A Survey: Analysis of Current Approaches in Anomaly Detection

Prashansa Chouhan
M.Tech (CSE), LNCT Bhopal
Affiliated to RGPV

Vineet Richhariya
HOD, CSE LNCT Bhopal
Affiliated to RGPV

ABSTRACT
An anomaly is abnormal activity or deviation from the normal behaviour. Anomaly detection is the process of removing these abnormal or anomalous behaviours from the data or services. Anomaly detection techniques are used to detect and discard anomalies from the data or services. In this survey paper we describe overview of some anomaly detection techniques which are on collective anomaly detection and clustering anomaly which are generated due to variety of abnormal activities such as credit card fraud detection, mobile phone fraud, banking fraud, cyber attack etc. an important aspect as the nature of anomaly. In existing paper introduced the concept of collective anomaly for network traffic analysis. It's used the variant of k-mean and x-mean algorithm for clustering network traffic and detects DOS attack. In the anomaly detection models anomalies are detected by comparing the tracing data with the actual data. On the basis of comparison deviations in the traced data or services are identified and they are considered as anomaly. To overcome these entire problems we proposed a novel technique to the combination of classification and Genetic based anomaly. We develop an efficient sampling technique which capture the underlying distribution of data and create a summary to be able to monitor high capacity network.

Keywords
Anomaly detection techniques, clustering, CAD, genetic and classification based technique.

1. INTRODUCTION
Data mining is collection of information which are identified the pattern and established the relationship between data and its multiple attributes. In a data mining basically done the extraction of implicit, previously unknown and potentially useful information from database. Data mining is sometimes called knowledge discovery. Knowledge discovery is a process that extracts implicit, potentially useful or previously unknown information from the data.

The term data mining is employed for methods and algorithms that allow analyzing data in order to find rules and patterns describing the characteristic properties of the information. Data mining techniques are attractive as they can be applied to any kind of data in order to learn more about hidden structures and correlations.

1.1 IDS
Intrusion detection systems (IDS) process large amounts of observing data. As an example, a host-based IDS examines log files on a computer (or host) in order to detect suspicious activities. A network-based IDS, on the other hand, searches network observing data for risky packets or packet flows. In the late 1990s, progress in data mining research and the necessity to find better methods for network and host based intrusion detection resulted in research activities attempting to deploy data mining techniques for anomaly and attack detection.

There are two types of IDSs: signature-based and anomaly-based. Signature-based IDSs exploit signatures of known attacks. Such systems require frequent updates of signatures for known attacks and cannot detect unknown attacks or anomalies for which signatures are not stored in the database.

