DISTANCE BASED ANALYSIS FOR DETECTION OF INTRUSIONS AND ANOMALIES

SHREEKANTH S
Assoc.Prof, CSE, GNITC, JNTUH, & Research Scholar in JNTU Hyderabad, INDIA,
e-mail:sreekanth.sreerrama@gmail.com

B. SAMIRANA ACHARYA
Asst.Prof, CSE, GNITC, JNTUH, Hyderabad, INDIA,e-mail:acharya501@gmail.com

B.NANDAN
Assoc.Prof, CSE, GNITC, JNTUH e-mail:bnandan@g[illegible]a.org

ABSTRACT :

An Intrusion Detection System (IDS) is one of these layers of defense against malicious attacks. In IDS a stream of data is inspected and rules are applied in order to determine whether some attack is taking place. Intrusion Detection Systems typically operate within a managed network between a firewall and internal network elements. This paper discusses the intrusion detection using data mining techniques such as classification, association and clustering. In this paper we mainly focused on clustering techniques and outlier detection.

KEYWORDS: IDS, classification, association and clustering, Outliers detection.

INTRODUCTION:

The word intrusion means t[illegible]

1. The act of intruding or the condition of being intruded on.

2. An inappropriate or unwelcome addition.

3. Law. Illegal entry upon or appropriation of the property of another.

Where intrude means that: To put or force in inappropriately, especially without invitation, fitness, or permission.

INTRUSION:

Based upon the above definitions, intrusion in the terms of information can be defined as when a user of information tries to access such information for which he/she is not authorized, the person is called intruder and the process is called intrusion.

1.1 INTRUSION DETECTION:

It is observed by Jones, Anita K., Sielken, Robert S(1999) that Intrusion detection is the process of determining an intrusion into a system by the observation of the information available about the state of the system and monitoring the user activities. Detection of break-ins or attempts by intruders to gain unauthorized access of the system is intrusion detection.

The intruders may be an entity from outside or may be an inside user of the system trying to access unauthorized information. Based upon this observation intruders can be widely divided into two categories; external intruders and internal intruders.

- External intruders are those who don't have an authorized access to the system they are dealing with.
- Internal intruders are those who have limited authorized access to the systems and they overstep their legitimate access rights. Internal users can be further divided into two categories; masqueraders and clandestine users.
- Masqueraders are those who use the identification and authorization of other legitimate users.
- Clandestine users are those who successfully evade audit and monitoring measures.

1.2. INTRUSION DETECTION SYSTEM:

Although intrusion detection technology is immature and should not be considered as a complete defense, but at the same time it can play a significant role in overall security architecture. If an organization chooses to deploy an IDS, a range of commercial and public domain products are available that offer varying deployment costs and potential to be effective. Because any deployment will incur ongoing operation and maintenance costs, the organization should consider the full IDS life cycle before making its choice. When an IDS is properly deployed, it can provide warnings indicating that a system is under attack, even if the system is not vulnerable to the specific attack. These warnings can help users alter their installation's defensive posture to increase resistance to attack. In addition, an IDS can serve to confirm secure configuration and operation of other security mechanisms such as firewalls. Within its limitations, it is useful as one portion of a defensive posture, but should not be relied upon as a sole means of protection. As e-commerce sites become attractive targets and the emphasis turns from break-ins to denials of service, the situation will likely worsen.

Intrusion detection with snort (2003) comprises that An intrusion detection system or IDS is any hardware, software or combination of both that monitors a system or network of systems for a security violation. Bace said that An IDS is often compared with a burglar alarm system. Just like a burglar alarm system monitors for any intrusion or malicious activity in a building facility, IDS keeps an eye on intruders in a computer or network of computers.

Figure 1 displays a generic intrusion detection system. From the audit data source the information goes to the pattern matching module for misuse detection and a profile engine to compare current profile with the normal behavior defined for the system. Pattern matching module interacts with policy rules to look for any signature defined in the policy. An anomaly detector distinguishes an abnormal behavior using the profile engine.

