Randomizing Greedy Ensemble Outlier Detection with GRASP

Lediona Nishani

A dissertation submitted to the faculty of the University of New York Tirana
in partial fulfillment of the requirements for the Doctoral degree in
Computer Science in the Department of Computer Science
in the Faculty of Informatics

Advisor
Assoc. Prof. Marenglen Biba, PhD

Tiranë, prill 2017

Randomizing Greedy Ensemble Outlier Detection with GRASP

Lediona Nishani

A dissertation submitted to the faculty of the University of New York Tirana
in partial fulfilment of the requirements for the Doctoral degree in
Computer Science in the Department of Computer Science
in the Faculty of Informatics

Tirane, 2017

Approved by:

Prof. Dr. Nertila Gjini

Prof. Dr. Fatmir Hoxha

Prof. Dr. Kozeta Sevrani

Assoc. Prof. Marenglen Biba

Assoc. Prof. Igli Tafaj

Abstract

Machine learning algorithms are first trained with reference input to "learn" its specifics, and this process may be supervised or unsupervised. Models are then deployed on unseen input for detection purposes. Machine learning methods enable systems to learn from experience, and often the system starts with some prior corresponding knowledge that analyzes and tests the data acquired.

Outlier detection methods have been extensively implemented in a variety of fields like medicine, reaction analysis, financial fraud detection, security counter terrorism, intrusion detection, etc. In addition, many other areas in the years to come will exploit the properties and advantages of outlier detection techniques. There are many definitions for denoting outliers. The major work elaborated so far in the field of outlier detection has embraced the statistical theoretical framework, (Hawkins, 1980): *"An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*". In the data-mining domain, outlier detection schemes are designed to mine exceptional activity behaviour of particular data.

Traditional schemes published on outlier detection have addressed the task of detecting outliers based on the underlying principles of statistical frameworks, such as LOF (Kriegel et al, 2009), LOCI (Papadimitriou et al. 2003), LoOP (Kriegel et al, 2009), etc. The task of discovering outliers has been tackled and covered by various approaches that can be divided into global and local outlier techniques. This feature has emerged from the type of a database considered as outlier of a particular data point. The difference here stands for the fact that the global methods take into consideration the entire database, while the local ones examine only some subset of the dataset or space. Local Outlier Factor (LOF) (Kriegel et al, 2009) is the first outlier detection algorithm that deals with local outliers and mechanisms that cannot be captured by the global methods. LOF has been developed as an emerging need for defining a precise degree for an object to be considered an outlier. The term local is referred to a concept of locality, in other words, the scale of isolation an object experiences to its neighbourhood. The LOF approach has given a careful evaluation of algorithm's performance in capturing abnormal activities in practical real world data. This is the first attempt to quantify the outliner-ness of objects in a dataset. LOF has been deployed in multidimensional datasets; in addition, every single item has been associated to a local factor, which shows the degree of each object of being outlier. This algorithm is related to the principle of density based clustering, and the local term means that limited objects in the surroundings are taken into consideration. The idea underlying this approach stands in the assumption that the item in a dataset included in a cluster is assigned a local outlier factor that appears to be close to 1. Every LOF object has the upper bound and the lower bound, which is the range of lying in the local outlier nature. Moreover, bounds close to each other are investigated while for items that do not satisfy this criterion, the approach attempts to sharpen the bounds. The new notion here is that the outlier is not a binary property. In order to calculate the LOF of each object, this approach presents a new notion, which is the distance, designated as k-distance of the object p.

The focus of this dissertation is to create an ensemble algorithm composed of the LOF family of algorithms that could exhibit an increase in performance, accuracy and diversity. We concentrate on diversity. Our contribution consists in designing a new algorithm upon the greedy ensemble algorithm by applying randomization with GRASP procedure. We have tested this new approach and provided evidence of enhancement in the diversity of results our novel mechanism has generated. Moreover, we have conducted thorough experiments for identifying the improvements and drawbacks of our approach.

To the best of our knowledge, this is a novel approach that has not been tackled from other research work. We hope to extend this work by adding other features to improve further our approach in terms of accuracy and diversity.

To my Family

Acknowledgements

No guide, no realization.

I would like to thank many persons for being helpful to me during the hard working period of my doctorate studies.

First of all, I am profoundly grateful to my advisor, Marenglen Biba who has followed and helped me in every step when encountering difficulties and has been of significant support to me during these years. He has contributed to sharing his thoughts and elaborating ideas regarding my dissertation. He has been patient enough to listen to my concerns and encouraged me when something was not going the proper way by sharing his research experience, which has been of substantial benefit to my dissertation.

I am very grateful to the University of New York in Tirana, which accommodated me in their environment and provided the necessary technical equipments for conducting crucial experiments and testing. Furthermore, I would like to thank all the professors that have delivered us lectures in the first year of the PhD studies by transferring and fostering scientific research skills, which were of real benefit to me for the upcoming doctoral years.

One particular help and support to me has been Dr. Erich Schubert from the University of Ludwig-Maximilians of Munich, member of the Research Unit of Database and Information System. We have exchanged many ideas, and he has helped when something in the code was not going properly.

I would like to thank my cousin, Eralda for helping me in my journey. We had almost the same related field of dissertation topic. In addition, I have discussed with her for particular issues of the dissertation, more specifically, to implement the ideas in the data mining platform. Without her help my research would have been incomplete.

An immense gratitude goes to my husband Ermal, with whom I have shared my concerns and thoughts during this hard working period. Talking about my dissertation and various issues, discussing with him every day, has induced to me novel ideas and positive energy for fulfilling this immense research work.

Finally, last but not least, a special appreciation goes to my family, especially my mother, which was the person who encouraged me in the beginning to pursue the doctoral degree in computer science. She has always believed in me and supported me in every step of my carrier. Subsequently, I would really want to thank my sister for being deeply caring during this time and for encouraging me to accomplish my dreams. All this work is dedicated to them, for their deep support throughout my life, for having fostered in me the spirit of curiosity and to always continue my path until I reach the destination.

