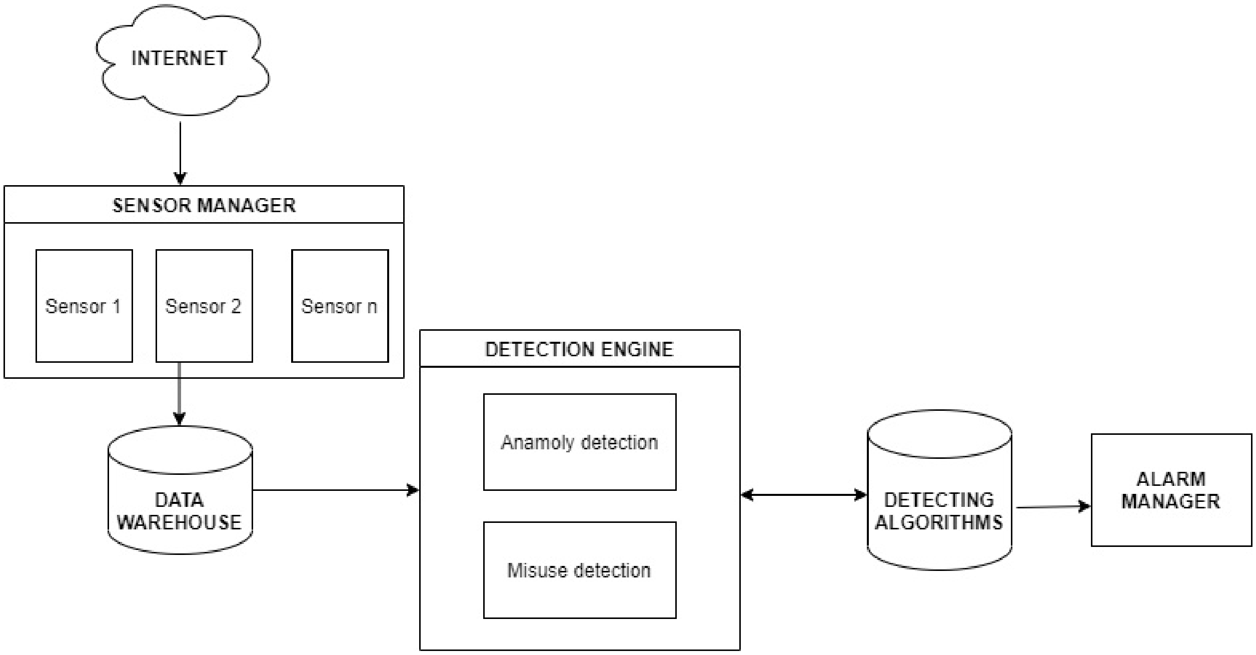

Figure 1: Framework for intrusion detection system

In compliance with the following criteria[14], IDPS may be classified: 1. Intruder type: This is both external and internal. Intruder type: An external attacker is one with no network or service access, while an internal extruder is one with allowed network access and minimal network permissions. 2. Intruding Type: In Chapter Two, there are different types of intrusions. 3. Detection technique: Three types of intrusion detection methods, mistreatment detection, detection of anomalies and specified Protocol analysis are usually used.[4]. The key difference in the design choices for IDSs, which is shown in the Figure 2, vary depending on several factors like based on detection approach the IDS can be either signature based, Anamaly based or hybrid IDS. From the source of data the IDS would be host based, network based or hybrid data. In the structure of IDS can be classified as centralized and distributed architecture. The IDS can be either real time or off line system based on time aspects.