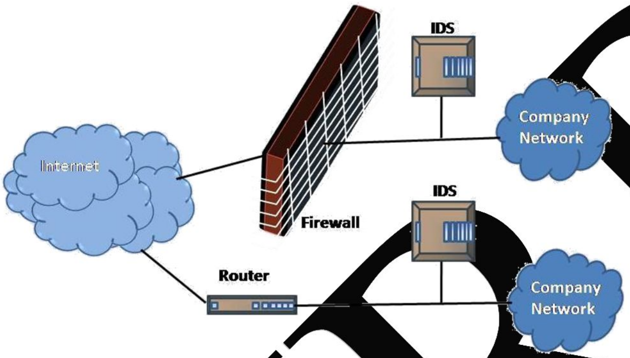

Figure1: Typical locations for an intrusion detection system

There are two types of intrusion detection systems according to the architecture. One which are implemented on the system they are monitoring and others which are implemented separately. This separate implementation has several advantages over the other approach.

- It keeps a successful intruder from disabling the intrusion detection system by deleting or modifying the audit records on which the system is based.
- It lessens the load associated with running the intrusion detection system on the monitored system.

The only disadvantage with this scheme is that it requires secure communication between the monitoring and monitored system.

IDPSes typically record information related to observed events, notify security administrators of important observed events and produce reports. Many IDPSes can also respond to a detected threat by attempting to prevent it from succeeding. They use several response techniques, which involve the IDPS stopping the attack itself, changing the security environment (e.g. Configuring a firewall) or changing the attack's content.

AN ANOMALYASED INTRUSION DETECTION SYSTEM:

In "A strict anomaly detection model for IDS" Sasha/Beetle said that An Anomaly-Based Intrusion Detection System is a system for detecting computer intrusions and misuse by monitoring system activity and classifying it as either normal or anomalous. The classification is based on heuristics or rules, rather than patterns or signatures, and attempts to detect any type of misuse that falls out of normal system operation. This is as opposed to signature based systems which can only detect attacks for which a signature has previously been created.

In order to determine what attack traffic is, the system must be taught to recognize normal system activity. This can be accomplished in several ways, most often with artificial intelligence type techniques. Systems using neural networks have been used to great effect. Another method is to define what normal usage of the system comprises using a strict mathematical model, and flag any deviation from this as an attack. This is known as strict anomaly detection.

Anomaly-based Intrusion Detection does have some short-comings, namely a high false positive rate and the ability to be fooled by a correctly delivered attack.

2.1 ANOMALY DETECTION:

Anomaly detection consists of first establishing the normal behavior profiles for users, programs, or other resources of interest in a system, and observing the actual activities as reported in the audit data to ultimately detect any significant deviations from these profiles. Most anomaly detection approaches are statistical in nature. a user's normal profile consists of a set of statistical measures. The measures used in NIDES are of the following types:

- Ordinal measure: A count of some numerically quantifiable aspect of observed behavior. For example, the amount of CPU time used and the number of audit records produced;
- Categorical measure: A function of observed behavior over a finite set of categories. Its value is determined by its frequency relative to other categories.

It can be further classified as:

i) Binary categorical measure: Whether the category of behavior is present (i.e., 0 or 1). This type of measure is sensitive in detecting infrequently used categories, such as changing one's password;

ii)Linear categorical measure: A score function that counts the number of times each category of behavior occurs. For example, command usage is a linear categorical measure, where the categories span all the available command names for that system.

To compute the deviations from the profile, IDES and NIDES use a weighted combining function to sum up the abnormality values of the measures. The profiles are also updated periodically (i.e., aged) based on the (new) observed user behavior to account for normal shifts in user behavior (for example, when a conference deadline approaches).

Anomaly detection systems can detect unknown intrusion since they require no a priori knowledge about specific intrusions. Statistical-based approaches also have the added advantage of being adaptive to evolving user and system behavior since updating the statistical measures is relatively easy. However, anomaly detection systems also have major shortcomings:

• The selection of the right set of system (usage) features to be measured can vary greatly among different computing environments;

• The fine tuning of the deviation threshold is very ad hoc;

• User behavior can change dynamically and can be very inconsistent;

• Some intrusions can only be detected by studying the sequential interrelation between events because each event alone can appear to be normal according to the statistical measures.

• A statistical-based system can be trained, over some period of time, by a deliberate intruder to gradually update the user profile to accept his intrusive activities as normal behavior!

