Volume-4,Issue-6, Dec - 2015
ISSN (O) :- 2349-3585

A Survey on Data Mining Classification for Intrusion Detection System

Ravi N. Jethva[1], Mr. Kaushal Madhu[2]
[1]P.G. Student, [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]L.J. Institute of Engineering and Technology, Ahmedabad, Gujarat, India.

Abstract – Intrusion Detection System using Data Mining algorithms is a wide scope of Research. Wherein, various classification techniques can be used for a better classification of Known and Unknown type of attacks. An IDS (Intrusion Detection System) monitors the network traffic and then sends the suspicious activity reports to the System Administrator. There are two types of Detection, Misuse Detection and Anomaly Detection. This paper presents and compares different techniques in terms of efficiency such as Fuzzy Logic, Neural Network, Rule based classification, Hoeffding Tree Classification.

Keyword – Intrusion Detection System, Hoeffding Tree Classifcation, Anomaly Detection.

I. Introduction

Computer Systems and Networks now a day are highly susceptive to Confidential and Sensitive data which are being received from and sent out to different networks. An IDS has been developed for such kind of network attacks which can access or alter confidential data or information. An IDS is a software application that monitors the underlying network for malicious activities. IDS is mainly focused on identifying possible incidents and to report them or to store them in the directory in order to prevent the same kind of incidents in future. Prevention techniques alone are not sufficient since it is impossible to have an absolute secure system, hence IDS is designed for identifying the attacks and to classify them for Historical Analysis purpose. New intrusions continually emerge and new techniques are needed to defend against those intrusions. IDS is the second line of defence, since it comes into the picture after an occurrence of an intrusion.[1] The main function of Data Mining techniques in Intrusion Detection is to classify the attacks as Normal or Malicious. Data Mining also provides Data Summarization and Visualization which provides efficient Historical Analysis.

II. Literature review

1. Data Mining based IDS Architecture

The IDS architecture consists of Sensors, Detectors, Data Warehouse, and Model Generator. IDS architecture is capable of gathering data and sharing of data. Also, it is designed for Data Analysis, Data Archiving, Model Generation and Distribution. Designated system is independent of Sensor data format and model representation.

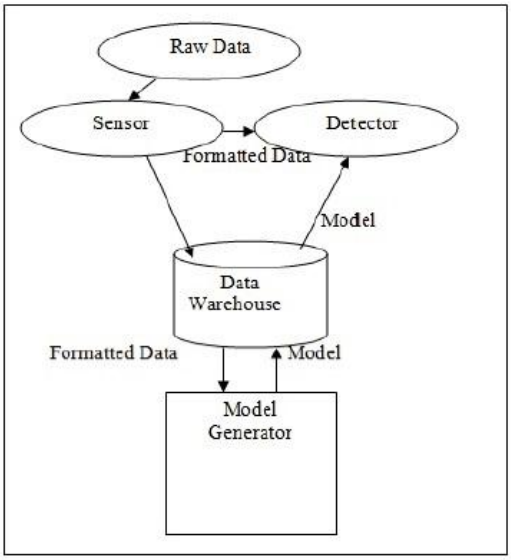

Fig 1. Data Mining based IDS Architecture [1]

1.1 Sensors

Sensor's task is to observe raw data on a monitored system. This raw data can be useful for Model Evaluation.

1.2 Detectors

Detector plays a role to take the processed data from sensors as an input and to process it in such a way that finds out if it is an attack. [1] It sends the results to Data Warehouse for reporting and analysis. There can be several detectors monitoring the same network system. It can be front-end or back-end detectors whose task are Simple intrusion detection and Trend analysis respectively.

1.3 Data Warehouse

Data Warehouse is a centralized storage for data and models and also facilitates integration from multiple sensors. IT correlates data/results from different IDS over a long period of time and hence the detection of large scale attacks become possible.

1.4 Model Generator

Main task of Model Generator is to provide Rapid Development of Intrusion Detection Models. For example, if an attack found to be an anomaly may have its exemplary data processed by the model generator, and the dataset can be used to generate a model that can detect new intrusions and it then can be distributed to detectors.

2. Classification of IDS

2.1 Host based IDS

It gets the audit data from host audit trails and detect attacks from single host.

