Data Mining Techniques Used in Cyber Security

Mrs. A. Meena
Assistant Professor in Computer Science,
AJK College of Arts and Science,
Navakkarai, Coimbatore – 641 105.

Abstract: Data mining is the way toward identifying patterns in big datasets. Data mining methods are vigorously utilized in logical research and additionally in business, generally to accumulate measurements and profitable data to upgrade client relations and marketing techniques. Data mining has likewise demonstrated a helpful apparatus in cyber security for finding vulnerabilities and social affair pointers for base lining.

***Keywords**: Data mining, techniques, cyber securities, etc.,*

*****

Introduction

Data mining is a procedure that includes analyzing information, anticipating future patterns, and making proactive, learning constructed decisions based with respect to vast datasets.

While the term data mining is typically regarded as an equivalent word for Knowledge Discovery in Databases (KDD), it's in reality only one of the steps in this procedure. The primary objective of KDD is to acquire valuable and regularly often obscure data from vast arrangements of information.

The entire KDD process includes four steps

- Pre-processing – selecting, cleaning, and integrating data
- Transformation – transforming information and consolidating it into forms appropriate for mining
- Mining – collecting, extracting, analyzing, and statistically processing data
- Pattern evaluation – identifying new and unusual patterns and presenting the knowledge gained from data mining
- Data mining helps you find new interesting patterns, extract hidden (yet useful and valuable) information, and identify unusual records and dependencies from large databases. To obtain valuable knowledge, data mining uses methods from statistics, machine learning, and artificial intelligence (AI), and database systems.

In recent years, many IT industry giants such as Comodo, Symantec, and Microsoft have started using data mining techniques for malware detection.

Data mining for malware detection

Data mining is one of the four recognition strategies utilized today to recognize malware. The other three are examining, action observing, and integrity checking.

When constructing a security application, designers utilize Data mining strategies to enhance the speed and nature of malware identification and additionally to expand the quantity of distinguished zero-day attacks.

Malware detection strategies

There are three strategies for detecting malware:

- ✓ Anomaly detection
- ✓ Misuse detection
- ✓ Hybrid detection

Anomaly Detection

Anomaly detection includes displaying the ordinary conduct of a framework or system with the end goal to distinguish deviations from typical usage patterns. Anomaly-based strategies can distinguish even already obscure attacks and can be utilized for characterizing marks for misuse detectors.

The principle issue with anomaly detection is that any deviation from the standard, regardless of whether it is a legitimate conduct, will be accounted for as an irregularity, therefore creating a high rate of false positives.

Misuse Detection

Misuse detection, otherwise called signature-based detection, distinguishes just known attacks dependent on models of their signatures. This strategy has a lower rate of false positives however can't recognize zero-day attacks.

Detection process

When using data mining, malware detection consists of two steps:

- ✦ Extracting features
- ✦ Classifying/clustering

In the initial step, different features, for example, API calls, n-grams, binary strings, and program behaviours are extricated statically and progressively to catch the attributes of the record tests. Feature extraction can be performed by running static or dynamic examination (with or without really running conceivably hurtful programming). A hybrid approach that joins static and dynamic investigation may likewise be utilized.

During classification and clustering, record tests are characterized into groups based on feature analysis. To order tests, it can utilize classification or clustering techniques.

To characterize document tests, we have to manufacture an classification model (a classifier) utilizing grouping algorithms, for example, RIPPER, Decision Tree (DT), Artificial Neural Network (ANN), Naive Bayes (NB), or Support Vector Machines (SVM). Clustering is utilized for grouping malware tests that have similar characteristics.

Utilizing machine learning procedures, every classification algorithm builds a model that represents to both benign and malicious classes. Preparing a classifier utilizing such document test accumulation makes it conceivable to recognize even recently released malware..

Data mining for intrusion detection

Besides distinguishing malware code, data mining can be successfully used to recognize interruptions and break down review results to identify abnormal examples. Malevolent interruptions may incorporate interruptions into systems, databases, servers, web customers, and working frameworks.