Lediona Nishani
Tirane, April 25th 2017

Contents

List of Figures

List of Tables

List of Algorithm

Abbreviations

ABOD	Angle-Based Outlier Detection, see Section 4.5.4
ABOF	Angle-Based Outlier Factor, see Section 4.5.4
aLOCI	approximate algorithm of LOCI, see Section 4.5.3
ANN	Artificial Neural Network, see Section 3.5
AUC	Area under the curve, see Section 6.2.1
BPN	Back Propagation Network, see Section 3.5
CDF	Cumulative Density Function, see Section 6.1.3
CP-KNN	Conformal Prediction K-Nearest Neighbor, see Section 3.6
DoS	Denial of Service, see Section 2.2
DSL	Dynamic Local Search, see Section 5.4.3
DTSVM	Decision Tree Support Vector Machine, see Section 3.3
EDTSVM	Enhanced Decision Tree Support Vector Machine, see Section 3.3
eSOMs	emergent Self Organizing Maps see Section 3.5
FSM	Finite State Machine, see Section 3.5
GNP	Genetic Network Programming Section 3.4
GRASP	Greedy Randomized Adaptive Search Procedure, see Section 5.4.5
HIDS	Host-based Intrusion Detection System, see Section 2.3
IBL	Instance-Based Learning 3.6
IDS	Intrusion Detection System
KDD	Knowledge Discovery Database, see section 3.3
k-dist	Distance to the k nearest neighbour see Section 4.5.2
kNN	k nearest neighbours of o
KDE	Kernel Density Estimates see section 6.1.7
KDEOS	KDEOS- Generalized Outlier Detection with Flexible Kernel Density Estimates see section 4.5.8
pdf	Probability Density Function,
SNNs(o; p)	Shared nearest neighbours of o and p,
LB-ABOF	Lower Bound ABOF, see Section 4.5.4
LDOF	The Local Distance-base Outlier Factor, see Section 4.5.6
LDF	
LOF	Local Outlier Factor see Section 4.5.2
LOCI	Local Correlation Integral Method, see Section 4.5.3
LoOP	Local Outlier Probabilities Outlier Detection, see Section 4.5.5
(LQV)	Learning Vector Quantization algorithm, see Section 3.5
MAC	Media Access Controller, see Section 3.3
MDEF	Multi-granularity Deviation Factor, see Section 4.5.3
MintPts	The neighbourhood density is determined by the distance to nearest neighbour Section 4.2.4

NIDS	Network Based Intrusion Detection System
OPTICS	Identifying Local Outliers see Section 4.5.1
PLOF	Probabilistic Local Outlier Factor, see Section 4.5.5
RCL	Restricted Candidate List, see Section 5.6
ROC	The Receiving Operator Curve, see Section 5.2 and 6.2.1
SIDS	Stack based Intrusion Detection System see Section 2.3
SLS	Stochastic Local Search, see Section 5.4
SOD	Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data, see Section 3.4.7
SOM	Self-Organizing Map, see Section 3.5
SVM	Support Vector Machines See section 3.3

Chapter 1

Introduction

1.1 Background and Motivation

Outlier detection is the field that stems from the notion of the “Outlier”. There are many definitions in literature given such as: “an *observation which deviates so much from other observations as to arouse suspicions that is was generated by a different mechanism”* stated by Douglas M. Hawkins (Hawkins, 1980). Grubs (1969) has pointed out that “an outlying observation is one that appears to deviate markedly from other members of the sample in which it occurs”. Furthermore, other scientists have provided definitions for an outlier such as Barnett and Lewis (1994): “an observation (or a subset of observations) which appears to be inconsistent with the remainder of that set of data. Various researchers such as Johnson and Wichern (1992) have viewed an outlier from the perspective of the pattern and have defined it as “an unusual observation that does not seem to belong to pattern of variability produced by other observations. These aforementioned statements for defining outliers can help us in understanding in depth from different points of view.

In order for an activity to be regarded as an outlier it must: occur rarely in a small sample of the dataset, it must be inconsistent and it can be distinguished from the other events. The outlier detection field is the domain which develops mechanisms for capturing these outlying activities and determining the degree of outlier-ness of each data points. There are many approaches that are explored for dealing with the scope. One prominent issue that arises in this case is that a data point is considered an outlier because it does not belong to a mathematical theory or because it is just different from its neighbors. An outlier can be possibly associated with a data error generated by a mechanism or it can be related to an event that has not occurred accordingly. If this is found as an error, it can justify the decision of removing the specific data object which is suspected to constitute an outlier.

It is crucial to highlight in this research that it appears than not all the outliers are a bad event that must be eliminated from the dataset. It happens that they may be a significant point of change that can show us the way to any problem or activity that is not occurring in the right direction. They are the connecting dots and noteworthy hints which can contribute in having a clear view of what is happening. In addition, where a person can deem a noise, another one can regard it as information (Ng, 1990). Therefore, we cannot draw a clear line whether a data point is considered an outlier or not because it depends on the circumstances when the judgment is made (Blakeslee, 1990). There are times when it is suggested to get rid of outliers for various scopes just like cleaning the data and generating a better characteristic of the phenomenon. In these cases, capturing these outliers identifies some faults of the system that should be eradicated. Nevertheless, there are circumstances or models in which some outliers can be of benefit because they can lead to indicating that the model or the approach is mistaken and inappropriate for analyzing a certain behavior. Therefore, in this context, rather than removing the outlier it is urged to change the approach or the mathematical theory which is not suitable for a specific task. In the information security domain or intrusion detection system, capturing outliers is very meaningful because they indicate where the system has been breached and penetrated. Hence, the outliers help the system administrator to flag out the flaws of the system. When carrying out outlier detection procedures, it is recommended that rather than regarding outliers as wrong and rubbish data, it is better to think of them as attention-grabbing, appealing data where we can derive significant information for a specific model. It is of paramount importance to analyze thoroughly the outlying activities in order to get a better sense of how and why data is diverging from the normal behavior.

Data mining as a field, it concerns the disclosing of the exceptional activity behavior of particular data (Tang et al. 2002). The traditional schemes that have been published on this subject, have addressed the task of detecting outliers based on the underlying principles of statistical framework LOF, LOCI, LoOP, etc. The task of discovering outliers has been tackled and covered by a variety of approaches that can be divided into global and local outlier techniques. This feature has emerged from the type of databases being regarded as outlier of a particular data point. The difference here stands in the fact that global methods take into consideration the entire database while the local ones examine only some subset of the dataset or space.

The term “labeling” is associated to the global approaches while determining a score or “scoring” is linked to the local outlier detection approaches. The global methods are based on the assumption that for a data point it is possible to assign a binary property while the local ones rather than a binary property, they assign a level of being an outlier which is the outlier factor. This factor determines to what extent a data point can be considered an outlier according to its neighborhood system and how isolated a data point can be.

Kriegel et al (2009) have provided the difference for both types of classification:

- Global approaches possess a reference set which includes the entire database and all the data points. These are based on the underlying assumption that there is only one normal mechanism. However, they lack identifying other outliers in the reference set and may produce false outputs.
- Local approaches lie in a reference set that enclose a small region space of the data objects. Nevertheless, problems arise when choosing the proper reference set or the proposed subset of data.

