

Performance Analysis of Data Mining using Various Detection Systems

P. Venkatesh and V. Sessa Sai Krishna

Department of Computer Science and Engineering, Sathyabama University, Chennai-119, India.

doi: <http://dx.doi.org/10.13005/bbra/2214>

(Received: 30 July 2015; accepted: 03 September 2015)

Nowadays it is most important to protect a high level data protection to make sure secure and believed communication of data between different groups. But it is not possible because of terror related activities, intrusion and outlier. It can be analyzed by using various detection systems of data mining. Different data mining algorithms plays an important role in various detection systems. In this paper we have used few algorithms named apriori algorithm and genetic algorithm for detecting various systems.

Key words: Data Mining, Intrusion Detection System, Terror Related Detection System, Outlier Detection Scheme and various Data Mining Algorithms.

Data mining tasks are classified by two type's namely predictive mining & descriptive mining. The expressive mining methods named Sequential Pattern discovery, Association, Clustering is used to discover individual interpretable outlines that depict the data. The analytical mining methods such as detection, deviation, Regression, classification, etc., are used to calculate unidentified or future values of other erratic ¹.

As a greatly application-ambitious area, data mining has included several methods from other areas such as algorithms, information retrieval, database and data warehouse systems, high-performance computing, statistics, pattern recognition, visualization, machine learning, and lots of application areas. It was shown in Figure 1. The interpositional nature of data mining research and development provides drastically to the triumph of data mining and its wide-ranging applications.

In this chapter, the examples of a number of controls that sturdily authority the growth of data mining techniques.

Different Detection Schemes

Intrusion Detection System

Intrusion is described as the efforts to avoid the security methods of a network or computer. The fundamental targets of network security are availability, integrity and confidentiality. Since integrity absorbs no deception in information, secrecy means solitude of the information and accessibility absorbs the charisma of the information in the perfect style, if it is required. Hence, Intrusion is a set of useless exploits intended to conciliation these protection targets. To avoid these exploits, intrusion avoidance only is not adequate. Consequently prior to Intrusion avoidance, Intrusion detection is required ¹.

An intrusion detection system (IDS) is a software mechanism or tool that observes system or network behaviors for spiteful actions or strategy breaches and creates details to a executive location. IDS appear in a selection of essences and move toward the target of identifying apprehensive traffic

* To whom all correspondence should be addressed.

in special approaches. This is based on host (HIDS) and network based (NIDS) intrusion detection systems. A few schemes can effort to stop an intrusion effort although it is neither needed nor anticipated of a screening scheme. IDPS are mainly alert on recognizing doable occurrences, cataloguing information on them and accounting efforts. In count, institutes utilize IDPSes for some other reasons, such as recognizing difficulties with protection rules, manuscript accessible perils and daunting entities from breaching protection rules. IDPSes is a required addition to the protection communications of practically all institute ². A typical location for an intrusion detection system was shown in Figure.2.

Terror Related Detection System

An Activities of Terrorist Typical is described as an entree to data related for terrorists and their followers. A common depiction of a

scheme based on the recommended methodology is obtainable in Figure 3. Every consumer beneath scrutiny is notorious as a consumer’s network having a unique protocol address. In this case, the detected protocol may be used to place the network and optimistically the alleged terrorist who may tranquil be logged on to the same network. The recommended method has two modes of operation:

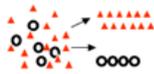
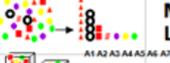
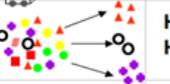
Studying usual terrorist activities

A set of website pages from terror-linked sites is downloaded and symbolized as a group of vectors used by the vector space model. The gathered information is used to obtain and symbolize the usual activities of the terrorists and their followers by relating methods of unsubstantiated clustering. In view of the fact that the IP addresses of downloaded pages are disregard, stirring the similar or same contents to a

Table 1. Differences in technology intrinsic to IPS and the IDS deployment

	Intrusion Prevention System	IDS Deployment
Placement in network infrastructure	Part of the direct line of communication (inline)	Outside direct line of communication (out-of-band)
System type	Active (monitor & automatically defend) and/or passive	Passive (monitor & notify)
Detection Mechanisms	1. Statistical anomaly-based detection 2. Signature detection: -Exploit-facing signatures -Vulnerability-facing signatures	1. Signature detection: -Exploit-facing signatures

Table 2. Data mining functionality

Techniques	Algorithm	Applicability
Classification 	Logistic Regression (GLM) Decision Trees Naïve Bayes Support Vector Machine	Classical statistical technique Popular / Rules / transparency Embedded app Wide / narrow data / text
Regression 	Multiple Regression (GLM) Support Vector Machine	Classical statistical technique Wide / narrow data / text
Anomaly Detection 	One Class SVM	Lack examples of target field
Attribute Importance 	Minimum Description Length (MDL)	Attribute reduction Identify useful data Reduce data noise
Association Rules 	Apriori	Market basket analysis Link analysis
Clustering 	Hierarchical K-Means Hierarchical O-Cluster	Product grouping Text mining Gene and protein analysis
Feature Extraction 	Nonnegative Matrix Factorization	Text analysis Feature reduction