2.2 MISUSE DETECTION:

Misuse detection consists of first recording and representing the specific patterns of intrusions that exploit known system vulnerabilities or violate system security policies, then monitoring current activities for such patterns, and reporting the matches. There are several approaches in misuse detection. They differ in the representation as well as the matching algorithms employed to detect the intrusion patterns. Some systems, for example NIDES [Lunt, 1993], use a rule-based expert system component for misuse detection. These systems encode known system vulnerabilities and attack scenarios, as well as intuitions about suspicious behavior, into rules. For example, one such rule is: more than three consecutive unsuccessful logins within five minutes is a penetration attempt. Audit data is matched against the rule conditions to determine whether the activities constitute intrusions. Another system, STAT [Il gun et al., 1995], uses state transition analysis for misuse detection. It represents and detects known penetration scenarios using state transition diagrams. The intuition behind this approach is that any penetration is essentially a sequence of actions that leads the target system from an initial normal state to a compromised state. Here a state in the state transition diagram is a list of assertions in terms of system attributes and user privileges. A transition is labeled by a user action (i.e., the signature action), for example, the acquisition of previously unheld privileges. Intrusions are detected in STAT when a final compromised state in the state transition diagram is reached.

2.3 PROBLEMS WITH CURRENT INTRUSION DETECTION SYSTEMS:

We measure the quality of IDS by its effectiveness, adaptability and extensibility. An IDS is effective if it has both high intrusion detection (i.e., true positive) rate and low false alarm (i.e., false positive) rate. It is adaptable if it can detect slight variations of the known intrusions, and can be quickly updated to detect new intrusions soon after they are invented. It is extensible if it can incorporate new detection modules or can be customized according to network system configurations. Current IDSs lack effectiveness. The hand-crafted rules and patterns, and the statistical measures on selected system measures are the codified "expert knowledge" in security, system design, and the particular intrusion detection approaches in use. Expert knowledge is usually incomplete and imprecise due to the complexities of the network systems. Current IDSs also lack adaptability. Experts tend to focus on analyzing "current" intrusion methods and system vulnerabilities. As a result, IDSs may not be able to detect "unknown" attacks. Developing and incorporating new detection modules is slow because of the inherent "learning curve". Current IDSs lack extensibility. Reuse or customization of IDS in a new computing environment is difficult because the expert rules and statistical measures are usually ad hoc and environment-specific. Since most current intrusion detection systems are monolithic, it is also hard to add new and complementary detection modules to existing IDS. Some of the recent research and commercial IDSs have started to provide built-in mechanisms for customization and extension. For example, both Bro [Paxson, 1998] and NFR [Network Flight Recorder Inc., 1997] filter network traffic streams into a series of events, and execute scripts, e.g., Bro policy scripts and NFR's N-Codes, that contain site-specific event handlers, i.e., intrusion detection and handling rules. The system administration personnel at each installation site must then assume the roles of both security experts and IDS builders because they are responsible for writing the correct event handling functions. Our first-hand experience with both Bro and NFR show that while these systems provide great flexibility, writing the scripts involves a lot of effort, in addition to learning the scripting languages. For example, there is no means to "debug" the scripts. These systems also handle a fixed set of network traffic event types. On a few occasions we were forced to make changes to the source code of the original IDS to handle new event types. We can attribute, to a very large extent, the poor qualities of current IDSs to the manual, ad hoc, and purely knowledge engineering development process. Given the complexities of network systems, and the huge amount of audit data generated by user and system activities, we need a more systematic and automatic approach to building IDSs.

2.4. APPLICATION OF DATA MINING IN INTRUSION DETECTION:

The goal of intrusion detection is to detect security violations in information systems. Intrusion detection is a passive approach to security as it monitors information systems and raises alarms when security violations are detected. Examples of security violations include the abuse of privileges or the use of attacks to exploit software or protocol vulnerabilities.

The applications of data mining hence are immense. Some common applications of data mining in business include the following

• Data mining concepts are in use for the Sales and marketing to provide better customer
Service, to improve cross-selling opportunities, to increase direct mail response rates.

• Customer Retention in the form of identification of patterns of defection and prediction of likely defections is possible through data mining.

• Risk Assessment and Fraud area also uses the data-mining concept for identifying the inappropriate or unusual behavior etc.