One key point worth mentioning herein is that there exist some techniques that exhibit both global and local properties. Outlier detection is the field which handles the problems or issues of capturing the outliers in a set of data and discerning data points that deviate to the normal profile of data. Most of the approaches leveraged in the outlier detection area are based on the full dimensional Euclidian data space for investigating the characteristic of the data objects and discovering prominent outlier data point. One crucial characteristic of complex datasets in the real world is that they cannot specify density with global factors. The local factor is of paramount importance in order to indicate outliers in different regions of the set of data.

Another field in which this dissertation is based is the Intrusion Detection Systems. In our daily life, computer systems exhibit vulnerable security aspects which can render these systems susceptible to malevolent intrusions. In order to fix this lack of security, come into play the intrusion detection systems. Prevention techniques are not enough for safeguarding completely a computer platform. Due to the fact that new scenarios of breaching the system are continuously being developed, new techniques for capturing them must emerge. IDS makes possible to detect even novel intrusions and vulnerabilities by scrutinizing the behavior of networks. They are dubbed as the second line of the defense, given that the IDS starts its work after a threat has been divulged (Sommer and Paxson, 2010).

An intrusion detection system embodies an important responsibility in disclosing malicious events in a computer system or network of computers. Intrusion detection refers to the process of observing and scrutinizing entire activities that are happening in a computer system for identifying abnormal behavior occurring in a network. This field of expertise is experiencing a dramatic growth as more expert hackers are being intruding and penetrating the networks. In addition, this research growth goes proportionally to the increase of sensitive data being processed every day.

Regarding intrusion detection systems, they embody a combination between hardware and software frameworks which capture the intrusions in a computer networking system. IDS can examine all the activities that occur in a network by collecting information in different fields inside a network. Usually, IDSs develop two preliminary assumptions in regarding the data set utilized as incoming information: the amount of normal data profile that overcomes the abnormal data quantified in numbers and to what extent the attack deviates from the normal data in terms of qualitative aspect.

Intrusion detection systems (IDSs) encompass software or hardware systems, which are incorporated in a network in order to monitor the generated traffic and to flag out the compromised activities that might occur. An IDS employs techniques for modeling and analyzing intrusive behavior in a computer system (Sommer and Paxson, 2010). We must make a difference between IDS and a firewall. While firewall filters all traffic between the internal network and the unreliable external network, IDS merely monitors and sniffs the network traffic. They cannot drop or ban the network packets (Maheshwar and Singh, 2013). In other words, IDS’s property is to detect the network for any vulnerability. The IDS presents a second wall of defense and can be combined together with firewall to detect and prohibit suspicious computer activities entering and compromising the internal network.

Many algorithms of machine learning have been deployed for intrusion detection purposes. Machine learning methods (Deepika et al, 2012) enable systems to learn from experience. Usually the system starts with some prior corresponding knowledge that analyzes and tests the data acquired. Machine learning techniques rely on explicit or implicit model that accommodate the analyzed patterns in order to be categorized. They can be divided into Genetic Algorithms, Fuzzy Logic, Neural Networks and Bayesian networks (Kaur et al, 2013).

Anomaly and misuse detection systems have often borrowed their schemes from the machine learning perspective respectively: information theory, neural networks, association rules, classification approaches, instance based-learning algorithms, artificial immune systems and many more. Machine learning algorithms are popular because they address many real world problems. They are based on explicit and implicit models facilitating the analysis of the patterns to be classified. Machine learning tools, which compose anomaly based detection systems, have demonstrated to achieve a significant increase of detection rate. This is due to the intrusion detection area, which operates with particular properties, thus making the machine learning approach harder and onerously.

Rather than finding similarities, machine-learning techniques perform better at identifying activities that do not pertain to the existing ones. The classic machine learning approach consists of a classification problem. Understanding and providing insights of what the system is doing, is the best way to enhance the performance of anomaly detection schemes. In intrusion perspective, we can always discover a variation that performs slightly better in a specific context. However, in this domain insights are of much more benefit than numbers.

In this context, we have decided to investigate the first algorithm that tackles the local outliers LOF. Our motivation has emerged from intrusion detection system due to the fact that information leakage has experienced a dramatic growth these years. Therefore, novel and developed techniques for counter measuring these intrusions are required. Our motivation relies in the fact that we have spotted some flaws or drawbacks in the greedy ensemble algorithm. This heuristic approach results in finding the accurate output by producing the same result of outliers with different algorithms. While searching for the best candidate, the greedy ensemble algorithm can be isolated in the local maxima. Consequently, it can lose and not capture some outliers which may be of great deal of benefit in global perspective of the dataset. We have thought that studying in depth this problem would lead to a significant piece of research that could help and could be our novel contribution to the ensemble techniques for outlier detection tasks.

1.2 Aims and Objectives

In this dissertation, we make an effort to work in a field that lies from LOF until the latest outlier detection algorithms in our days. We aim to provide to the body of knowledge a state-of-the-art research work in the field of outlier detection methods from the first traditional efforts till the most well-known mechanisms investigated from the data mining research community. This dissertation aims to explore what has been done in the outlier detection area and to analyze the behavior of outliers in various circumstances and considering them as indicators in the performance of system.

After providing a sound theoretical framework and deepening our knowledge in this interesting field, this dissertation aims to narrow down in the density based approaches, which are represented from the local methods. We have chosen this sub-direction because it searches for outliers that can be found in a small sample of data and in small clusters. We are thrilled by this direction of research and we delve deeper into the family of the Local Outlier Factor algorithms. We have observed profoundly all LOF variants and looked for gaps and novel paths of research by finding issues that have not been tackled yet.

We have investigated in depth the greedy ensemble algorithm (Schubert, 2013). It is claimed from authors that this simple greedy ensemble appeared to perform pretty well. However, it lacks in terms of diversity because it always strives to find the best configuration of outlier detectors. Hence, we intend to fill this gap by optimizing the diversity property through randomization procedures.

Stochastic Local Search (SLS) are methods carried out in order to find solutions for the combinatorial optimization problems (Hoos & Stützle, 2005). The SLS algorithms can perform well on hard problems; in addition, they are simple and can be implemented smoothly in the variety of optimization problems. SLS are known for their robustness, thus making them very appealing for real-world tasks. We aim to focus specifically in Greedy Randomized Adaptive Search Procedure (GRASP) (Resende & Ribeiro, 1995) pertaining to the group of SLS because of its convenience in the implementing process. One disadvantage of the greedy approach is to obtain candidate solution with a low degree of diversity. GRASP attempts to alleviate this shortcoming by randomizing the construction method in order to yield a substantial variety of good starting points. If we compare the greedy heuristic approach with GRASP, we identify that the constructive phase of GRASP does not require the solution to be the best with the maximal properties. On the contrary, it picks in a random way one of the components, which has been considered to be the best of highly ranked solutions components.