There are two types of intrusion attacks that can detect using data mining methods:

- ⇒ Host-based attacks, when the intruder focuses on a particular machine or a group of machines
- ⇒ Network-based attacks, when the intruder attacks the entire network (for instance, causing a buffer overflow

To recognize host-based attacks, we have to investigate features extracted from projects, while to distinguish network-based attacks, we have to break down system movement. What's more, much the same as with malware identification, we can search for either irregular conduct or instances of misuse?

Data mining for fraud detection

We can distinguish different kinds of fraud utilizing data mining strategies, from money related misrepresentation to media communications fraud and PC interruptions. Fraudulent activities can be recognized with the assistance of directed and unsupervised learning.

With administered adapting, every single accessible record are delegated either false or non-fake. This arrangement is then utilized for preparing a model to recognize conceivable fraud. The fundamental disadvantage of this strategy is its failure to identify new sorts of attacks. Unsupervised learning techniques help distinguishes protection and security issues in information without utilizing statistical analysis.

Data mining pros and cons

Using data mining in cyber security lets you

- ✦ process large datasets faster;
- ✦ create a unique and effective model for each particular use case;
- ✦ apply certain data mining techniques to detect zero-day attacks.

While this list of the benefits is impressive, there are also certain drawbacks you need to know about:

- ⇒ Data mining is complex, resource-intensive, and expensive
- ⇒ Building an appropriate classifier may be a challenge
- ⇒ Potentially malicious files need to be inspected manually
- ⇒ Classifiers need to be constantly updated to include samples of new malware
- ⇒ There are certain data mining security issues, including the risk of unauthorized disclosure of sensitive information

Data mining encourages rapidly examining large datasets and naturally finding hidden patterns, which is crucial with regards to making a compelling anti-malware arrangement that is ready to distinguish already obscure threats. Be that as it may, the last consequence of utilizing data mining techniques dependably relies upon the nature of information we utilize.

When utilizing data mining in cyber security, it's significant to utilize just quality information. Be that as it may, getting ready databases for examination requires a considerable measure of time, exertion, and assets. You have to clear the entirety of your records of copy, false, and inadequate data previously working with them. Absence of data or the nearness of copy records or errors can fundamentally reduce the viability of complex data mining methods. Just utilizing precise and complete information can guarantee high calibre of examination.

Conclusion

Data mining has extraordinary potential as a malware detection tool. It enables to dissect large amount of data and concentrate new learning from it.

The primary advantage of utilizing data mining methods for recognizing malicious programming is the capacity to distinguish both known and zero-day attacks. Notwithstanding, since a formerly obscure however authentic movement may likewise be set apart as potentially fraudulent, there's the likelihood for a high rate of false positives.

References

[1] Data Mining for Security Applications : Bhavani Thuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen.

[2] Rakesh Agrawal, Tomasz Imieliski, and Arun Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data.

[3] Daniel Barbara and Sushil Jajodia, editors. Applications of Data Mining in Computer Security. Kluwer Academic Publishers.

[4] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and J Sander. Lof: identifying density-based local outliers. In Proceedings of the 2000 ACM SIG-MOD international conference on Management of data.

[5] Varun Chandola and Vipin Kumar. Summarization {compressing data into an informative representation. In Fifth IEEE International Conference on Data Mining.

[6] Thuraisingham, B., “Web Data Mining Technologies and Their Applications in Business Intelligence and Counterterrorism”, CRC Press, FL, 2003.

[7] Chan, P, et al, “Distributed Data Mining in Credit Card Fraud Detection”, IEEE Intelligent Systems.

[8] Lazarevic, A., et al., “Data Mining for Computer Security Applications”, Tutorial Proc. IEEE Data Mining Conference, 2011.

[9] Thuraisingham, B., “Managing Threats to Web Databases and Cyber Systems, Issues, Solutions and Challenges”, Kluwer, MA 2004 (Editors: V. Kumar et al).

[10] Thuraisingham B., “Data Miming, Privacy, Civil Liberties and National Security”, SIGKDD Explorations, 2012.