DATA MINING AND INTRUSION DETECTION:

ID using Data Mining (IDDM), use as basis the audited data from different sources , activity indexes (from normal and intrusion activity) and algorithms to search significant patterns; enabling the construction of misuse and anomaly detection models based on an intelligible set of rules. The raw data is archived and sampled in discrete records according to the attributes. Data mining programs are subsequently used over the traffic records to compute patterns. The connections and the patterns are then analyzed to construct additional features, getting an empirical and iterative approach. One of the most critical and success determining selections is the related with the data mining technique:

3.1. CLASSIFICATION : L. Kaufman and P. Rousseeuw said that Classification categorizes the data records (training data set) in a predetermined set of classes (Data Classes) used as attribute to label each record; distinguishing elements belonging to the normal or abnormal class (a specific kind of intrusion), using decision trees or rules. This technique has been popular to detect individual attacks but has to be applied with complementary fine-tuning techniques to reduce its demonstrated high false positives rate. With support tools as RIPER (a classification rule learning program) and using a preliminary set of intrusion features, accurate rules and temporal statistical indexes can be generated to recognize anomalous activity. They have to be inspected, edited and included in the desired model (frequently misuse models).

3.2. ASSOCIATION RULES: Associations of system features finding unseen and / or unexpected attribute correlations within data records of a data set, as a basis for behavior profiles.

3.3. CLUSTERING: discovers complex intrusions occurred over extended periods of time and different spaces, correlating independent network events. The sets of data belonging to the cluster (attack or normal activity profile) are modeled according to pre-defined metrics and their common features. It is especially efficient to detect hybrids of attack in the cluster, showing high performance when are processed features computationally expensive. With other techniques is able to re-train itself reclassifying the existing clusters and generating new ones.

3.3.1. EVALUATION OF CLUSTERING

An objective function is used for evaluation of clustering methods. The choice of the function depends upon the application, and there is no universal solution of which measure should be used. Commonly used a basic objective function is defined as:

$$f(P,C)=\sum_{i=1}^{n} d(x_i,c_{pi})^2$$

$$f(P,C)=\sum_{i=1}^{n} d(x_i,c_{pi})^2$$

Where P is partition and C is the cluster representatives, d is a distance function. The Euclidean distance and Manhattan distance are well-known methods for distance measurement, which are used in clustering context. Euclidean distance is expressed as:

$$\sqrt{\sum_{i=1}^{k}(x_1^i - x_2^i)^2}$$

$$d(x_1,x_2)=\sqrt{\sum_{i=1}^{k}(x_1^i - x_2^i)^2}$$

and Manhattan distance is calculated as:

$$d(x_1,x_2)=\sum|x_1^i - x_2^i|$$

$$d(x_1,x_2)=\sum|x_1^i - x_2^i|$$

3.3.2 HIERARCHICAL CLUSTERING:

Hierarchical clustering methods build a cluster hierarchy, i.e. a tree of clusters also known as dendogram. A dendrogram is a tree diagram often used to represent the results of a cluster analysis. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) as shown in Figure 2. An agglomerative clustering starts with one-point clusters and recursively merges two or more most appropriate clusters. In contrast, a divisive clustering starts with one cluster of all data points and recursively splits into non overlapping clusters.

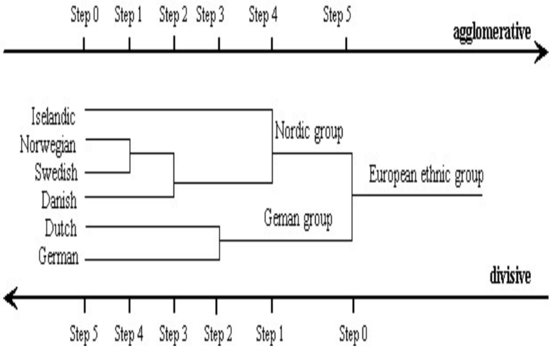

Figure 2. Example of dendogram

In “Cluster analysis”, Brian S. Everitt said that he process continues until a stopping criterion (frequently, the requested number M of clusters) is achieved. Hierarchical methods provide ease of handling of any form of similarity or distance, because use distance matrix as clustering criteria. However, most hierarchical algorithms do not improve intermediate clusters after their construction. Furthermore, the termination condition has to be specified. Hierarchical clustering algorithms include BIRCH and CURE.