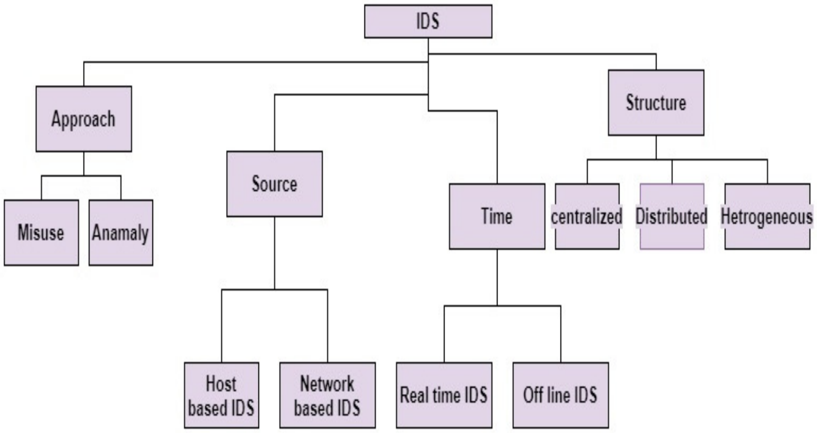

Figure 2: Classification of intrusion detection system with different aspects

Table 1 offers a brief description of the methods of intrusion detection, including anomaly based, stateful protocol, rule based, supervised machine learning and unsupervised learning methods their benefits and disadvantages, along with examples.

Table 1: Categorization of Various intrusion detection systems

Type	What is it?	Pros	Cons	Example
Anomaly-Based ID	This is a behavior-based testing method that gets its feedback from operating system-generated audit logs. This method of strategy searches for behavioural differences that could suggest masquerading.	i)Ability to recognize and understand and decrease the false alarm rate of unidentified attacks ii)Uses the collected actions statistical test to assess intrusion	i)The accuracy of identification is dependent on the amount of activity or features obtained. ii)Less successful due to continuous changes in tracked activities in the complex world	i)Level of processor usage for a host during a given period of time ii)Average number of emails sent by a profiled user Routing traffic levels
Stateful Protocol Analysis	Stateful review of protocols recognizes protocol modification. This adopts preset uniform profiles generated based on agreed meanings of protocol activity created by suppliers and market leaders unlike the anomaly detection process.	i)Adds stateful attributes to routine study of protocols ii)Recognizes abnormal command sequences	i)Work consuming for the tracing and study of the protocol state ii)Attacks that do not breach the features of commonly agreed protocol conduct can not be observed.	Monitoring requests with its corresponding response
Rule-based detection	This involves making decisions based on rule sets which are defined by domain experts. They can detect known attacks but are incapable of detecting novel attacks. Also, with increase in network traffic, finding and coding rule sets is both difficult and time-consuming	Can easily detect known attacks	Unable to detect unknown attacks Finding and coding rule sets is both difficult and time wasting	Detecting flood type attacks,SEM.DAP

Supervised Machine Learning (ML)	It needs no model design, just like in the case of anomaly detection. Instead, it can learn dynamic and malicious models.	i) Capable of analyzing different and harmful models	i)They are not used in a controlled situation because they need adequate supply of marked naming data	i)Linear regression, ii)Random Forest, iii)SVM
Unsupervised Machine Learning	This intrusion detection approach includes unlabeled data construction models	i)Unlabeled data may not be needed on the domain specialist	i)Testimony is a legal is not as strong as ML controlled	i)K-Means ii)Apriori

3.2 Data Mining

An algorithm of data mining is a series of heuristic and calculative data mining models based on data.[15] Determining the correct or designed specifically algorithm to apply to solve some certain issue can be a challenge. Although new approaches can be used for the same tasks, each algorithm produces different results, and some algorithms may deliver even more than one type of performance. Some algorithms can assign one or more discrete variables depending on the other attributes of the data collection. The data is predictable. Some algorithm works regression functions and can forecast more or more constant variables on the basis of other data set attributes. Some algorithms may action for implementing, divide data into groups or clusters of things with similar properties, as Microsoft has pointed out [15]. While some algorithms can be associative by finding similarities between different attributes in a collection, some can be used for sequence analysis processes, which can be used to summaries sequences or episodes in data, such as a web path float. All of the algorithms mentioned above can therefore be divided into two broad categories: supervised learning and unregulated research algorithms. The following sections address briefly the two categories: regulated and unmonitored instruction. The most popular data mining methods used to design IDSs are illustrated in Figure 3. A description, capabilities and weakness of some of the data mining algorithms can be found in the following sub sections.