The proposed framework is based on the combination of greedy ensemble algorithm with the GRASP in order to optimize and increase its level of diversity and detection rate. We had this idea due to the logic that best outlier results sometimes do not lead to best outcomes. It happens that while searching for the best result, it may find out just the local maximum data point, but not the global maximum of the whole dataset. Therefore, in order to escape the local maxima, we need to employ randomization techniques into the greedy ensemble approach, which exhibits greediness properties. By carrying out randomization techniques, we make sure that the search will continue in a random data point. We have conceptualized that this kind of methodology can lead to a substantial increase of diversity. The main goal of this dissertation is to come up with a novel ensemble algorithm that can increase the diversity of the outlier results. When using a heuristic greedy ensemble algorithm, mainly it is aimed to find the most accurate outliers; however, there are some algorithms that generate the same outlier results. They are not capable of capturing new outliers which can be very interesting to be analyzed and can lead to a better understanding of the observation. Therefore, we can compromise in the accuracy and boost the diversity by inducing a randomization process. Our scope is to enhance the detection rate by almost capturing all the outliers in a dataset by implementing this idea in a data mining environment.

We aim to undertake several scientific experiments in order to witness the optimization incorporated into the greedy ensemble algorithm with GRASP procedure. Our experiments begin from running the simple heuristic greedy ensemble till executing the novel randomized greedy ensemble algorithm by trying various outlier algorithms, matching and comparing the results. Finally, we provide significant evidence of a slight increase in performance and detection rate from our novel algorithm when comparing the previous approach. We have utilized ELKI Data Mining Framework, open source software based in JAVA, which is very appropriate for the scope of our research. ELKI is designed for comparing and evaluating various algorithms in terms of precision and recall and enables fair benchmarking of a variety of data mining algorithms. In concluding of our dissertation, we reasonably indicate some open issues that we intend to deal with in the future work.

1.3 Structure of the Dissertation

This dissertation is structured as follows:

- The first chapter is an introduction in the field divided into two crucial aspects respectively the background and motivation. The theoretical framework that our dissertation is founded from are data mining and machine learning algorithms in the field of outlier detection are of paramount importance and specifically the LOF (Local Outlier Factor) which is the direction that has been followed and narrowed down.
- Chapter 2 presents various intrusion detection systems. Intrusion detection systems are widespread in various domains just like fraud detection, healthcare; however, we have described all the types of intrusion detection system and highlighted the ones which are employed in the outlier detection domain.
- Chapter 3 consists of main algorithms for Machine learning for Intrusion Detection systems. It reviews these algorithms divided in four types of algorithm that are used for intrusion detections systems purposes. In different sections, we discuss these algorithms pertaining to the classification approaches, association rule mining, neural networks and instance based learning from the intrusion detection perspective.
- Chapter 4 presents outlier detection algorithms as a subfield of the machine learning for intrusion detection. After reviewing various directions of outlier detection in density based, distance-based and cluster based, we have been concentrated in the family of density based approaches for outlier detection with the most representation algorithm LOF Local Outlier Factor, we have reviewed the most substantial algorithm that have derived from this.
- Chapter 5 provides substantially the logic and the preliminary assumption of the methodological framework of this dissertation. It is thoroughly described the design of the novel algorithm that we have come up with and our main contribution to the body of knowledge related to the field of LOF and outlier detection. The combined pure greedy

ensemble algorithm with the GRASP randomizing procedure in ensemble of algorithm constitutes the foundation of our novel idea presented in this dissertation.

- Chapter 6 consists of evaluating experiments being conducted across the ELKI data mining platform. In these experiments, it is investigated the randomized ensemble and compared with the greedy heuristic approach. It is tested in terms of accuracy and diversity especially, because our aim is to increase the diversity by obtaining more various outliers.
- Chapter 7 draws significant conclusions of this research in terms of outliers and developing novel enhanced algorithms for outlier detection purposes which can be applied in various fields of usage: fraud detection, image recognition, financial forecast, weather forecast etc. Along with the conclusions, we have provided further future gaps in this field of ensemble algorithm that we were not able to address.

Chapter 2

Intrusion Detection Systems

2.1 Overview

Intrusion detection has experienced a dramatic attention with an extensive long history of research and scientific attempts for protecting networks. In this section, we explore and categorize intrusion detection systems. We provide a general description of the most prominent IDSs and their classification in terms of data resource and model of intrusion (Azad and Jha, 2013).

2.2 Intrusion Detection Systems

For computer systems, security is paramount important issue, and there can be some gaps which can render these systems vulnerable to intrusions. In order to fix these lacks of security come into play. Prevention techniques are not enough for safeguarding completely a computer platform. Due to the fact that new scenarios of breaching the systems are continuously being developed, new techniques for capturing them must emerge. IDS makes possible to detect even novel intrusions and vulnerabilities by scrutinizing the behavior of networks. They are dubbed as the second line of the defense, given that the IDS starts its work after a threat has been divulged (Sommer and Paxson, 2010).

An intrusion detection system embodies an important responsibility in disclosing malicious events in a computer system or network of computers. In order comprehend the role and functions of an intrusion detection system, it is of benefit to first specify the meaning of intrusion as a central keyword. An intrusion is a kind of malicious activity that attempts to compromise and breach the security aspects of a computer like:

-**data integrity** - this attribute consists of the scope of the data to be ensured during transmission from the source to the destination by not being modified.

-**data confidentially** – consists of making sure that while being transmitted the data can be accessible only to the authorized receiver (D’silva and Vora, 2013).

- **data availability** (D’silva and Vora, 2013) - when the network system is guaranteeing that the data is reachable by the authorized system based on the demand that the user is asking for the data.

With regard to the intrusion detection term, it refers to the process of observing and scrutinizing the entire activities that are happening in a computer system for identifying abnormal behavior occurring in a network. This field of expertise is experiencing a dramatic growth as more expert hackers are intruding and penetrating networks and goes proportionally to the increase of sensitive data being processed every day. Intrusion detection systems embody a combination between hardware and software frameworks which capture the intrusions in a computer networking system. IDS can examine all the activities that occur in a network by collecting information in different fields inside a network. Usually, an IDS develops two preliminary assumptions in regards to the data set being utilized as incoming information: the amount of normal data overcoming the abnormal data, and to what extent the attack deviates from the normal data.