OUTLIER ANALYSIS:

“What is an outlier?” Very often, there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers. Outliers can be caused by measurement or execution error. Alternatively, outliers may be the result of inherent data variability. The salary of the chief executive officer of a company, for instance, could naturally stand out as an outlier among the salaries of the other employees in the firm. Many data mining algorithms try to minimize the influence of outliers or eliminate them all together. This, however, could result in the loss of important hidden information because one person’s noise could be another person’s signal. In other words, the outliers may be of particular interest, such as in the case of fraud detection, where outliers may indicate fraudulent activity. Thus, outlier detection and analysis is an interesting data mining task, referred to as outlier mining.

Outlier mining has wide applications. As mentioned previously, it can be used in fraud detection, for example, by detecting unusual usage of credit cards or telecommunication services. In addition, it is useful in customized marketing for identifying the spending behavior of customers with extremely low or extremely high incomes, or in medical analysis for finding unusual responses to various medical treatments. Outlier mining can be described as follows: Given a set of n data points or objects and k, the expected number of outliers, find the top k objects that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data. The outlier mining problem can be viewed as two sub problems: (1) define what data can be considered as inconsistent in a given data set, and (2) find an efficient method to mine the outliers so defined. The problem of defining outliers is nontrivial. If a regression model is used for data modeling, analysis of the residuals can give a good estimation for data “extremeness.”

The task becomes tricky, however, when finding outliers in time-series data, as they may be hidden in trend, seasonal, or other cyclic changes. When multidimensional data are analyzed, not any particular one but rather a combination of dimension values may be extreme. For nonnumeric (i.e., categorical) data, the definition of outliers requires special consideration.

"What about using data visualization methods for outlier detection?" This may seem like an obvious choice, since human eyes are very fast and effective at noticing data inconsistencies. However, this does not apply to data containing cyclic plots, where values that appear to be outliers could be perfectly valid values in reality. Data visualization methods are weak in detecting outliers in data with many categorical attributes or in data of high dimensionality, since human eyes are good at visualizing numeric data of only two to three dimensions.

In this section, we instead examine computer-based methods for outlier detection. These can be categorized into four approaches: the statistical approach, the distance-based approach, the density-based local outlier approach, and the deviation-based approach, each of which are studied here. Notice that while clustering algorithms discard outliers as noise, they can be modified to include outlier detection as a by-product of their execution. In general, users must check that each outlier discovered by these approaches is indeed a "real" outlier.

4.1. DISTANCE-BASED OUTLIER DETECTION:

The notion of distance-based outliers was introduced to counter the main limitations imposed by statistical methods. An object, **o**, in a data set, D, is a distance-based (DB) outlier with parameters pct and dmin,11 that is, a **DB(pct;dmin)**-outlier, if at least a fraction, pct, of the objects in D lie at a distance greater than dmin from **o**. In other words, rather than relying on statistical tests, we can think of distance-based outliers as those objects that do not have "enough" neighbors, where neighbors are defined based on distance from the given object. In comparison with statistical-based methods, distance based outlier detection generalizes the ideas behind discordancy testing for various standard distributions. Distance-based outlier detection avoids the excessive computation that can be associated with fitting the observed distribution into some standard distribution and in selecting discordancy tests.

For many discordancy tests, it can be shown that if an object, **o**, is an outlier according to the given test, then **o** is also a DB(pct, dmin)-outlier for some suitably defined pct and dmin. For example, if objects that lie three or more standard deviations from the mean are considered to be outliers, assuming a normal distribution, then this definition can be generalized by a DB(0:9988, 0:13••) outlier.

Several efficient algorithms for mining distance-based outliers have been developed. These are outlined as follows.

INDEX-BASED ALGORITHM: Given a data set, the index-based algorithm uses multidimensional indexing structures, such as R-trees or k-d trees, to search for neighbors of each object **o** within radius dmin around that object. Let M be the maximum number of objects within the dmin-neighborhood of an outlier. Therefore, onceM+1 neighbors of object **o** are found, it is clear that o is not an outlier. This algorithm has a worst-case complexity of O(n2k), where n is the number of objects in the data set and k is the dimensionality. The index-based algorithm scales well as k increases. However, this complexity evaluation takes only the search time into account, even though the task of building an index in itself can be computationally intensive.

