Supervised Outlier Detection research papers

Here I have chosen all the Supervised Outlier detection model for my research as it is feasible to execute and models like Boltzmann, Data Mining , Mahala Nobis-distance based outlier detection, NN nearest neighbor, and K means are easy to train in comparison to the former model. Moreover, over my research and validation practically I have identified their significance over particular scenarios in terms of when to use and with what data sets.

Paper 1:

Network anomaly detection with the restricted Boltzmann machine

This research paper focusses on Discriminative Restricted Boltzmann Machine where it demonstrates a model to classify the network abuses happening over the wide area network. This model has been devised strategically to ensure that the data (network confidentiality) remains safe and secure both from internal as well as external influences. The main problem for this research has been listing out the criteria or factors for judgement for normal and abnormal network behaviors that would lead to proper classification of the data.

In this the focus is put on semi-supervised anomaly detection, with a view of investigating whether normal traffic behavior (and, conversely, anomalous behavior) shares some inherent similarity that we can use to characterize. The tool which has been selected for this analysis is the Discriminative Restricted Boltzmann Machine, a network of stochastic neurons behaving according to an energy-based model. The network couple has an ability to express variability of data, given by generative models, where the discriminative classifiers are used for classifying the data variables correctly over a network.

The main advantages of this approach is that ,”it is not restricted to any specific environment, or a priori knowledge base, and that it can enable the detection of any type of unknown anomalous events, being effective in coping with the so-called zero-day attacks.” (UgoFiore, FrancescoPalmieri, AnielloCastiglione, & Santis, 2012)

The main disadvantage arises when there is a form of inherent characteristic of deep similarity between all normal traffic and some anomalous traffic; is where the model fails. Moreover, this model assumes that the traffic changes over a longer period although this may not be the case in real world. It is also criticized as being heavily.

The primary concept depends upon the availability of a model aiming at characterizing current behavior, based on a statistical idealization of past events. There play two main challenges here, namely the nature of network traffic being complex and unpredictable and dynamic background of network model in play which makes the anomalies continuously evolving.

The following model revolves around network(namely Wide Area Network) Data.

Paper 2:

# **Survey on Anomaly Detection using Data Mining Techniques**

In this research article the Data mining techniques have been taken into consideration for detecting vulnerabilities while transfer and storage of data from one platform to another. As these days, data playing a crucial part in substantial economy has made it prime to make it security and check all its vulnerabilities to be mitigated as efficiently as possible. “Anomaly detection uses these data mining techniques to detect the surprising behavior hidden within data increasing the chances of being intruded or attacked.” (ShikhaAgrawal & JitendraAgrawal, 2015).

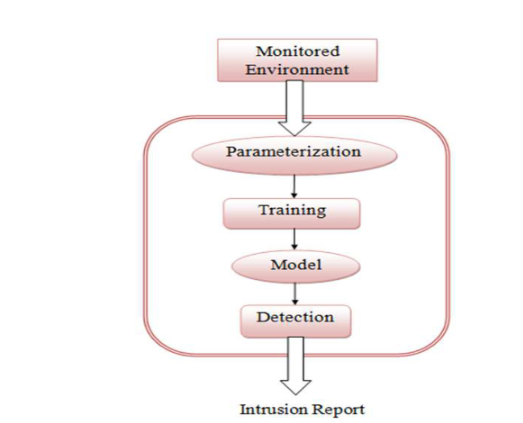
This research focusses on various techniques like clustering, classification, Data Mining and intrusion detection.

Here, too the randomness of anomalies is taken into consideration and various hybrid approaches are taken into effect for this method in order to work in place.

The below figure represents the process flow of how the model goes onto using mining as its solution for this problem.

The main advantage of this model is that is uses hybrid approaches like Cascading supervised techniques, combining supervised and unsupervised techniques which really help with the feasibility issue of the problem.

The disadvantage is that in this research work fusion or combination of already existing algorithms are that have been proposed and that the new and unknown or dynamic outliers’ detection is difficult.



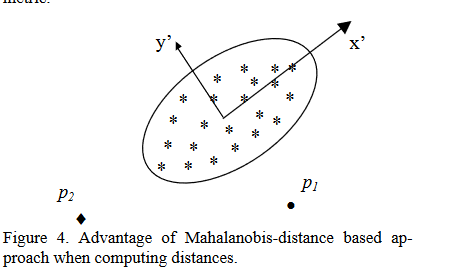
For this model the Data Transactions (Of an institution most likely a firm/organization) is taken into consideration i.e. large data set.