The most commonly occurred intrusion threats work throughout the network by exploiting the network protocols in order to affect the target system. These connections are regarded as abnormal while the other connections in which can be utilized the network protocols are marked as normal connections and links. The attacks are classified into four main categories: **DoS-Denial of Service** when the hacker attempts to not consent the authorized users to access their service which are eligible for in a specific system (Djenouri et al, 2005), **Probe – Surveillance and probing-** in this attack, it is investigated the network in order to spot some well known vulnerabilities of the system on which the attacker intends to compromise. These attacks are of a very significant importance for the attacker who plans to attack in the future (Djenouri et al, 2005). **RCL-Remote to local**, it is refers to unauthorized hackers who can obtain local access to a specific system from a remote machine and then compromise the target machine and make use of its weaknesses. The last most developed attack is the **U2R-use to Root**, which aims to attack and obtain super user privileges. This happens when the specific machine is already under attack.

Intrusion detection systems (IDS) encompass software or hardware systems, which are incorporated in a network in order to monitor the generated traffic and to flag out the compromised activities that might occur. An IDS employs techniques for modeling and analyzing intrusive behavior in a computer system (Sommer and Paxson, 2010). We must make a difference between IDS and a firewall. While firewall filters all traffic between the internal network and the unreliable external network, IDS merely monitors and sniffs the network traffic. They cannot drop or ban the network packets (Maheshwar and Singh, 2013). In other words, IDS's property is to detect and to sniff the network for any network vulnerability. The IDS presents a second wall of defense and can be combined together with a firewall to detect and prohibit suspicious computer activities entering and compromising the internal network.

An IDS is composed of its core element known as the sensor (engine that analyzes data). The sensor retrieves data from three major sources directions: Own IDS knowledge base, Syslog and Audit. The ultimate goal of the sensor is to filter the information retrieved from data and to drop any irrelevant data obtained from the event set associated with the protected system. In other words, it warns the system when suspicious activity occurs.

The concept of an intrusion is a group of activities that aim to breach the data integrity, or make computer system vulnerable to fraudulent behaviors. An intrusion detection system is designed to safeguard all the network interactions in order to disclose well known attacks or even to discern for unknown threats that havened exists before. The major responsibility for IDS is to inform and flag out any susceptible behaviors to the system administrator.

In terms of technology used for detection and identification of the suspicious activity, IDSs are classified in signature-based and anomaly-based types. The both approaches have been vastly covered by various research scientists. Nevertheless the misuse detectors are found to be in the form of signature base system that has the objective to scan the traffic of a network for characteristic byte sequences. The major function of intrusion detection is to defend the network, and to examine intrusions among normal audit data (Maheshwar and Singh, 2013); this can be deemed as a classification problem. Below it is provided a graphical representation of the state-of-the-art of Intrusion Detection systems and approaches that it consists of.

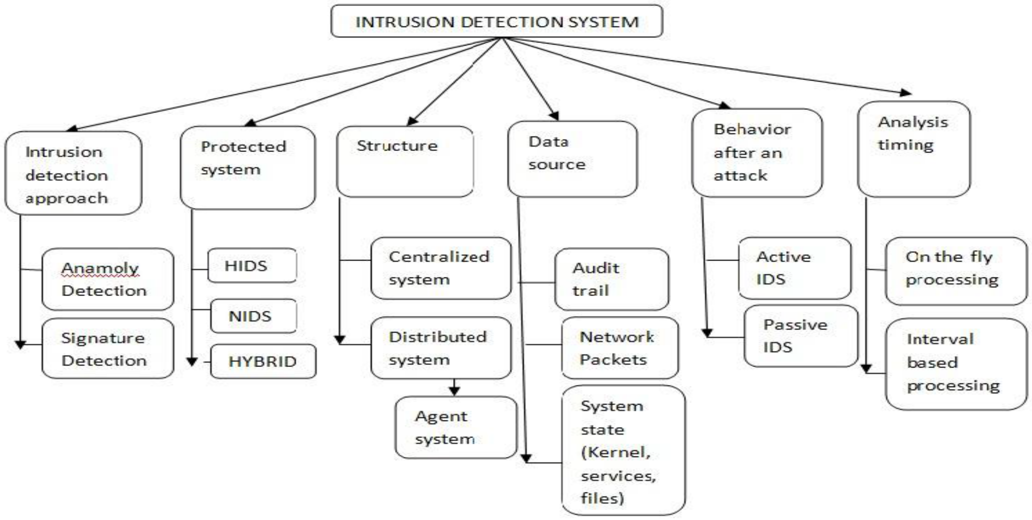


Figure 2.1: State-of-the-Art of Intrusion detection System (Azad and Jha, 2013)

One major objective of an intrusion detection system (IDS) is to disclose undesired activities with high precision and detection rate while providing low false positive results. It makes use of all the statistical data from the network. The intrusion detection architecture (Mitrokotsa et al, 2007) could be distributed or cooperative distributed and hierarchical. In distributive IDS a set of independent IDS agent create the IDS system. Every local IDS agent has two major elements (Mitrokotsa et al, 2007): the data collection, a component which selects the local audit data and the activity logs as well and intrusion detection engine which detects the local abnormal data by having a reference local audit data. This scenario functions as follows: If the Intrusion Detection Engine finds an intrusion then this leads to the activation of the Response Engine. The *Intrusion Response Engine* is in charge of sending local and global alarm in order to notify its surrounding node and its global nodes.

A distributed IDS collects the data and examines it in multiple hosts, in contrast with the centralized IDS where the data is gathered and investigated in a centralized fashion. The two strategies of collecting the data may apply to host-based and network-based intrusion detection as well or it can incorporate a combination of the two methods. IDS can provide reaction in two distinct manners: Active where some actions are developed some actions for reaction to the intruding event. This can be a logging of service, disruption of the connection, etc. Regarding the passive reaction, it can produce alarms and warnings. It carries out audit information and examination in IDS in the real time and in the interval based time. The real time intrusion detection systems executes frequently for detecting intruders and generates outcomes in real time which is related to the fact that the IDS can respond very quickly to the intruding activity. Besides the real-time IDSs, security researchers have developed periodical IDS, which consist of executing periodical intrusions occurring on specific intervals.

One of key challenge an IDS has to overcome is the fact of being overloaded with data packets. In wireless networks, it is of paramount importance to choose the features that can best specify the conduct of the network. When it comes to wireless networks, it is composed of many nodes, which can interact with other nodes throughout messages. Hence the node's performance can be generated by observing the traffic of the network. Every node is comprised as an elementary unit which monitors its surrounding neighbors and constructs a profile along the offline training. Consecutively, this built profile is utilized as a threshold to identify anomalous behavior in the whole network. This scenario is appropriate for small networks.

An IDS in order to accomplish it goals employs one or more detection engines (Snap et al, 1999. These engines specify whether the target event matches an abnormal or a normal profile activity. These detection engine frameworks can be categorized as follows (Panos, 2010):

a) signature-based systems function based on predefined set of patterns in order to capture threats

b) specification-based engines which depend on a set of constraints that may be a specification of precise steps of operation of a particular program and aim to observe the carrying out of the aforementioned programs concerning these restriction

c) anomaly-based engines which incorporate specific models of data just like normal profile of data of the nodes' conduct and flag out those nodes which experience a substantial deviation from these patterns.