NESTED-LOOP ALGORITHM: The nested-loop algorithm has the same computational complexity as the index-based algorithm but avoids index structure construction and tries to minimize the number of I/Os. It divides the memory buffer space into two halves and the data set into several logical blocks. By carefully choosing the order in which blocks are loaded into each half, I/O efficiency can be achieved.

CONCLUSION:

Intrusion detection systems (IDSs) attempt to identify computer system and network intrusions and misuse by gathering and analyzing data. IDSs have traditionally been developed to detect intrusions and misuse for wired systems and networks. More recently, IDSs have been developed for use on wireless networks. These wireless IDSs can monitor and analyze user and system activities, recognize patterns of known attacks, identify abnormal network activity, and detect policy violations for WLANs. Wireless IDSs gather all local wireless transmissions and generate alerts based either on predefined signatures or on anomalies in the traffic.

Outlier detection and analysis are very useful for fraud detection, customized marketing, medical analysis, and many other tasks. Computer-based outlier analysis methods typically follow either a statistical distribution-based approach, a distance-based approach, a density-based local outlier detection approach, or a deviation-based approach.

REFERENCES:

1) Muazzam siddiqui, "*high performance data mining techniques for intrusion detection*" B.E. NED University of Engineering & Technology, 2000

2) Jones, Anita K., Sielken, Robert S., "*Computer system intrusion detection: A survey",* Technical Report, Computer Science Dept., University of Virginia, 1999.

3) Koziol, Jack, "*Intrusion detection with Snort*", Sams Publishing, 2003.

4) Bace, Rebecca Gurley, Intrusion detection", Macmillan Technical Publishing, 2000.

5) Wenke Lee, *"A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems "*

6) Manoj[1] and Jatinder Singh[2] ,"*Applications of Data Mining for Intrusion Detection*" 1Ph.D., Research scholar, Singhania University, India 2Dean, Engg., DBIEM, Moga, India

7) Scarfone, Karen; Mell, Peter (February 2007). "*Guide to Intrusion Detection and Prevention Systems (IDPS)*". Computer Security Resource Center (National Institute of Standards and Technology) (800–94). Retrieved 1 January 2010.

8) Wang, Ke. *"Anomalous Payload-Based Network Intrusion Detection"*. Recent Advances in Intrusion Detection. Springer Berlin. doi:10.1007/978-3-540-30143-1_11. Retrieved 2011-04-22.

9) Sasha/Beetle , "*A strict anomaly detection model for IDS "* Phrack 56 0x11

10) Perdisci, Roberto; Davide Ariu, Prahlad Fogla, Giorgio Giacinto, and Wenke Lee (2009). *"McPAD : A Multiple Classifier System for Accurate Payload-based Anomaly Detection". Computer Networks, Special Issue on Traffic Classification and Its Applications to Modern Networks* 5 (6): 864–881.

11) Dary Alexandra Pena Maldonado , *Data Mining: A New Intrusion Detection Approach GIAC Security Essentials Certification Practical Assignment Version No 1.4 Option 1* By:,June 19th 2003

12) Abraham, Tomas. "*IDDM: Intrusion Detection Using Data Mining".* Department of Defense- DSTO Electronics and Surveillance Research Laboratory. May 2001.

13) Svetlana Cherednichenko " *Outlier Detection in Clustering*" 24.01.2005University of Joensuu, Department of Computer Science, Master's Thesis Brian S. Everitt, "Cluster analysis". Third Edition, 1993.

14) S. Giha, R. Rasstogi and K. Shim, "*CURE: an efficient clustering algorithm for large databases*". In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, pages 73 – 84, June 1998.

15) L. Kaufman and P. Rousseeuw, "*Finding Groups in Data: An Introduction to Cluster Analysis*". John Wiley Sons, New York, USA, 1990.

16) E. Paquet, *"Exploring anthropometric data through cluster analysis*". Published in Digital Human Modeling for Design and Engineering (DHM), pages, Rochester, MI,June, 2004.

17) Jiawei Han , Micheline Kamber *"Data Mining: Concepts and Techniques"* Second Edition, University of Illinois at Urbana-Champaign.