Paper 3:

A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection

In this research paper the anomaly detection against computers and network infrastructure attacks. In this a wide range of attacks are taken into account which include the intentional (formal and informal) as well as unintentional attacks (which may consist of induced, faults, defects etc.).

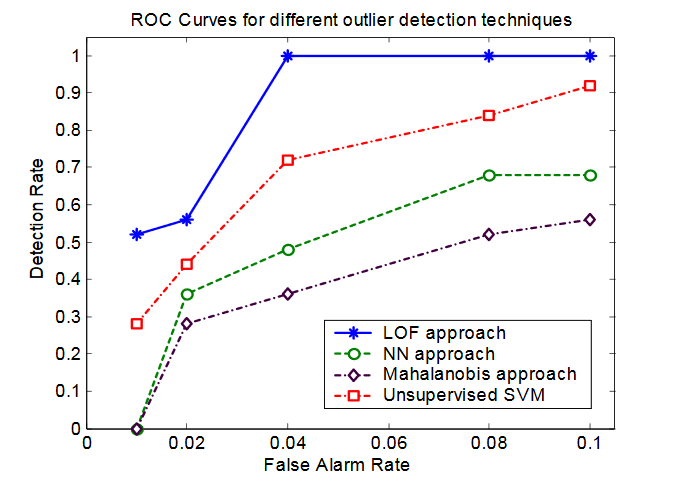
“Several existing supervised and unsupervised anomaly detection schemes and their variations are evaluated on the DARPA 1998 data set of network connections [9] as well as on real network data using existing standard evaluation techniques as well as using several specific metrics that are appropriate when detecting attacks that involve a large number of connections.” (Aleksandar Lazarevic, 2003)



The advantage of this model is that it lays down models like mining outliers using K-th nearest neighbor, Mahala Nobis-distance based outlier detection which proved to be much efficient for ROC as evident from the above diagram.

The disadvantage for this model is “several slow probing attacks that scan the hosts (or ports) using a much larger interval than 5 seconds (e.g. one scan per minute hour). Therefore, these attacks time-based features.” (Aleksandar Lazarevic, 2003)

For simplification a procedure for extracting statistical content based and temporal features is also implemented.

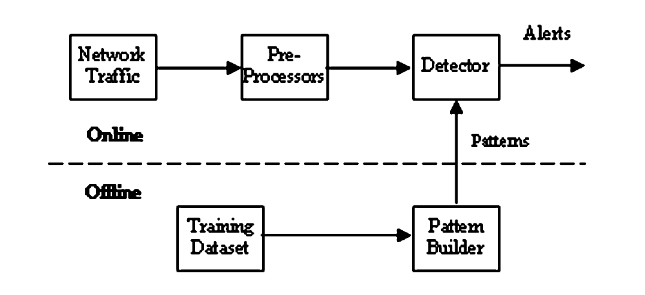


Here the variate data of a cyber security firm is taken and evaluated on it.

Paper 4:

Random-Forests-Based Network Intrusion Detection Systems

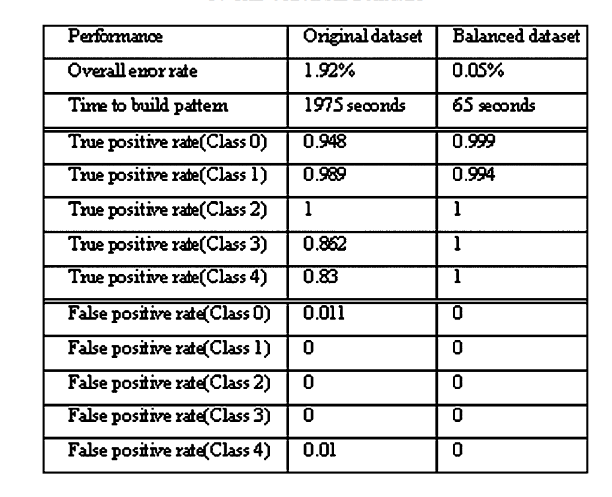
This research paper focusses on detecting novel intrusions in the network like IDs (intrusion detection systems) which are rule based. Here they primarily focus on the algorithm random forest which is a data mining algorithm which is used in misuse, anomaly and hybrid-network based IDs. “After building the patterns of network services by the random forest’s algorithm, outliers related to the patterns are determined by the outlier detection algorithm. The hybrid detection system improves the detection performance by combining the advantages of the misuse and anomaly detection.” (Jiong Zhang, 2008)



The major advantages of this model is that it performs well when the false positive rate is low and the presented hybrid system improves the overall performance of the IDs. The random forests algorithm can build the patterns automatically instead of coding rules manually.