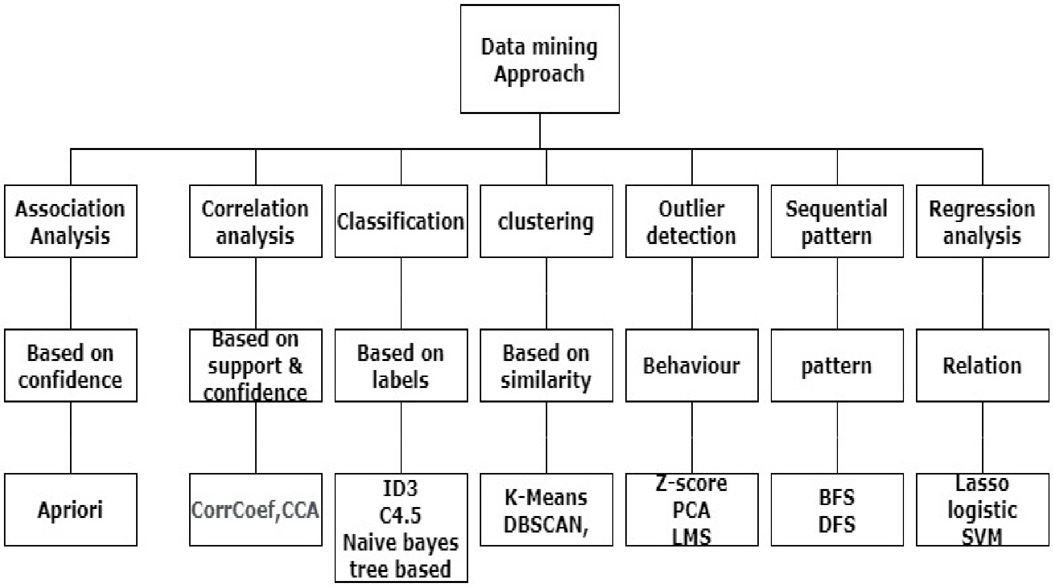

Figure 3: Various data mining approach for building IDSs

The supervised learning algorithms are those for which before the algorithm is run, the class attributes for the dataset are known. These data are referred to as data on marking or instruction [16]. This set is made up of tuples (x , y) where x is a vector and y is the class, always a scalar attribute. Learning supervised produces an x to y mapping model. The challenge is to find a mapping of m.) (to m(x) = y. It is also presented with the unlabeling or test data set where instances are unknown in the form(x,?) and y values. In view of the learning of m.) (and of x from an unlabelled instance, the prediction of a mark for an unlabeled case [5] can be calculated by m(x).

3.2.1 Decision Tree Learning(C4.5,ID3,CART)

In decision tree learning method each unleaf node in a tree constitutes a component and each branch represents a value the feature can take. Instances are categorised by following the path that begins in the root node and terminates on a leaf by following branches based on instance feature values.[17]. The construction of decision trees is heuristically oriented.

Advantage

Multiple decisions trees can be taken out of the same data collection. They can both accurately estimate the class attributes on all instances on the dataset. They are implicitly screening variables or selector of functionality. It takes comparatively limited effort from users in data planning.

Limitations

Decision trees are recurrently built on training data using the greedy upstream model, with sequential features chosen.[17] They appear to over-sticks training data, rendering them marginally weak predictors without adequately pouncing or restricting tree growth. Low performance because you need to 'rewrite' the tree each time you want the CART model to be modified and poor data resolution with complicated variable relations. There are only two choices for each node (left-right), therefore, variable relationships cannot be learned in decision-making. Practically classification limited.