2.2 Distributed IDS

It gathers audit data from multiple hosts and possibly the network that connects hosts. It can detect attacks involving multiple hosts.

2.3 Network-based IDS

It uses the network traffic as audit data source and relieves the burden on the hosts that usually provide normal computing services. It can detect attacks from within a network.

3. Classification of IDS Techniques

3.1 Misuse Detection

It catches the intrusions in terms of the characteristics of known attacks or system vulnerabilities. It extract feature from known intrusions and integrate the Human knowledge. The rules are pre-defined however, it has disadvantage that it cannot detect novel or unknown attacks.

3.2 Anomaly Detection

It detects any action that significantly deviates from the normal behaviour. Sometime assume the training audit data does not include intrusion data. Any action that significantly deviates from the normal behaviour is considered intrusion. It has a disadvantage that when a noise (intrusion) data is in training data, it will make a misclassification.

4. Data Mining based approaches

For Intrusion Detection, Data Mining is used for classification and association of rules that describes the frequent association between two database records to indicate an attack on the network.

4.1 Supervised Learning-based approaches

For supervised learning for intrusion detection, there are mainly supervised neural network (NN)-based approaches & support vector machine (SVM)-based approaches. [1]

4.2 Unsupervised Learning-based approaches

An example of unsupervised learning for intrusion detection includes K-means-based approaches and self-organizing feature map (SOM). [1]

III. Related Work

A. Rule based Data Mining algorithm

In 2015, Kailas Elekar, M. M. Waghmare, Amrit Priyadarshi proposed this algorithm. In this paper, they focused on Rule based classification techniques such as:

- Decision Table
- JRip
- OneR
- Part
- ZeroR

The comparison of these Rule-based classification is presented and the performance is measured on metrics using WEKA tools and KDD-CUP dataset. The classification performance is evaluated using cross validation dataset. The comparison is also analysed based on False and Correct attack ratio.

- **Decision Table:** It uses tabular representation for describing and analysing situations. The decision is taken depending upon number of conditions.
- **JRip:** It is based in association rules with reduced error pruning (REP), and is a kind of Decision tree techniques. In REP for rules algorithms, the training data is split into a growing set and a pruning set. Hence Pruning of erroneous dataset provides an efficient classification. [2]
- **OneR:** It is a simplest association rlues, which involves only one attribute in condition part. It works well with real-world data.
- **Part:** It combines divide and conquer strategy. Incomplete tree are built in each step and rule is build using best leaf.
- **ZeroR:** It is the simplest classification method for simply predicting majority category (class). It relies on the target and ignores all predictors. It has no predictability power.

After the comparative analysis, PART emerged to be the efficient classifiers among others.

Classifiers	Attack Category				
	DOS	PROBE	U2R	R2L	Normal
Decision Table	99.99	97.85	32.75	88.00	99.28
PART	99.99	**99.36**	62.06	96.88	**99.96**
ZeroR	**100.00**	0.00	0.00	0.00	0.00
JRip	99.99	**99.36**	**67.24**	**97.60**	99.94
OneR	99.99	36.81	0.00	65.60	94.60

Fig 2. Category wise Percentage Attack Detection using Rule based classification technique [2]

B. Stream Data Mining and Drift Detection Method

In 2015, Manish Kumar, M. Hanumanthappa proposed this algorithm. The focus was on Concept Change Detection and Drift Detection. Concept drift is defined as a change in the underlying data generation process.

In the context of classification, concept drift is the change in statistical properties of the target variable, which the model is trying to predict, over time.The distribution generating the items of a data stream can change over time.

Concept drift is an unforeseen substitution of one data source S1 with another source S2. As concept drift is assumed to be unpredictable periodic seasonality is usually not considered as a concept drift problem. [3] The concept of this is the uncertainty about the future.

