Classification, Clustering And Intrusion Detection System

Manish Joshi
MCA Final Year
Amity University, Uttar Pradesh
Under Guidance of Ms. Shruti Nagpal

Abstract
Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Classification and clustering techniques in data mining are useful for a wide variety of real time applications dealing with large amount of data. Some of the applications of data mining are text classification, selective marketing, medical diagnosis, intrusion detection systems. In information security, intrusion detection is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource. Intrusion detection systems are software systems for identifying the deviations from the normal behavior and usage of the system. They detect attacks using the data mining techniques-classification and clustering algorithms. In this report I discuss various approaches based on classification and clustering techniques and the use of clustering techniques in intrusion detection. I also discuss some of the application of data mining.

1. Introduction
Clustering is the process of organizing data into meaningful groups, and these groups are called clusters. Clustering groups (clusters) objects according to their perceived, or intrinsic similarity in their characteristics, Even though it is generally a field of unsupervised learning, knowledge about the type and source of the data has been found to be useful in both selecting the clustering algorithm, and for better clustering. This type of clustering generally aims to give tags to the clusters and has found use in fields like content mining. Classification techniques analyze and categorize the data into known classes. Each data sample is labeled with a known class label. Classification is a classic data mining task, with roots in machine learning.

Clustering can be seen as a generalization of classification. In classification we have the knowledge about the object, the characteristics we are looking for and the classifications available. So classification is more similar to just finding "Where to put the new object in". Clustering on the other hand analyses the data and finds out the characteristics in it, either based on responses (supervised) or more generally without any responses (unsupervised).

An intrusion detection system (IDS) inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. They can be categorized into -

1.1 Misuse Detection Vs. Anomaly Detection System
In misuse detection, the IDS analyze the information it gathers and compares it to large databases of attack signatures. Essentially, the IDS look for a specific attack that has already been documented. Like a virus detection system, misuse detection software is only as good as the database of attack signatures that it uses to compare packets against. Misuse detection systems model attacks as a specific pattern and are more useful in detecting known attack patterns.
In anomaly detection, the system administrator defines the baseline, or normal, state of the network, traffic load, breakdown, protocol, and typical packet size. The anomaly detector monitors network segments to compare their state to the normal baseline and look for anomalies.
.

1.2 Network-Based vs. Host-Based System
In a network-based system, or NIDS, the individual packets flowing through a network are analyzed. The NIDS can detect malicious packets that are designed to be overlooked by a firewall??s simplistic filtering rules. In a host-based system, the IDS examines at the activity on each individual computer or host.

1.3 Passive Systems Vs. Reactive Systems
In a passive system, the IDS detect a potential security breach, logs the information and signals an alert. In a reactive system, the IDS respond to the suspicious activity by logging off a user or by reprogramming the firewall to block network traffic from the suspected malicious source.
Data mining approaches can be applied for both anomaly and misuse detection. The data sample is a set of system properties, representing the behavior of the system/user.

Classification techniques are used to learn a model using the training set of data samples. The model is used to classify the data samples as anomalous behavior instance or the normal behavior instance. Clustering techniques can be used to form clusters of data samples corresponding to the normal use of the system. Any data sample with characteristics different from the formed clusters is considered to be an instance of anomalous behavior. Clustering based techniques can detect new attacks as compared to the classification based techniques.

2. Data Mining Classification Methods

The data mining consists of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. The Classification process involves following steps:

1. Create training data set.
2. Identify class attribute and classes.
3. Identify useful attributes for classification (relevance analysis).
4. Learn a model using training examples in training set.
5. Use the model to classify the unknown data samples.

The commonly used methods for data mining classification tasks can be classified into the following groups.

2.1 Decision Trees (Dt's)

A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached. Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

2.2 Neural Networks

Neural networks (NN) are those systems modeled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input. An artificial neural network consists of connected set of processing units. The connections have weights that determine how one unit will affect other. Subset of such units act as input nodes, output nodes and remaining nodes constitute the hidden layer. By assigning activation to each of the input node, and allowing them to propagate through the hidden layer nodes to the output nodes, neural network performs a functional mapping from input values to output values. The mapping is stored in terms of weights over connection.

2.3 Naive Bayesian Classifiers

Naive Bayesian classifiers use the Bayes theorem to classify the new instances of data. Each instance is a set of attribute values described by a vector, X = (x1, x2 . . . , xn). Considering m classes, the sample X is assigned to the class Ci if and only if

$P(X|Ci)P(Ci) > P(X|Cj)P(Cj)$

for all j in (1, m) such that j =6 i.