3.2.2 Naive Bayes Classifiers (NB)

The Bayesian network is a framework that computes probabilistic relations between interest variables[18].Based on the implementation of Baye 's theorem with strict (native) autonomy assumptions between the characteristics, naive classification devices are a family of simple probabilistic classifications. The Bayesian intruder technique in tandem with statistical schemes, is commonly used, offering multiple benefits, such as the ability to encode interdependencies between variables and to forecast events, and the ability to combine both prior information and data.

Advantage

Easy for using. Efficient if the set is sufficiently large. Learning capacity; the findings became increasingly reliable with the increasing training kit (intelligence)[17]. Naive Bayes can also surpass more advanced classification approaches, considering its simplicity. When NB is simply believed to have conditional freedom, a Naive Bayes classification converges faster than discriminatory models such as logistic regression, because you would require fewer training results. Naïve Bayesian classification scheme simplifies the equations and displays high precision and speed when applying to massive datasets. Bayesian classifiers produce good outcomes as the emphasis is not on exact probabilities, but on the detection of instance groups[18].

Limitations

Does not take into account the word sequence (Non-relevant word feature. It cannot research practical interaction. The Bayesian theorem is based on this approach and especially suitable if the dimension of the inputs is high. computational stress is considerably higher[18] The solution to Naïve Bayes is to still be predicted

3.2.3 K-Nearest Neighbor (KNN)

There are no criteria needed by KNN for its function. To measure the distance between neighbors [19], Euclidean distance is used. The basic concept behind the KNN classification algorithm used to classify a new data instance into classes that have already been observed is shown in Figure 11, based on its relative distance to either class. The green squares represent the usual type of behavior and the red triangles display the abnormal class of behaviour, every newly discovered unknown instance (blue hexagon) can now be categorized on the basis of the number of closest maximum neighbours from either class. This new instance is, therefore, graded as a recognized class. The number of nearest neighbours that are used for classification is k.

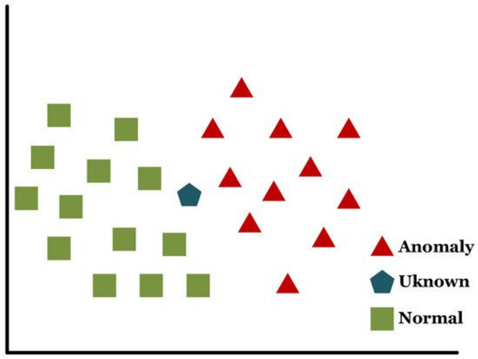

Figure 4: Principle classification of K-Nearest Neighbor[20]

Advantage

It is very simple to understand the K-NN algorithm and equally easy to implement. K-NN is an algorithm that is non-parametrical. K-NN does not specifically construct any model, it merely labels the learning from historical data based on the new data entry. A memory based solution is k-NN. K-NN can be used both for problems with classification and regression.

Limitations

K-NN can also be very simple to implement, but the efficiency or speed of the algorithm declines very rapidly as the dataset increases. KNN works well with a limited number of input variables, but the K-NN algorithm fails to predict the output of the new sample as the amount of independent variables. With K-NN, one of the main problems is selecting the optimum number of neighbours to be considered when classifying the new data entry. On imbalanced data, k-NN doesn't perform well. The K-NN algorithm is very sensitive to outliers because, based on distance parameters, it simply selects the neighbours. Inherently, K-NN has no capacity to deal with the missing value problem.

3.2.4 k-Means Clustering

Which is an unsupervised algorithm focused on the discovery of k clusters in sample data. Based on its characteristics, every instance of sample data is allocated to a specific cluster. Using the estimation of centroids as per squared Euclidean distance, the samples are distributed over k clusters according to their features. The method proceeds iteratively until no improvements can be made to the clusters[21][22]. The selection of an acceptable k value and the assumption that the sample dataset is distributed equally over the k clusters serve as limitations for the algorithm of clustering k-means.