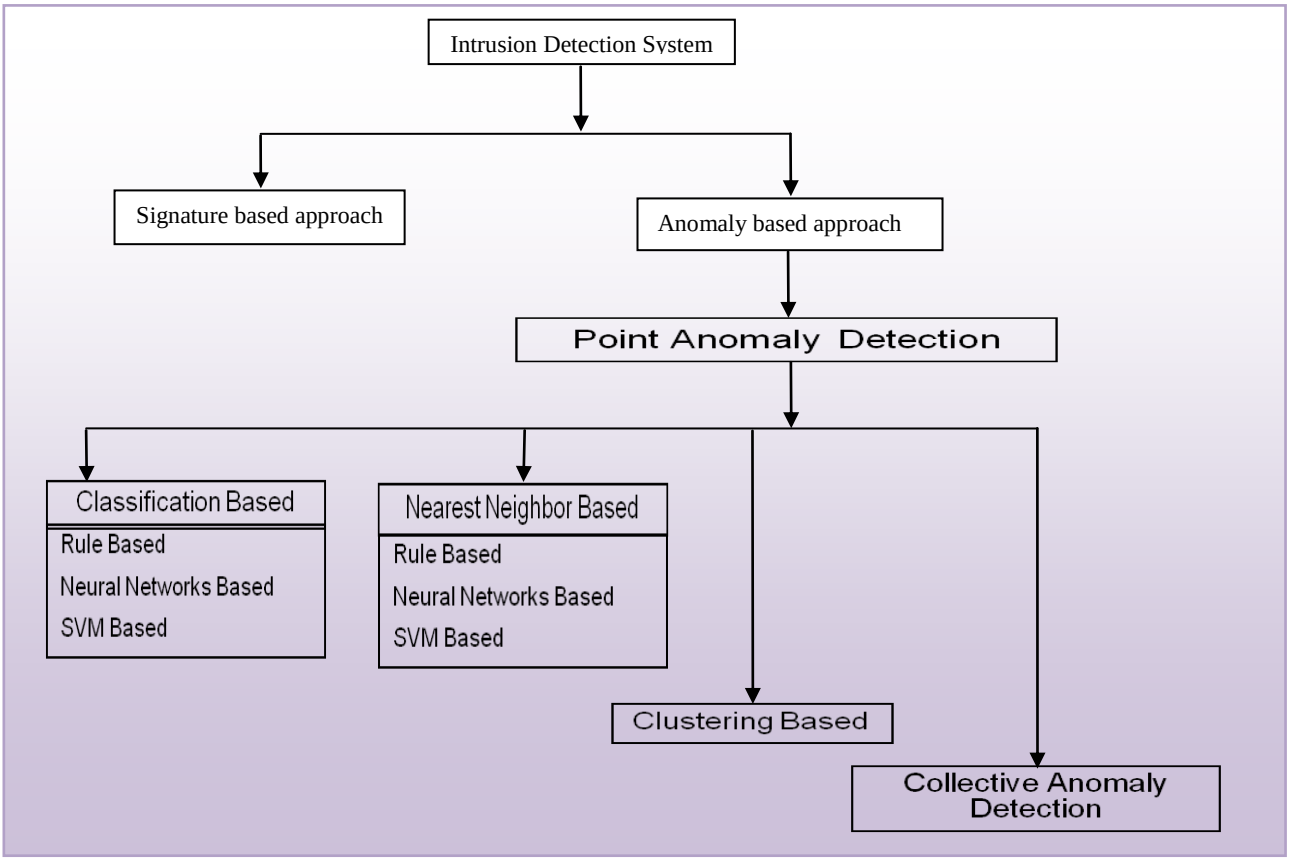

Figure 1: Classification of IDS

Intrusion detection systems are software's used for recognizes the intended or unintended use of the system resources by unauthorized users. They can be classified into misuse detection systems and anomaly detection systems. Misuse detection systems representation attacks as a definite pattern and are more valuable in detecting known attack patterns. If the intrusion happens through learning, then the anomaly detection system may find out the intruder's behaviour and for this reason may fail. Being more generalized and having a extensive possibility as compared to misuse detection systems, most of the present techniques focus on anomaly detection systems. Data mining approaches can be applied for both anomaly and misuse detection. Clustering techniques can be used to form clusters of data samples equivalent to the usual exploit of the method. Clustering based techniques can identify new attacks as compared to the classification based techniques.

1.2 Anomaly Detection and its Issues

Anomaly detection is about finding the normal practice patterns from the audit data, while misuse detection is on the subject of encoding and matching the intrusion patterns via the audit data. Anomaly detection refers to the important problem of finding non-conforming patterns or behaviours in live traffic data. These non-conforming patterns are often known as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains. In practice, it is very difficult to precisely detect anomalies in network traffic or normal data. So, anomaly is an interesting pattern due to the effect of traffic or normal data while noise consists of non-interesting patterns that hinder traffic data analysis.

- **Number of Attributes:** since an object may have many attributes, it may have anomalous values for some attributes; an object may be anomalous even if none of its attribute values are individually anomalous.
- **Global Vs Local Perspective:** an object may seem unusual with respect to all objects, but not with respect to its local neighbours.
- **Degree of Anomaly**: some objects are more extreme anomalies than others; it's desirable to have some assessment of the degree to which an object is anomalous (outlier score).
- **One at Time Vs many at Once:** is a better remove anomalous object one at time or identify a collection of objects together. Two distinct problems: masking, where the presence of an anomaly masks the presence of other; swamping, where normal objects are classified as outliers.
- **Evaluation:** find a good measure of evaluation for the process of anomaly detection when class labels are available and when class labels are not available (precision, recall, FP-rate, accuracy).
- **Efficiency:** calculate the computational cost of the process of anomaly detection scheme.