A considerable number of IDS literature and research works that have been elaborated are focused on the anomaly based detection because of its benefit that it can impose regarding the attacks, which appear to be novel and not tackled before. This advantage is associated by a substantial number of challenges just like the high false alarm rates, the requirement of the adaptability into the dynamic networks circumstances. There is a significant piece of work which has aimed to tackle these restrictions by incorporating novel techniques and strategies (Xenakis et al, 2011) (Sun et al, 2007) (Mishra et al, 2004.) (Azer et al 2005) (Li, and Wei, 2004).

2.3 Types of IDS

There are some various approaches for implementing intrusion detection systems. Below we provide a thorough review of the most compelling techniques area of Intrusion Detection System.

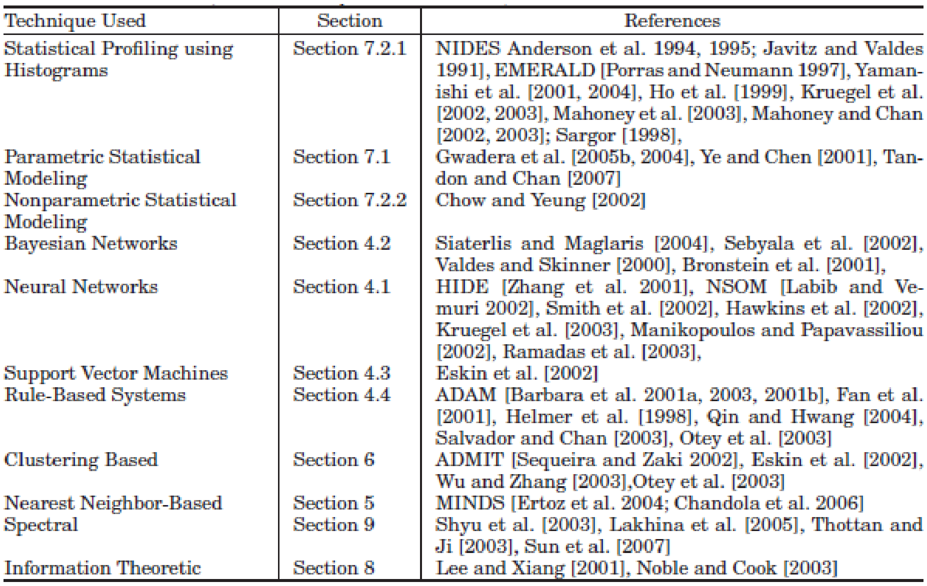
2.3.1 Stack based Intrusion Detection System (SIDS)

Intrusion detection systems based on the stack of the protocols are responsible for integrating precisely with TCP/IP protocol stack, which enables the system to observe packets as they are distributed through the OSI layers. IDSs pull the packet when they identify any compromised behavior before the application can process the packet.

2.3.2 Network Based Intrusion Detection System

The Network-based Intrusion Detection systems (NIDS) are the most traditional approaches and they are classified based on the style of the detection that they carry out. These kinds of intrusion detection systems monitor the network packets and detect network attacks. NIDS listen to the packets in a segment of the network allowing them to detect distributed attacks. NIDS monitor traffic of the network; it consists of sensors to detect packets, a data analyzer to make sense of data. It generates alarms when it encounters suspicious activity. These systems cope with detecting intrusions in the networks (Kaur and Singh, 2015). The threats are divulged through particular methods that model the data and detect anomalous sequences that can constitute a set of anomalies.

Its major drawback is not being aware of the behavior in the internal environment (Gwadera et al. 2005b, 2004). The cause of these abnormal data can come from outside where hackers need network information in order to breach the networks. The generated data can change in terms of granularity, because it can be produced from packet level, routers etc. The data to be examined from a network-based intrusion detection system is high dimensional typically with a mix of continuous attributes (Chandola et al, 2009). One challenge that these systems have to overcome is the nature of anomalies that continues to change and be altered during time. Hence, they have to adapt to the networks in order to escape the actual intrusion detection system.



Technique Used	Section	References
Statistical Profiling using Histograms	Section 7.2.1	NIDES Anderson et al. 1994, 1995; Javitz and Valdes 1991], EMERALD [Porras and Neumann 1997], Yamanishi et al. [2001, 2004], Ho et al. [1999], Kruegel et al. [2002, 2003], Mahoney et al. [2003], Mahoney and Chan [2002, 2003]; Sargor [1998],
Parametric Statistical Modeling	Section 7.1	Gwadera et al. [2005b, 2004], Ye and Chen [2001], Tandon and Chan [2007]
Nonparametric Statistical Modeling	Section 7.2.2	Chow and Yeung [2002]
Bayesian Networks	Section 4.2	Siaterlis and Maglaris [2004], Sebyala et al. [2002], Valdes and Skinner [2000], Bronstein et al. [2001],
Neural Networks	Section 4.1	HIDE [Zhang et al. 2001], NSOM [Labib and Vemuri 2002], Smith et al. [2002], Hawkins et al. [2002], Kruegel et al. [2003], Manikopoulos and Papavassiliou [2002], Ramadas et al. [2003],
Support Vector Machines	Section 4.3	Eskin et al. [2002]
Rule-Based Systems	Section 4.4	ADAM [Barbara et al. 2001a, 2003, 2001b], Fan et al. [2001], Helmer et al. [1998], Qin and Hwang [2004], Salvador and Chan [2003], Otey et al. [2003]
Clustering Based	Section 6	ADMIT [Sequeira and Zaki 2002], Eskin et al. [2002], Wu and Zhang [2003],Otey et al. [2003]
Nearest Neighbor-Based	Section 5	MINDS [Ertoz et al. 2004; Chandola et al. 2006]
Spectral	Section 9	Shyu et al. [2003], Lakhina et al. [2005], Thottan and Ji [2003], Sun et al. [2007]
Information Theoretic	Section 8	Lee and Xiang [2001], Noble and Cook [2003]