Advantage

K-means are straightforward to fix and unknown classes of data across complex data sets are defined. The implementation of K-means can easily adapt to the modifications. For a significant larger dataset, K-means is ideal because it is calculated much quicker than a particular dataset. Its output relies on the shape of the clusters. In hyper-spheric clusters, K-means function well.

Limitations

K-means do not permit an indicated that the best of clusters to be formed, and you should decide on the clusters beforehand for effective results. Even when the input signal has different models, it generates a uniform cluster. The final results would be completely altered by modifying or rescaling the dataset by either normalisation or standardisation.[23]

3.2.5Ensemble Learning (EL)

As shown in Figure 5, EL operates by drawing on the strengths of different classifiers through a combination of their outcomes and then producing a simple majority for classification. This increases the precision of classification by a combination of the outputs of different homogeneous classifiers[24]. EL is focused on the study[140], where it would be found that for its accuracy, every ML classification algorithm relies on the application and associated data. Therefore, no ML problem can be formulated as "one size fits all solutions" and EL like combinations can be better suited for generalised applications to optimise accuracy by reducing variance and avoiding overfitting.

4.Datasets Available for Network Intrusion

4.1 KDD cup 99

This is an improvement in the DARPA DARPA98 dataset, introduced by an IDS programme undertaken by the Lincoln Laboratory (Massachusetts Institute of Technology) of MIT to determine IDSs that discriminate between usual incoming and attack conations. After some screening in the IKD CUP 99 dataset [25], this dataset was submitted in the International Data Mining Tools Competition (Stolfo, S.J.; Fan, W.; Lee, W.; Prodromidis, A.; Chan, P.K. In Proceedings of the DARPA Information Survivability Conference and Exposition, January 2000). For the last two decades, several scholars have used this dataset. The lack of alternatives has led to the KDD CUP 99 dataset[25] having been established as a comprehensive benchmark for classifier precision. However, KDD-99 has several drawbacks, including ageing, distorted goals, and lack of stationary relationship between training and testing results, pattern redundancy and irrelevant functionality.

4.2 NSL- KDD

NSL-KDD is an attempt to resolve the shortcomings of KDD-99 by the researchers who published their work in[25]. It is a more balanced resampling of KDD-99 where the focus is placed on instances that classifiers trained on the basic KDD-99 are anticipated to miss. Nevertheless, as everyone authors themselves admit, there are still shortcomings in the dataset, such as its lack of representation of low footprint attacks.[26].

4.3 DDoS - 2016

Data collected in a managed environment (using Network Simulator NS2) with four malicious network attack forms are presented in the dataset: HTTP Flood, UDP flood, DDOS Using SQL injection (SIDDOS), and Smurf. There are 27 attributes, 5 classes (4 attack classes and one regular traffic class) and 734,627 records in the dataset.[27]

4.3 CICIDS 2017

The Canadian Center for Cyber security has made the dataset public. Two kinds of usage profiles and multi-stage attacks, such as Heart bleed, and a number of DoS and DDoS attacks were used in the development methodology. Using the CIC Flow Meter method, it has 80 network traffic properties which are extracted. The abstract human behaviour of 25 users working with HTTP , HTTPS, FTP, SSH, and email protocols was focused on user profiles, with the goal of generating background traffic. The traffic was generated over a limited period of time (5 days).

4.4 CSE-CIC-IDS 2018

Six types of network attacks are protected by the dataset: Botnet, brute-force, Denial of Service (DoS), Distributed DoS (DDoS), infiltration and web attacks. Depending on virtual web browsers, which obtain abstract representations of network events and behaviours, the dataset was developed. Fifty network nodes with 420 computers and 30 servers were used to coordinate an attack on the victim's infrastructure. Using the CICFlowMeter-V3 tool, the dataset contains 84 network traffic features extracted from network traffic.[28]