Anomaly Detection main approach are statistical approach, Proximity-Based, Density-Based, Clustering-Based.

1. **Statistical Approaches** are model-based which are based on building a probability distribution model consider how likely objects are under that model an outlier is an

object that has a low probability with respect to a probability distribution model of the data.

2. **Proximity-Based Approaches** Simplest way to measure proximity distance to the k-nearest neighbours outlier score is given by the distance to its k-Nearest Neighbours. An object is an anomaly if it is distant from most points (k) this scheme is simple it's easier to determinate a "good" measure than to determinate the statistical distribution.
3. **Density-Based** outlier detection is closely related to Proximity-**Based** outlier detection since. Density is usually defined in terms of Proximity.
4. **Clustering** used to find groups of (strongly) related objects Anomaly Detection used to find objects that are not (strongly) related to other objects.

2. LITERATURE REVIEW

2.1 Supervised Approaches

In this approach, a predictive model is developed based on a training dataset (i.e., data instances labeled as normal or attack class). Any unseen data instance is comparedagainst the model to determine which class it belongs to. There are two major issues that arise in supervised anomaly detection. First, the anomalous instances are far fewer in number compared to normal instances in the training data. Issues that arise due to imbalanced class distributions have been addressed in the data mining and machine learning literature [5]. Second, obtaining accurate and representative labels, especially for the anomaly class is usually challenging. A number of techniques have been proposed that inject artificial anomalies in a normal dataset to obtain a labeled training dataset [6]. Other than these two issues, the supervised anomaly detection problem is similar to building predictive models. We now discuss some of the most common incremental supervised anomaly detection approaches.

Ren et al. [2]: The authors propose a new anomaly detection algorithm that can update the normal profile of system usage dynamically. The features used to model a system's usage pattern are derived from program behaviour. A new program behaviour is inserted into old profiles by density-based incremental clustering when system usage pattern changes. It is much more efficient compared to traditional updating by re-clustering. The authors test their model using the 1998 DARPA BSM audit data, and report that the normal profiles generated by their algorithm are less sensitive to noise data objects than profiles generated by the ADWICE algorithm. The method improves the quality of clusters and lowers the false alarm rate.

2.2 Semi-supervised Approaches

In semi-supervised approach, the training data instances belong to the normal class only. Data instances are not labelled for the attack class. There are many approaches used to build the model for the class corresponding to normal behaviour. This model is used to identify anomalies in the test data. Some of the detection methods are discussed in the following.

Burbeck et al. [3]: ADWICE (Anomaly Detection With fast Incremental Clustering) uses the first phase of the BIRCH clustering framework [5] to implement fast, scalable and adaptive anomaly detection. It extends the original clustering algorithm and applies the resulting detection mechanism for analysis of data from IP networks. The performance is demonstrated on the KDD99 intrusion dataset as well as on data from a test network at a telecom company. Their experiments show good detection quality (95%) and acceptable false positives rate (2.8 %) considering the online, real-time characteristics of the algorithm. The number of alarms is further reduced by application of the aggregation techniques implemented in the Safeguard architecture[1].

Rasoulifard et al. [4]: It is important to increase the detection rate for known intrusions and also to detect unknown intrusions at the same time. It is also important to incrementally learn new unknown intrusions. Most current intrusion detection systems employ either misuse detection or anomaly detection. In order to employ these techniques effectively, the authors propose an incremental hybrid intrusion detection system. This framework combines incremental misuse detection and incremental anomaly detection. The framework can learn new classes of intrusion that do not exist in data used for training. The framework has low computational complexity, and so it is suitable for real-time or on-line learning. The authors use the KDDcup99 intrusion dataset to establish this method.

2.3 Unsupervised Approaches

Unsupervised detection approaches do not require training data, and thus are most widely applicable. These techniques make the implicit assumption that normal instances are far more frequent than anomalies in the test data. If this assumption is not true, such techniques suffer from high false alarm. Most existing unsupervised anomaly detection approaches are clustering based. Clustering is a technique to group similar objects. It deals with finding structure in a collection of unlabeled data. Representing the data by fewer clusters necessarily leads to the loss of certain finer details, but achieves simplification. In anomaly detection, clustering plays a vital role in analyzing the data by identifying various groups as either belonging to normal or to anomalous categories. There are many different clustering based anomaly detection approaches in the literature.