Input:	S: A data stream of examples C: Classifier
Output:	W: A window with the examples selected to train classifier C

1: Initialize $(i, p_i, s_i, ps_{\min}, p_{\min}, s_{\min})$;
2: $newDrift \leftarrow false$;
3: $W \leftarrow \phi$;
4: $W' \leftarrow \phi$;
5: **for all** examples $x_i \in S$; **do**
6: **if** prediction $C(x_i)$ is incorrect **then**
7: $p_i \leftarrow p_i + (1.0 - p_i)/i$;
8: **else**
9: $p_i \leftarrow p_i - (p_i)/i$
10: compute s_i using equation 1
11: $i \leftarrow i + 1$;
12: **if** $i > 30$ (approximately normal distribution) **then**
13: **if** $p_i + s_i \leq ps_{\min}$ **then**
14: $p_{\min} \leftarrow p_i$;
15: $s_{\min} \leftarrow s_i$;
16: $ps_{\min} \leftarrow p_i + s_i$;
17: **if** drift detected (3) **then**
18: Initialize $(i, p_i, s_i, ps_{\min}, p_{\min}, s_{\min})$;
19: $W \leftarrow W'$;
20: $W' \leftarrow \phi$;
21: **else if** warning level reached (2) **then**
22: **if** $newDrift \leftarrow true$ **then**
23: $W' \leftarrow \phi$;
24: $newDrift \leftarrow false$
25: $W' \leftarrow W' \cup \{x_i\}$;
26: **else**
27: $newDrift \leftarrow true$
28: $W \leftarrow W \cup \{x_i\}$;

Fig 3. Rule Induction Algorithm [3]

Upon detecting a drift, Rule Induction Algorithm (Fig.3) uses rule qualities to determine which rules should be removed from the model and which instances should be removed from the old data. It is thus important to observe the effect of the choice of rule quality on performance.

C. Integrated Fuzzy GNP Rule Mining with Distance-based classification.

Genetic Network Programming is an evolutionary algorithm. It has directed graph structure which is different from Genetic Algorithm and Genetic Programming [4].In this paper, a new classification method is proposed. In which, it extracts rules from training data and two-dimensional projection is done based on distance.

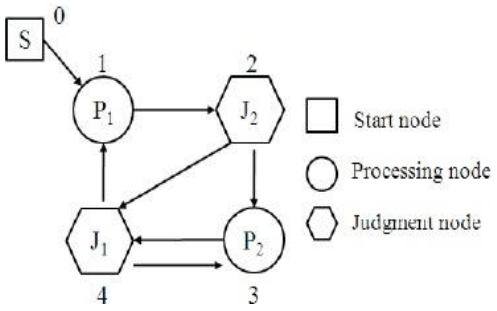

Fig 4. GNP Structure [4]

Genetic Operators:

- Individual Identity(IVI)
- Chromosome Fitness(CF)
- Individual Fitness(IVF)
- Upgrade Index(UI)

Hence, by projecting multi-feature space into a two dimensional average matching degree space, the new classification method fully utilize the distances of an observed dataset to its closest neighbour points.

D. Signature-Based Anomaly Intrusion Detection using Integrated Data Mining Classifiers

In this paper, the focus was mainly on Naïve Bayes and Random forest to decrease false alarms as well as to generate signatures based on detection.

Naïve Bayes: NB classifier is an uncomplicated probability classifier based on independent or rational assumption, Bayes theorem, and independent characteristic model. NB is easy to implement and applied, and also capable to handle continuous data and missing attribute values. [5]

Random Forest: Random Forest (RF) fuses more than on Decision Trees, and extracts a single tree to produce prediction. [5] Therefore, RF can be viewed as an ensemble learning method where various models are employed to achieve finer prognostic performance.

E. Integrated Fuzzy GNP Rule Mining with Distance-based classification.

Genetic Network Programming is an evolutionary algorithm. It has directed graph structure which is different from Genetic Algorithm and Genetic Programming [4].In this paper, a new classification method is proposed. In which, it extracts rules from training data and two-dimensional projection is done based on distance.

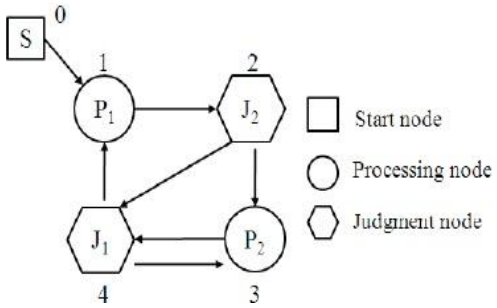

Fig 4. GNP Structure [4]

Genetic Operators:

- Individual Identity(IVI)
- Chromosome Fitness(CF)
- Individual Fitness(IVF)
- Upgrade Index(UI)

Hence, by projecting multi-feature space into a two dimensional average matching degree space, the new classification method fully utilize the distances of an observed dataset to its closest neighbour points.