4.5 LITNET-2020

A new annotated network benchmark dataset acquired from the real-world academic network, LITNET-2020. The dataset offers real-world examples of network traffic that is common and under attack. We define and evaluate the dataset's 85 network flow features and 12 forms of attack. By using statistical analysis and clustering techniques, we present the analysis of the dataset features. Our findings show that it is possible to efficiently use the proposed feature set to classify various attack groups in the dataset. For research purposes, the network dataset presented is made publicly accessible.[29]

5. Research methodology

The bulk of testing techniques are split in the phases of data processing and data analysis. Figure 5 summarizes a comprehensive schematic of the method. The method of data preparation involves preprocessing and partitioning of training and test data. Data review requires preparing the proposed model and the success evaluation by means of test data for the proposed model. Data pre-processing on any network attack data collection such as KDD cup 99, NSL-KDD data set, DARPA dataset etc. was first carried out. 70 percent of the data collection was used for preparation and the remaining 30 percent for research in the data partitioning process. The training data were used to train the various supervised or unsupervised learning models during the training process. In the test phase, both normal and irregular input data were reconstructed and there was a comparison of the difference between the inputs and the reconstructed data.

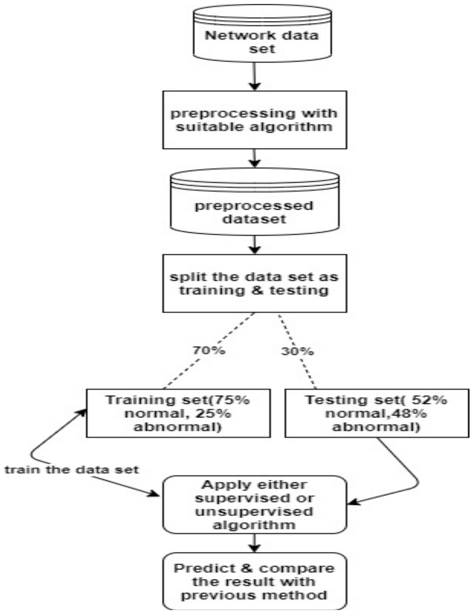
Network data set
preprocessing with suitable algorithm
preprocessed dataset
split the data set as training & testing
70%
30%
Training set(75% normal, 25% abnormal)
Testing set(52% normal,48% abnormal)
train the data set
Apply either supervised or unsupervised algorithm
Predict & compare the result with previous method

Figure 5: Research methodology for various research proposals

6. Results and Discussion

Most of the papers reviewed describe novel approaches for IDSs, which are based on a data mining or experiments. The table below is a description of the researchers' Data mining architectures. The dataset used for validation of methods is also defined. The suggested solutions have not been adequately comparable. Owing to the use of multiple datasets or data subsets, such similarities may not be informative. In addition, during data processing processes or the kind of attacks observed by particle wide IDS, there are several variations. Only sections of the following algorithms are listed in this article because of the wide range of available methods. Table 2 below covers the comparison of work conducted on Data mining techniques.

Table 2 comparison of recent works on data mining

Authors	Data set	Method	Attacks	Methodology	Year
(Singh & Singh, 2014)	NSL-KDD Kyoto 2006+	Online sequential extreme learning machine (OS-ELM)	Dos u2r r2l probe	Computational time and memorandum specifications approach minimized. Multiple topological parameters of the proposed architecture of the data mining are based on neuron count modulation in the hidden layer. Comparison of the effects of the suggested	2014