Mohiuddin Ahmed[1] describe on collective anomaly detection and clustering anomaly which are generated due to variety of abnormal activities such as credit card fraud detection, mobile phone fraud, banking fraud, cyber attack etc. an important aspect as the nature of anomaly. In existing paper introduced the concept of collective anomaly for network traffic analysis. It's used the variant of k-mean and x-mean algorithm for clustering network traffic and detects DOS attack.

B.Senthilnayaki, K.Venkatalakshmi[7] describe on genetic algorithm and classification algorithm for anomaly detection intrusion detection system using soft computing techniques to offer effective security through the provision of detection accuracy, fast processing time, ability to adapt and exhibit fault tolerance. In this paper, intelligent algorithms for intrusion

detection are proposed which detect the network attacks as normal or anomaly based attacks by performing effective preprocessing and classification. This system uses a new genetic

algorithm approach for pre-processing and Modified J48 classification algorithm to identify the intended activities. The new genetic based feature selection algorithm proposed in this

paper is helpful to identify the important features needed to classify the normal and anomaly records.For this IDS, we propose a new genetic based feature selection algorithm which

reduces the 41 features of the KDD Cup data set into 9 important features by applying the fitness value as a threshold. Moreover, we perform classification using a modified decision tree algorithm which has been developed by enhancing the existing J48 decision tree algorithm. 99 dataset suffers from major weakness due to the presence of redundant records. These to redundant records reduce the detection rate and accuracy. KDD'99 dataset has 41 features with classes labeled as either normal or anomaly with specific attack type.

Table 1: Comparative analysis of Anomaly detection technique

S. No.	Auther	Anomaly Detection Algorithm	Advantages	Limitations
1	**Mohiuddin Ahmed**	k-mean and x-mean Clustering technique	Suitable for large and dynamic systems where it is difficult to monitor application-level knowledge of services	Performance degradation of services and issue of scalability
2	**Ren**	Density-based Supervised learning technique	Simple and easy to implement for anomaly detection	Detect only primary causes of anomalies
3	**Burbeck**	Aggregation techniques	Its implement fast, scalable and adaptive anomaly detection	Require vast amount of specific domain knowledge
4	**Rasoulifard**	Incremental Hybrid Detection Semi-supervised Approaches	It is simple, as it is much easier to model the correct use of a protocol than to model its misuse	It suffers from high false positive rate
5	**B.Senthilnayaki**	Genetic algorithm and classification algorithm	The major advantages of the proposed IDS are reduction in false positive and fast classification.	it is difficult to precisely model all behaviours since anomaly based detection can detect only known attacks.

3. CHALLENGES AND SOLUTION

Based on our survey of published papers on incremental anomaly detectors, we observe that most techniques have been validated using the KDD99 intrusion datasets in an offline mode. However, the effectiveness of an ANIDS based on incremental approach can only be judged in a real-time environment. Following are some of the research issues we have identified in this area.

- Most existing IDSs have been found inadequate with new networking paradigms currently used for both wired and wireless communication. Thus, adaptation to new network paradigms needs to be explored.
- Most existing methods are dependent on multiple input parameters. Improper estimation of these parameters leads to high false alarm rates.
- The clustering techniques that are used by anomaly detectors need to be faster and scalable when used on high dimensional and voluminous mixed type data.
- Lack of labelled datasets for training or validation is a crucial issue that needs to be addressed. The KDD99 datasets are out-of-date. New valid datasets need to be created and made available to researchers and practitioners. Developing a reasonably exhaustive dataset for training or validation for use in supervised or semi-supervised anomaly detection approaches is a challenging task.
- Estimation of unbiased anomaly scores for periodic, random or busty attack scenarios is another challenging issue.
- Lack of standard labelling strategies is a major bottleneck in the accurate recognition of normal as well as attack clusters.
- Development of pre or post-processing mechanisms for false alarm minimization is necessary.
- Handling of changing traffic pattern remains a formidable problem to address.