Table 2.1: Examples of network-based intrusion detection

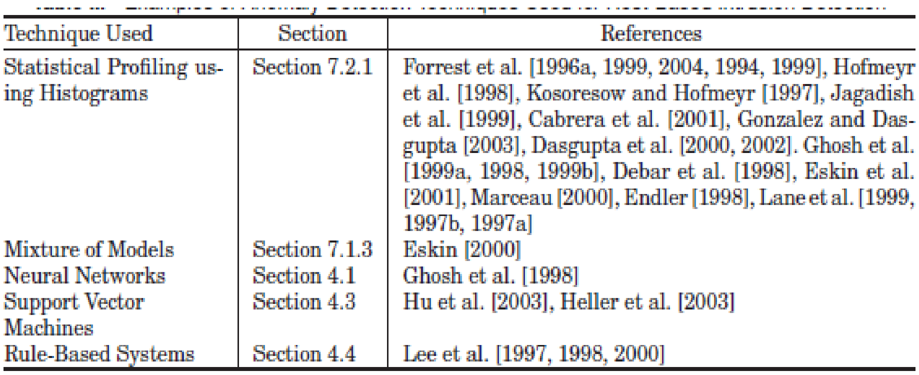
2.3.3 Host-based Intrusion Detection System

The Host-based Intrusion Detection Systems (HIDS) detect intrusive activities and malevolent behavior on the host. It controls the privileged access of the host in order to monitor components of host that are not accessible to other systems. It monitors internal and external activity from the computer. It is viewed as an individual device; this approach is not aware of what is happening in the whole picture of network environment. Its major disadvantage is that HIDS cannot detect attacks targeted to the host that does not have HIDS installed.

With the development of Internet, hackers are developing novel malware every day. While new malware is developed, IDSs are also evolving and becoming more sophisticated. They need to be a step ahead compared to hackers, with regard to improving and being able to detect the dramatic growth of attacks. IDSs detect fraudulent behavior by looking for known weaknesses and known attack patters (signature-based) or normal behavior (anomaly-based). Therefore, it is difficult to indicate all potential attacks because IDSs need to know all probable attacks in order to achieve a satisfying protection. Whenever IDSs encounters novel attacks, IDS manufacturers develop rules and signature for that novel threat. In practice, some manufactures may not be aware of new attacks distributed in the network. Another pitfall of IDS is the management of its sensitivity. Many IDSs generate high false alarm rates, in other word they flag out many intrusion alerts, which do not consist of any malevolent behavior. This leads to hurdles on handling and inspecting alerts from security professionals. On the other hand, if we lower the sensitivity of IDS we might encounter the problem of missing attacks of paramount importance, which can lead to insecure networks and hosts. A major challenge in this case is to determine the optimal sensitivity of IDS.

In the most of host-based Intrusion Detection System various type of data are gathered for different activities, just like operating system log, the traffic of the network or other behavior and action that the network conducts. These data may have an unseal profile because of the malicious code that they are composed of. Host-based intrusion application generally generates discrete data because of the numerous events which diagnose various attacks and behaviors are extracted from a collection of discrete instance (Singh et al, 2009). Below in the table 2.2, it is provided a summary of some compelling research works, which have employed the host based intrusion detection scenario.

The host based intrusion detection systems handle the operating system call traces (Chandola, 2009). The deviated patterns are translated into malevolent behaviors. The key factor for making the distinct between normal and abnormal activity is the occurrence or the frequency of the events. In the host-based method, all the activities pertain to the same alphabet; the factor of occurrence that determines one data point from another. The operating system calls can be produced by programs (Hofmeyr et al, 1998) or by users system (Lane and Brodley, 1999). The alphabet is composed of 183 system calls (Chandola, 2009). Various programs execute various system calls in specific sequences. The data can be categorized in various levels just like program level and user level. In the host based intrusion detection system, it is entailed to deal with the sequential nature of data. The broad spectrum of these techniques has to model data sequence or to calculate similarities between patterns.



Technique Used	Section	References
Statistical Profiling using Histograms	Section 7.2.1	Forrest et al. [1996a, 1999, 2004, 1994, 1999], Hofmeyr et al. [1998], Kosoresow and Hofmeyr [1997], Jagadish et al. [1999], Cabrera et al. [2001], Gonzalez and Dasgupta [2003], Dasgupta et al. [2000, 2002]. Ghosh et al. [1999a, 1998, 1999b], Debar et al. [1998], Eskin et al. [2001], Marceau [2000], Endler [1998], Lane et al. [1999, 1997b, 1997a]
Mixture of Models	Section 7.1.3	Eskin [2000]
Neural Networks	Section 4.1	Ghosh et al. [1998]
Support Vector Machines	Section 4.3	Hu et al. [2003], Heller et al. [2003]
Rule-Based Systems	Section 4.4	Lee et al. [1997, 1998, 2000]

Table 2.2: Examples of host based techniques for intrusion detection

2.4.4 Anomaly Based Intrusion Detection System

Anomaly Based Intrusion Detection Systems (Panos et al, 2011) operates on the concept that the attack behavior diverges from the normal profile behavior. Firstly, it identifies the baseline of normal profile, and then the new event is compared with the normal behavior. If the new activity deviates from the normal profile then it is regarded as anomalous by generating an alarm. The variations between the normal profile and the monitoring feature are analyzed by making use of various techniques like statistical analysis, machine learning and data mining techniques. Anomaly based IDS systems suffer from high rates of false positive alarms and can introduce heavy processing overheads on computation resources. Their main advantage is the detection of unknown attacks. Anomaly based detection has to be adaptive to be able to face the dynamic change of the network. Their normal profile should represent the normal network operation. The dynamic change must be incorporated immediately into the normal profile. Most current IDSs employ both techniques to achieve better detection capability.

Anomaly detection is responsible for discovering patterns in a specific dataset that do not follow the characteristic of a normal profile of data. The patterns are called anomalies or dubbed outliers, and constitute vulnerabilities to several areas. The anomaly detection approaches are categorized in many approaches according to the availability of the predefined data (Tan et al, 2005). Many of the prior systems employ signature based methods to identify unknown and exceptional behave or, for the reason that they generate a very small number of the false positives rates when compared to the anomaly based models. However, the anomaly based techniques are more preferred because these incorporate the ability to capture zero-day attacks or in few words novel attacks that have been known in existing methods. From the data mining perspective, the anomaly base detection are categorized in three major groups which are based on the assumption that a domain expert has to label the data in normal or abnormal profile or to design an anomaly detection model.

One key challenge for the anomaly based detection systems is the huge volume of data and a trend of an exponential increase of the sensitive data (Panos et al, 2010). These techniques must evolve and be computationally effective in order to deal with increase of data size. Another concern to be discussed thoroughly is the high false alarm detection rate which is due to the large sizes of input. Marked data corresponds to the normal behavior available, while there are no disposed labels for intrusion. This is due to the fact that data points constitute a low percentage of false alarms; this can make the investigation very compelling for data analysts and professional security engineers. Therefore, semi supervised and unsupervised techniques are preferred in this area.