				technique: ANN, AdaBoost, Natural Bayes and ELM	
Kim et al. (2016)	KDD Cup 1999	LSTM +RNN	Dos u2r r2l probe	ML System with KDD Cup 99 subset training results. This essay comprises professional mentorship on the data mining parameter eters to enhance the approach suggested. Comparison results: GRNN, PNN, RBNN, KNN, SVN, Bayesian	2016
Pandeeswari and Kumar (2016)	KDD Cup 1999	Hybrid: Fuzzy means clustering (FMC) - clustering of incoming data NN - trained based on FMC output	Dos u2r r2l probe	Proposed hybrid cloud-based method. Results vs.: Naïve Bayes and ANN	2016
Wang, et al. (2017)	Self-created USTC-TFC2016	CNN	Gmail Cridex	Method of imagery taking traffic information from the network-CNN training input. The use of raw data trafc. This document contains a new dataset of network trafc, generated in accordance with USTC-TK2016 data preprocessing package, developed by the author, USTCTFC2016 (around 3.71 Gb).	2017
Du et al. (2017)	HDFS logs OpenStack logs VAST	LSTM	DOS Port scan	System logs focused as training data process. system logs. Check performed even on the 2011 dataset VAST challenge	2017
Yin et al. (2017)	NSL-KDD	Recurrent NN	Dos u2r r2l probe	Form of research of various recurrent neural network topologies. Comparison descriptions with other ML approaches such as: J48, Naïve Bayes, SVM, NB Tree Run Woodland, MLP.	2017
Ashfaq et al. (2017)	NSL-KDD	Fuzziness based algorithm	Dos u2r r2l probe	Method of semi-controlled ML. Results compared to: J48, Naive Bayes, NB tree, Random Forests, Random tree, SVM. Classification of two classes: standard versus attack.	2017
Shone et al	KDD Cup 99 NSL-KDD	Non-symmetric deep AE Random forest	Dos u2r r2l probe	The AE and Random Forest Stacked Process. A detailed comparison of DBN findings by each KDD data set hazard.	2018

7. Conclusion and Future Enhancement

The paper presents an overview of the literature review carried out for intrusion detection systems to current data mining use. We wanted to do so because cybersecurity is an emerging topic of study and the continuing advancement in the field of data studies is a reality. For the development of new models data mining algorithms are commonly used. One of the main problems in intrusion detection systems is the development of a reactive framework for any new and low-frequency attacks. Public sources currently available are not an appropriate source for such a case of usage. One of the interesting methods is to concentrate on unique forms of attacks and to plan responses for them explicitly, as seen in many papers reviewed. This may improve the adaptability of strategies to new forms of threats. Furthermore, the massive volume of data processed in the world every day should be discussed. IDSs to be generated in the future must be immune to data volume problems

References

[1] Hoque, B. B. (2014). Network attacks: taxonomy, tools and systems. *J Netw Comput Appl* , 307–324.

[2] Ng, J., Joshi, D., Shankar, & Banik. (2015). Applying Data Mining Techniques to Intrusion Detection. *12th International Conference on Information Technology - New Generations,*.

[3] K-KR, C. (2011). The cyber threat landscape: challenges and future research directions. *Comput Secuity* , 719–731.

[4] Singh, & Singh, M. (2014). Analysis of host-based and network-based intrusion detection system. *Comput. Netw. Inf. Secur* , 41–47 .

[5] Tieleman T, H. G. (2012). Rmsprop: divide the gradient by a running average of its recent magnitude. *COURSERA Neural Netw Mach Learn* , 26–31.

[6] Khan, & Madden. (2014). One-class classifcation: taxonomy of study and review of techniques. *Knowl Eng Rev* , 345–374.

[7] Barekatain, Dehghani, & Pourzaferani. (2015). An energy-aware routing protocol for wireless sensor networks based on new combination of genetic algorithm & k-means. *Procedia Comput. Sci* , 552–560.

[8] Karimov, & Ozbayoglu. (2015). Clustering quality improvement of k-means using a hybrid evolutionary model. *Procedia Comput. Sci.* , 38–45.

[9] Krishnasamy, Kulkarni, & Paramesran. (2014). A hybrid approach for data clustering based on modified cohort intelligence and K-means. *Expert Syst. App* , 6009–6016.

[10] Alrawashdeh, & Purdy. (2016). Toward an online anomaly intrusion detection system based on deep learning. *15th IEEE international conference on machine learning and applications (ICMLA)* (pp. 195–200.). USA: IEEE.