It's providing services to users in the proper and normal form; anomaly detection becomes important and interested area for research work. For anomaly detection so many techniques are developed and these techniques are broadly divided into three categories: - statistical, data mining based and machine learning based anomaly detection technique. Intrusions or attacks can be detected based on the collection of information from a network or a host (normally, in and out traffic). This problem has been formulated in from the pattern recognition view point. Incremental approaches are used to make the system faster in terms of training as well as testing of the instances. There are enormous problems in developing incremental techniques for network anomaly detection due to the dynamic updation of normal as well as attack profiles. We observe that the main importance of incremental network anomaly detection lies in dynamic profile updation for both normal and attack, reduced memory utilization, faster and higher detection rate, and improved real time performance. To address all these issues, we present a comprehensive survey of incremental approaches for network anomaly detection.

To overcome these entire problems we proposed a novel technique to the combination of classification and Genetic based anomaly. We develop an efficient sampling technique which capture the underlying distribution of data and create a summary to be able to monitor high capacity network.

4. CONCLUSION

In this paper we have discussed about various anomaly detection techniques which can be used in different conditions. The above mentioned techniques can differentiate between normal and anomalous behaviour of services on the basis of comparison between them. When performance of our data or service is deviate from the normal path it is considered as anomaly. Once anomaly is detected in the data it can be

removed using any suitable detection technique. Anomaly detection is an interesting arena of computer and network security. It is also applied in various application domains. It is regarded as one of the fundamental problems of data mining as well. In this paper, we have summarized anomaly detection techniques along with various research direction and application domains. We have described existing approaches for CAD.

5. REFERENCES

[1] Mohiuddin Ahmed, Abdun Naser Mahmood, "Network Traffic Analysis based on collective anomaly detection" 2014 IEEE 9th Conference on Industrial Electronics and Applications.

[2] F. Ren, L. Hu, H. Liang, X. Liu, and W. Ren, "Using density-based incremental clustering for anomaly detection," in Proceedings of the 2008 International Conference on Computer Science and Software Engineering. Washington, DC, USA: IEEE Computer Society, 2008, pp. 986–989. [Online]. Available: http://dx.doi.org/10.1109/CSSE.2008.811

[3] K. Burbeck and S. Nadjm-tehrani, "ADWICE - anomaly detection with real-time incremental clustering," in Proceedings of the 7th International Conference on Information Security and Cryptology, Seoul, Korea. Springer Verlag, pp. 4007-424, 2004.

[4] A. Rasoulifard, A. G. Bafghi, and M. Kahani, Incremental Hybrid Intrusion Detection Using Ensemble of Weak Classifiers, in Communications in Computer and Information Science. Springer Berlin Heidelberg, November 23 2008, vol. 6, pp. 577–584. [Online]. Available: http://10.1007/978-3-540-89985-3

[5] M. V. Joshi, I. T. J. Watson, and R. C. Agarwal, "Mining needles in a haystack: Classifying rare classes via two-phase rule induction," SIGMOD Record (ACM Special Interest Group on Management of Data), Vol. 30, No. 2, pp. 91-102, 2001.

[6] J. Theiler and D. M. Cai, "Resampling approach for anomaly detection in multispectral images," in Proc. SPIE, pp. 230–240, 2003.

[7] B.Senthilnayaki, K.Venkatalakshmi, A. Kann an, "An Intelligent Intrusion Detection System Using Genetic Based Feature Selection and Modified J48 Decision Tree Classifier" 2013 Fifth International Conference on Advanced Computing (ICoAC)

[8] P. Laskov, C. Gehl, S. Kr¨uger, and K.-R. M¨uller, "Incremental support vector learning: Analysis, implementation and applications," Journal of Machine Learning Research, vol. 7, pp. 1909–1936, 2006.

[9] S. Jiang, X. Song, H. Wang, J.-J. Han, and Q.-H. Li, "A clustering-based method for unsupervised intrusion detections," Pattern Recognition Letters, vol. 27, pp. 802–810, 2006.

[10] H. Cheng, P.-N. Tan, C. Potter, and S. A. Klooster, "Detection and characterization of anomalies in multivariate time series," in Proceedings of the SIAM (SDM), pp. 413– 424, 2009