The disadvantages are as the outlier detection only works when most data are normal. We use misuse detection to filter out known intrusions. However, this cannot guarantee that most activities are normal after removing known intrusions. Moreover, it may undermine the hybrid system. Second, some intrusions with high degree of similarity cannot be detected as outliers by the anomaly detection. (Jiong Zhang, 2008)

The evaluation matrix is summarized as follows:



Here again a network structure data is taken into consideration for evaluating the model.

Paper 5:

**Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach**

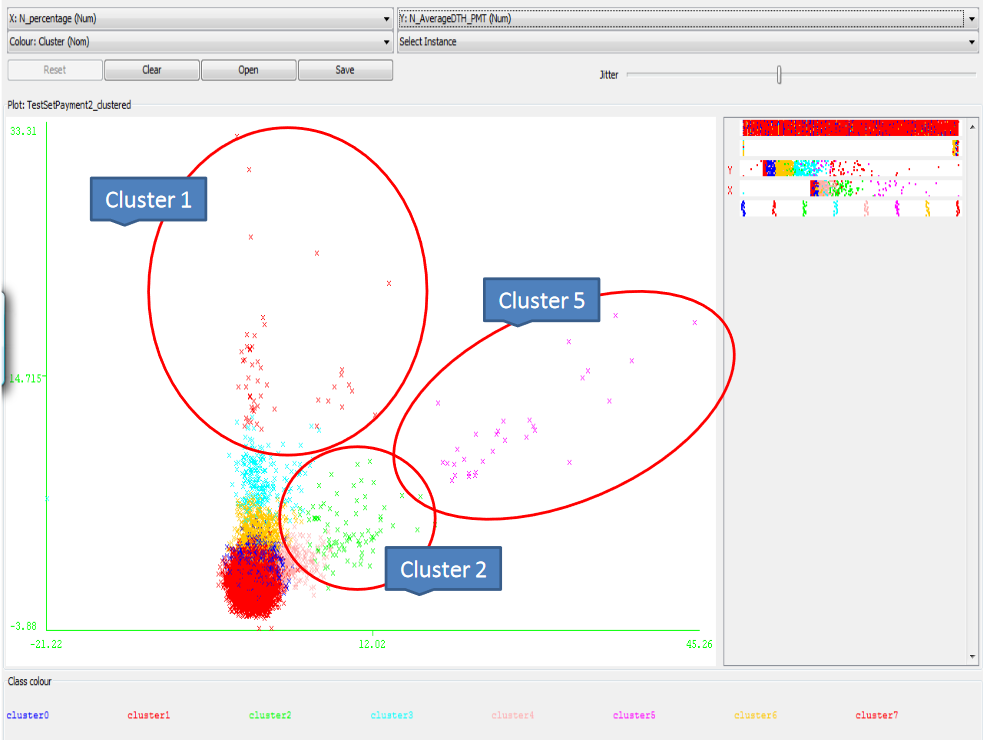
In this Research paper, it focusses on cluster analysis in the accounting domain and finding discrepancies in the audit data. It works to examine the clustering technology to automate the process of fraud detection and filtering during the process of automation. A simple K-means clustering approach is devised for modeling this problem and it is run on an open source software, WEKA.

“Outliers are identified in two ways. First, that have low probability of being a member of a cluster (i.e. are far away from other cluster members) are identified as outliers. The probability of 0.6 is used as a cut-off point. Second, clusters with small populations (less than 1% of the total) are considered outliers.” (Thiprungsri. & Vasarhelyi, 2011)

The main advantages of this model are that it is a promising technique that can be integrated in a schema of continuous system monitoring and assurance which reveals acceptable clusters, and efficient measurement of distance from the clusters.

The only disadvantage for this model is that it may flag transactions identified via other methodologies.

The evaluation matrices can be seen from the graph below and visualized as it efficiently forms all the clusters and identifies data, so it is a good tactic.



Here the Audit data that is large and messy is taken into consideration for evaluation.

# References

Aleksandar Lazarevic, L. E. (2003). A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. *Proceedings of the 2003 SIAM International Conference on Data Mining*, 12.

Jiong Zhang, M. Z. (2008). Random-Forests-Based Network IntrusionDetection Systems. *IEEE*, 111.

ShikhaAgrawal, & JitendraAgrawal. (2015). Survey on Anomaly Detection using Data Mining Techniques. *Procedia Computer Science*, 713.

Thiprungsri., S., & Vasarhelyi, M. A. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *The International Journal of Digital Accounting Research*, 84.

UgoFiore, FrancescoPalmieri, AnielloCastiglione, & Santis, A. (2012). Network anomaly detection with the restricted Boltzmann machine. *ScienceDirect*, 23.