[11] Dada. (2017). A hybridized SVM-kNN-pdAPSO approach to intrusion detection system. *Fac Semin Ser Univ Maid* , 48–54.

[12] Al-Yaseen, W. L., Othman, Z. A., & Nazri, M. Z. (2017). Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. *Expert Systems with Applications* , 67,296-303.

[13] Kemmerer, & Vigna. (2002). Intrusion detection: a brief history and overview. *IEEE Computer Society* , 27–29 .

[14] Santos, Chandra, Phani, Ratnakar, Baba, & Sudhakar. (2013). Intrusion detection system- types and prevention. *International journal of computer science and information technology* , 77-82.

[15] Asanka, D. (2019). *introduction to SQL Server Data Mining.* sql shack.

[16] Liao, Lin, Lin, & Tung. (2013). Intrusion detection system: a comprehensive review. *network and computer applkications* , 16-24.

[17] Zafarani, Abbasi, & Liu. (2014). *Social Media Mining*. Cambridge University Press.

[18] Swarnkar, & Hubballi. (2016). OCPAD: One class Naive Bayes classifier for payload based anomaly detection. *Expert systems applications* , 330-339.

[19] Soucy, & Mineau. (nov-dec 2001). A simple KNN algorithm for text categorization. *In Proceedings of the 2001 IEEE International Conference on Data Mining*, (p. 29). USA.

[20] essam, hasnet, & all, e. (2020). A Review of Intrusion Detection Systems Using Machine and Deep Learning in Internet of Things: Challenges, Solutions and Future Directions. *electronics* , 177.

[21] Bhuyan, Bhattacharyya, & Kalita. (2013). Network anomaly detection: methods, systems and tools. *Ieee communications surveys & tutorials* , 303-336.

[22] Kanjanawattana. (2019). A Novel Outlier Detection Applied to an Adaptive K-Means. *International Journal of Machine Learning and Computing* , 9(5).

[23] Vinnikov, & Shalev. (2014). K-means Recovers ICA Filters when Independent Components are Sparse. *International Conference on Machine*, (p. vol 32). china.

[24] Woźniak, M.; Graña, M.; Corchado, E. A survey of multiple classifier systems as hybrid systems. Inf. Fusion 2014, 16, 3–17. (2014). *inf.fudion* , 3-17.

[25] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. (July 2009). A detailed analysis of the KDD CUP 99 data set. *In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications* (pp. 1–6). Canada: IEEE.

[26] McHugh, J. (2000). Testing intrusion detection systems: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. 262–294.

[27] Rajagopal, Kundapur, & Hareesha. (2020). A stacking ensemble for network intrusion detection using heterogeneous datasets. *Security and Communication Networks* .

[28] https://www.unb.ca/cic/datasets/ids-2018.html, U. C.-C.-I. (2020).

[29] robertas, & etc. (2020). LITNET-2020: An Annotated Real-World Network Flow Dataset for Network Intrusion Detection. *Electronics, Machine Learning Techniques for Intelligent Intrusion Detection Systems* , 9(5)800.

[30] Stolfo, S.J.; Fan, W.; Lee, W.; Prodromidis, A.; Chan, P.K. In Proceedings of the DARPA Information Survivability Conference and Exposition. (January 2000). *Cost-based modeling for fraud and intrusion detection: Results from the JAM project.*, (pp. Volume 2, pp. 130–144.).

[31] Gupt, M., & Shrivastava. (Feb 2015). Intrusion Detection System based on SVM and Bee Colony. *International Journal of Computer Applications* , 111,10.

[32] Meira2, J., & Andrade, R. (n.d.). Performance evaluation of unsupervised techniques in cyber attack anomaly detection . *Journal of Ambient Intelligence and Humanized Computing* .

[33] Sharafaldin, Lashkari, H., & Ghorbani. (2018). A Detailed Analysis of the CICIDS2017 Data Set. *ICISSP* (pp. 172-188). switzerland: springer.