F. Signature-Based Anomaly Intrusion Detection using Integrated Data Mining Classifiers

In this paper, the focus was mainly on Naïve Bayes and Random forest to decrease false alarms as well as to generate signatures based on detection.

Naïve Bayes: NB classifier is an uncomplicated probability classifier based on independent or rational assumption, Bayes theorem, and independent characteristic model. NB is easy to implement and applied, and also capable to handle continuous data and missing attribute values. [5]

Random Forest: Random Forest (RF) fuses more than on Decision Trees, and extracts a single tree to produce prediction. [5] Therefore, RF can be viewed as an ensemble learning method where various models are employed to achieve finer prognostic performance.

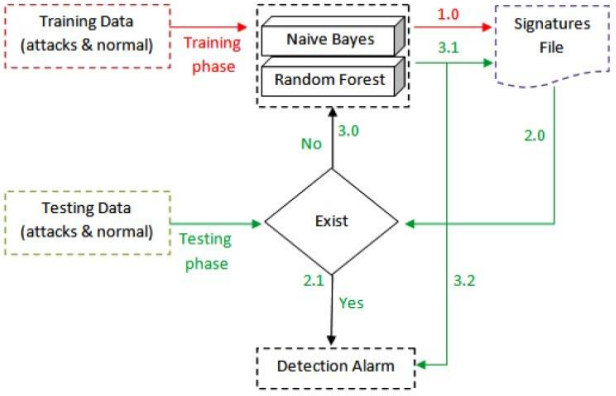

Fig 5. Signature based Anomaly Intrusion Detection Model [5]

This model focuses on detecting familiar and unfamiliar attack behaviour more precisely, include reducing the computation duration. The experiment result demonstrates that the detection capabilities and duration of the proposed method is more significant to be employed as intrusion detection scheme in contrast to conventional anomaly-based detection system (CADS).

A. Hoeffding Tree classification

In this paper, the researcher focussed on different data streaming techniques such as Hoeffding, Naïve Bayes, and Classifier ensemble. Among them Hoeffding proves to be most effective for Stream Data Mining classification.

Hoeffding algorithm is a decision tree learning algorithm and an effective way of classification of data points. It consists of the test node, root node and the leaf nodes, where each leaf node denotes prediction of class. Hoeffding algorithm

combines the data into a tree while the model is being built incrementally, even at that time we can use to classify data.

These are some advantages discussed, but there are also some disadvantages of this algorithm. If ties occur in the dataset, then holding fails to classify data into the tree.

Naïve Bayes is the probability based classifier while Hoeffding is Decision Tree algorithm. From the results comparison, it is found that Naïve Bayes gives more accurate results but takes more time, while Hoeffding gives nearby accuracy as of the Naïve Bayes but takes less time in giving results. Hence, for future scope, one can combine Hoeffding with any other techniques for a better classification.

Sr. No.	Author(s)/Year	Paper Title	Techniques	Results
1	Kailas Elekar, M. Waghmare, Amrit Priyadarshi, 2015	Intrusion Detection System using Stream Data Mining and Drift Detection Method[2]	Decision Table, JRip, ZeroR, OneR, PART	Efficient classification for particular types of attacks such as DOS, PROBE, U2R, R2L
2	Manish Kumar, Dr. Hanumanthappa, 2013	Intrusion Detection System using Stream Data Mining and Drift Detection Method[3]	Stream Data Mining, Concept Drift Detection	Efficient for Stream Data mining classification, suitable for Heterogeneous set of data
3	Nannan Lu, Shingo Mabu, Tuo Wang, Kotaro Hirasawa, 2012	Integrated Fuzzy GNP Rule Mining with Distance-based Classification for Intrusion Detection System[4]	Fuzzy Rule, Genetic Network Programming	Fuzzy GNP rule combination projects multi-feature space into a two dimensional degree space and detects anomaly detection
4	Warusia Yassin, Hazura Zulzalil, 2014	Signature-Based Anomaly Intrusion Detection using Integrated Data Mining Classifiers[5]	Naïve Bayes, Random Forest	Detection capabilities and duration of the proposed method is more significant to be employed as IDS than those used in conventional IDS (CADS)
5	Ketan Sanjay Desale, Chandrkant Kumathekar, Arjun Chavan, 2015	Efficient Intrusion Detection System using Stream Data Mining Classification Technique[6]	Hoeffding, Naïve Bayes, Classifier Ensemble	Naïve Bayes has more accuracy but takes more time where Hoeffding has nearby accuracy but takes less time