First, we need to determine the meaning of an anomaly and provide details for this term, which is of fundamental importance to disclosing the properties of anomaly detection systems. There have been explored a vast definitions for anomaly (D'silva and Vora, 2013), (Grubbs (1969) as an outlying observation. Outlier is one that appears to deviate markedly from other members of the sample in which it occurs. Another definition is provided from Barnett and Lewis (1994) according to them it is an observation (or a subset of observations) which appears to be inconsistent with the remainder of that set of data. John (1995) have stated that "an outlier can be also considered as a surprising veridical data, a situation in which a point otherwise belonging to class A, but in actual is placed in class B, thereby making the true (veridical) classification of that point surprising to the observer. Due to Aggarwal and Yu (2001) outliers may be regarded as noise points which lie outside a set of undefined clusters or alternatively outliers may be determined as the points that lie in the outer space of the clusters set. Others definition and attempts to define the anomaly as concepts has been developed from Chandola et al. (2009), which have considered them as patterns in data that do not conform to well defined notion of the normal behaviors.

Savage (2014) has pointed out that anomalies are determined as regions of the network whose structure differs from the expected normal model. An anomaly is described as unusual behavior which showcases a distinguished activity compared to the others which are taking place at the same environment. It can be interchangeably used instead of anomaly the term outlier or abnormality or exception. One can find some misconceptions regarding the concept of anomaly when reading various literature and definitions about anomalies.

In the Aggarwal and Yu (2001), it is advocated that anomalies distinguish from noisy data which is regarded as a random error and has no relevance with the process of analyzing the data. The noisy data can be found in the fraud detection of credit cards, the noisy data can be referred to one person's purchase actions. Therefore, these noisy data cannot be considered as anomalies because it would not be cost effective to examine every interactions and activity of individuals. In contrary, it is applied the removal of the noise before conducting the anomaly detection process.

Anomaly detection has been considered closely with novelty detection, which is concerned with the detection of novel behaviors that have not been addressed before. The anomalies which are occurring in the data can lead to substantial concerns which need to be addressed specifically and accurately. There are extensive examples that illustrate the fundamental importance just like some attackers may build a set of false identities and exploit them to interact with random users

(Mahoney and Chan, 2003). These anomalies are very sensitive because if we tackle them, it may be a tremendous loss of data. However, in our daily life, the detection strategy and the scenario of being divulged together with its countermeasure are experiencing a significant rise.

In the Fig. 2.2, there are depicted anomalies in a two dimensional set of data. The data is divided into two clusters thereby two normal profile data region N1 and N2. The majority of the data points are positioned into these two normal spaces and hence the data objects, which are located significantly far away from these normal data regions, are reported as anomalies behavior. There are many scopes and intentions for introducing anomalies data for example it may be for credit card fraud purposes, cyber intrusion, hactivism or crashing and denying the service to a system. These reasons make sense to the security analysts and they draw a profound attention to data scientists, as well.

Anomaly detection has some similarities, but it is differs from the noise removal (Teng et al, 1990) and noise accommodation in (Rousseeuw and Leroy 1987), both techniques tackle the noisy data. On the other hand, the noise can be described (Chandola et al, 2009) as an occurrence that is out of interest to the data scientists, however, it exhibits a burden for discovering the real outliers. The noise removal aims to remove and get rid of the entire set of unwanted data points before any data examination has been conduct. Regarding, noise accommodation, it specifies the immunizing a statistical model of estimation versus the anomalies monitoring (Huber, 1974). One other direction relevant to the area of the anomaly detection is deemed the novelty detection.

Markou and Singh (2003a, 2003b), Saunders and Gero (2000), rely on dealing with novel previously unobserved anomalies. The difference among novel patterns and anomalies is that the novel patters are integrated into the model of normal data after the process that has detected them.

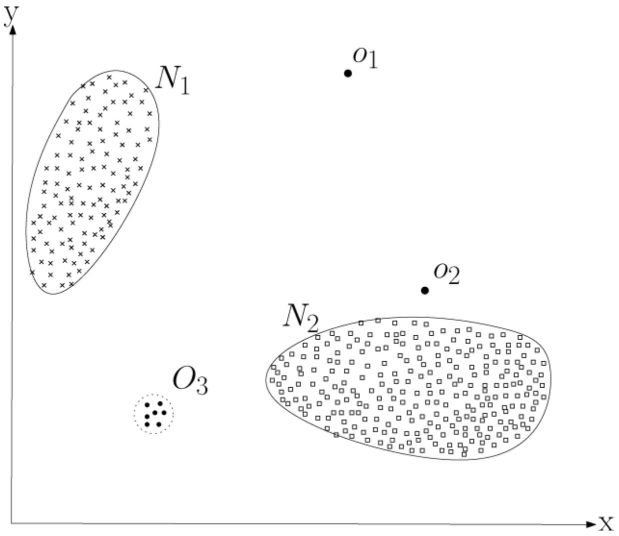


Figure 2.2: A demonstration of data set divided into two clusters and the presence of anomalies

2.3.4.1 Related work of anomaly detection approaches

There is a large body of work focused on anomaly detection for symbolic sequences (Forrest, et al, 1999), (Gao et al, 2002), (Gonzalez and Dasgupta, 2003), (Hofmeyr et al, 1998) (Lee and Stolfo, 1998), (Lee, et al, 1997), (Michael and Ghosh, 2000), and (Sun et al, 2006) (Budalakoti et al, 2006). Their implementations have been in various applications and fields. One of earliest methods was introduced in (Forrest et al, 1996), which exploited lookahead pairs in order to define anomalous sequence of system calls. Another piece of work is presented in (Hofmeyr, 1998) which proposed a window based techniques STIDE and demonstrated that STIDE can have more advantages over the lookahead pairs based techniques in terms of performance in operating systems domain. Extensive techniques and approaches have been introduced for identifying anomalies in the system data call using Hidden Markov Models (Qiao et al, 2002). Some others have utilized Finite State Automata (FSA) (Michael and Ghosh, 2000) and classification mechanisms, as well RIPPER (Lee and Stolfo, 1998), (Lee et al, 1997).

All these research works have demonstrated the applicability of specific models focused only on the system call and some other have conducted a comparative analysis with the STIDE approach. Even though, there exist many reviews and comparative analysis of the traditional schemes for anomaly detection (Hodge and Austin, 2004), they are not focused on the sequence anomaly detection. In (Forrest et al, 1999), it is presented a comparative evaluation of models on the system call intrusion detection dataset comparing from prominent techniques respectively: STUDE, t-STIVE, HMM based and RIPPER. Markovian techniques designate a probabilistic anomaly score to each event incorporating modeling techniques such Finite State Automata (FSA), Hidden Markov Models and Probabilistic Suffix Trees (PST).

Figure 2.3: The structure of relevant fields in which anomaly detection is applied