The sample belongs to the class with maximum posterior probability for the sample. For categorical data P(Xk|Ci) is calculated as the ratio of frequency of value xk for attribute Ak and the total number of samples in the training set. For continuous valued attributes a Gaussian distribution can be assumed without any loss of generality. In naive Bayesian approach the attributes are assumed to be conditionally independent In spite of this assumption, naive Bayesian classifiers give satisfactory results because focus is on identifying the classes for the instances, not the exact probabilities. Application like spam mail classification, text classification can use naïve Bayesian classifiers. Theoretically, Bayesian classifiers are least prone to errors. The limitation is the requirement of the prior probabilities. The amount of probability information required is exponential in terms of number of attribute, number of classes and the maximum cardinality of attributes. With increase in number of classes or attributes, the space and computational complexity of Bayesian classifiers increases exponentially

2.4 Fuzzy Sets

Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today's databases: imprecision, non-specificity, inconsistency, vagueness, etc. Fuzzy sets exploit uncertainty in an attempt to make system complexity manageable. As such, fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems.

3. Clustering Methods

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired.

In general, clustering methods may be divided into two categories based on the cluster structure which they produce.

3.1 Non Hierarchical

The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap. These methods are sometime divided into *partitioning* methods, in which the classes are mutually exclusive, and the less common *clumping* method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar however the threshold of similarity has to be defined.

3.2 Hierarchical

The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or *divisive* methods. In *agglomerative* methods, the hierarchy is build up in a series of N-1 agglomerations, or Fusion, of pairs of objects, beginning with the un-clustered dataset. The less common *divisive* methods begin with all objects in a single cluster and at each of N-1 steps divide some clusters into two smaller clusters, until each object resides in its own cluster. Some of the important Data Clustering Methods are described below.

4. Partitioning Method

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative. Example of portioning method: *k*-means clustering

4.1 K-means Clustering

is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. Given a set of observations ($\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets ($k \leq n$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where

$\boldsymbol{\mu}_i$ is the mean of points in S_i.

4.2 K-mediod Clustering

k-mediod is a classical partitioning technique of clustering that clusters the data set of *n* objects into *k* clusters known *a priori*. It is more robust to noise and outliers as compared to *k*-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances. The most common realization of *k*-mediod clustering is the **Partitioning Around Mediods (PAM)** algorithm and is as follows:

Initialize: randomly select *k* of the *n* data points as the mediods associate each data point to the closest mediod. ("closest" here is defined using any valid distance metric, most commonly Euclidean distance, Manhattan distance or Minkowski distance)

For each mediod *m*

For each non-mediod data point *o*

Swap *m* and *o* and compute the total cost of the configuration

Select the configuration with the lowest cost.

Repeat steps 2 to 5 until there is no change in the mediod.

5. Intrusion Detection Systems

An Intrusion Detection System (IDS) is the device or software application that monitors network and system activities for malicious activities and produces reports to a Management Station.

The characteristics of a good intrusion detection system are:

1. High detection rate.
2. Less false alarms.
3. Less CPU cycles.
4. Quick detection of intrusion.

The user profile, system behavior comprising of the statistics related to the network, CPU, memory, processes, software and applications used by the users constitute the test data for the intrusion detection system.

5.1 Anomaly based intrusion detection system

Anomaly detection was proposed for Intrusion 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. Rest of the section discusses the approaches for anomaly

detection based on the classification and clustering techniques.

5.1.1 Naive Bayesian Approach

A **naive Bayes classifier** is a simple probabilistic classifier based on applying **Bayes' theorem** with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, naïve Bayes classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature, given the class variable.

5.1.2 Neural Networks

Neural Networks can be used for anomaly detection. The approach consists of maintaining a database of a sequence of system calls made by each program to the operating system, used as the signature for the normal behavior. If online sequence of system calls for a program differ from the sequence in the database anomalous behavior is registered. If significant percentage of sequences does not match then alarm for intrusion is raised.

Back propagation network is trained with training set of sequences of system calls. A leaky bucket algorithm is used to capture the temporal locality of the anomalous sequences. When closely related anomalous sequences are faced, counter gets a large value and when a normal sequence is obtained the counter gradually drops down to zero. This leads to intrusion detection only when a lot of similar anomalous sequences are obtained, thereby representing the behavior of intruder.

5.1.3 Hierarchical Clustering

Hierarchical clustering is based on the core idea of objects being more related to nearby objects than to objects farther away. As such, these algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

6. Conclusion

Classification is a classic data mining task, with roots in machine learning. Classification is a field of supervised learning and we have the knowledge about both the characteristics we are looking for and the classifications available. Classification methods are typically strong in modeling interactions.

Clustering can be seen as a generalization of classification and it is unsupervised learning. Clustering techniques can be used for intrusion detection, as they can detect unknown attacks also. They are useful for misuse detection as well as anomaly detection systems.

Intrusion detection systems are one of the key areas of application of data mining techniques. Network based (NIDS) are easy to deploy and can monitor many host based (HIDS) detects at application layer and no trouble with encryption.

References

a. Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann

b. www.it.iitb.ac.in/~kaushal/downloads/seminarreport.pdf

c. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/

d. http://www.ecmlpkdd2006.org/kyriakopoulou.pdf

e. Clustering and Classification-Wikipedia

f. Intrusion Detection System-Wikipedia

g. http://www.sans.org/reading_room/whitepapers/detection/understanding-intrusion-detection-systems_337