TABLE 1 - COMPARISON OF CLASSIFICATION ALGORITHM

Conclusion

Intrusion Detection is used for detecting known and unknown attacks on a network system. Since the increasing size of traffic and amount of data storage is huge, we can't go for a simple network algorithm or techniques such as Firewall to detect intruders in our system. To efficiently classify and analyse kind of intrusion in a particular system, some sort of data mining classification techniques needs to be employed which will not only report the intruders to the System Administrators, but also store the data for historical analysis and for generating evolutionary models for unknown attacks. In past, various classification techniques has been employed but they all fall short somewhere or other. Depending on the application or kind of data, one can go for one or more combination of those techniques in order to achieve an efficient classification for a scalable Intrusion Detection System. In future, we can combine Hoeffding technique with GNP Rule in order to design a highly efficient Intrusion Detection System.

References

1. Shilpreet Sigh, Meenakshi Bansal, "A Survey on Intrusion Detection System in Data Mining", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume No. 2, Issue No. 6, June 2013
2. Ketan Sanjay, Desale Chandrakant, Namdev Kumathekar,"Efficient Detection System using Stream Data Mining Classification Technique", 2015 International Conference on Computing Communication Control and Automation, 978-1-4799-68923/15 © 2015 IEEE
3. Kailas Elekar, M.M. Waghmare, Amrit Priyadarshi "Use of rule base data mining algorithm for Intrusion Detection ", International Conference on Pervasive Computing (ICPC) , -1-4799-6272-3/15 (c)2015 IEEE
4. Manish Kumar , Dr. M. Hanumanthappa "Intrusion Detection System using Stream Data Mining and Drift Detection Method ", IEEE – 31661 4th ICCCNT 2013
5. Warusia Yassin, Azizol Abdullah " Signature-Based Anomaly Intrusion Detection using Integrated Data Mining Classifiers", 2014 International Symposium on Biometrics and Security Technologies (ISBAST), 978-1-4799-6444-4/14 ©2014 IEEE
6. Hai Jin, Jianhua Sun, Hao Chen, Zongfen Han, " Efficient Intrusion Detection System using Stream Data Mining Classification Technique", Proceedings of the 10th IEEE International Workshop on Future Trends of Distributed Computing Systems (FTDCS'04) Technologies (ISBAST), 0-7695-2118-5/04 © 2004 IEEE
7. S. Devaraju, S .Ramakrishnan "Detection of Accuracy for Intrusion Detection System using Neural Network Classifier" International Journal of Emerging Technology and Advanced Engineering(ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
8. C. So, N. Mongkonchai, P. Aimtongkham, K. Wijitsopon and K. Rujirakul " An Evaluation of Data Mining Classification Models for Network Intrusion Detection", DICTAP 2014, Page No: 90 – 94
9. J. Sanejunthichai, "Real Time Network Communication Data Analysis System in Order to Detect Internet Worm by Using Decision Tree Technique," In Proceedings of National Symposium on Applied Computing Technology and Information System, pp.118–124, 2011
10. Damon Sotoudeh, Aijun An,CIKM'10 Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Pages 769-778, 2010
11. Parekh S P, Madan B S, Tugnayat R M, "Approach for Intrusion Detection System Using Data Mining", Journal

of Data Mining and Knowledge Discovery, ISSN: 2229–6662 & ISSN: 2229–6670, Volume 3, Issue 2, 2012, pp.-83-87.

12. N. Lu, S. Mabu and K. Hirasawa, Integrated rule mining based on Fuzzy GNP and probabilistic classification for intrusion detection, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 15, No. 5, pp. 495-505, 2011.
13. P. Louvieris, N. Clewley, and X. Liu, "Effects-based feature identification for network intrusion detection," Neuro computing vol. 121, no. 0, pp. 265–273, 2013.
14. Bifet, Albert. "Mining Big Data in Real Time", Informatica37, pp: 15-20, 2013.