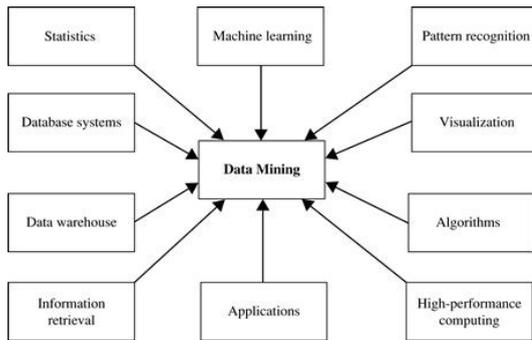


Fig. 1. Statistics of Data Mining

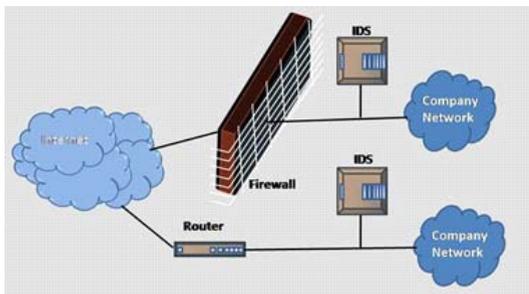


Fig. 2. Typical locations for an intrusion detection system

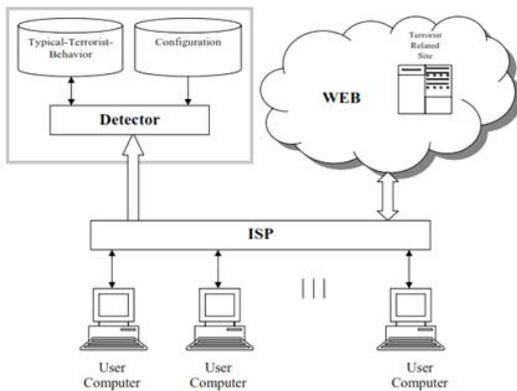


Fig. 3. Detection methodology

new address, as regularly passed out by terror-linked sites. It won't change the detection exactness of the innovative technique.

Observing users

In this mode is planned by comparing the content of data accessed by users to the Usual Terrorist Activities to detecting terrorist users. Access vector is used to convert the content of the data accessed by a user on the Web into a vector. When the comparison among the

access vector and the Usual Terrorist Activities is over a predefined doorsill, an alarm is concerned. Since the system does not require storing either the visited IP addresses or the actual content of the viewed pages, the privacy of regular users is preserved. Due to widespread dimensionality decrease processes, the access vectors don't clasp adequate data to reinstate the real content of the web page. New types of viewed content are disregarded by the system. Figure 3 shows the detection methodology of data mining process.

Outlier Detection Scheme

Outlier Detection based on Distance

The parameters p and D are used. It is well suited for circumstances where the experimental circulation does not robust at all regular circulation and the most significant on outlier based on distance is that it is exactly described for k -dimensional datasets of k for all the values. It has two easy algorithms, that are having a difficulty of O, N is the number of objects and k is dimensionality in the dataset. Datasets are supported by these algorithms with more than that two attributes ^{3,4,5}.

Local Outlier Detection scheme based on Density

It allocates to all the point an amount to be an outlier. This amount is called the local outlier factor (LOF) of a point. That is local and the amount depends on how secluded the point is with respect to the neighboring environs. Outliers are data points with high LOF ranges while data points with low LOF ranges are expected to be regular with respect to their environs in LOF algorithm. Low-density environs are an indication of high LOF and therefore high possible of individual outlier ^{6,7}.

Data Mining: Algorithms

The Apriori Algorithm

The crucial Apriori algorithm ⁸ discovers regular item sets for Boolean connection laws, getting as effort a database T of contracts and the least sustain for the laws. Apriori property is used, if an item set I is not regular, the item set $I * A$ is also not regular. That is every non blank detachments of a regular point situate should besides be regular.

Genetic Algorithm

A Genetic Algorithm (GA) is a indoctrination method that imitates organic progression as a trouble-solving approach.

Darwinian's principle of development is based on this algorithm and endurance of robust to optimize an inhabitant of applicant clarifications through a pre described condition. An advancement and ordinary selection used GA that applies a genetic material similar to data arrangement and progress the genetic materials using collection, recombination and alteration machinists. The progression regularly commenced with erratically produced inhabitants of genetic materials, which symbolize every probable elucidation of a trouble that are measured applicant explanations. As of all genetic materials dissimilar points are programmed as characters, numbers or bits. These points can be submitted to as genetic materials. Three reasons will have crucial brunt on the efficiency of the algorithm and the purposes, if used GA for solving different problems. They reasons are the representation of individuals, the GA parameters and the fitness function. The GA is engaged in intrusion detection to develop a set of union regulations from system review data, and the sustain-assurance construction is employed as a robustness function to critic the eminence of all the laws. GA is it is vigorous to noise if it has good properties, nature erudition potentials; no incline data is needed to discover the universal finest or sub-finest explanation. High molest recognition rate and low sham-constructive rate are the benefits of GA methods. Hence GA for Intrusion Detection is used ⁹.

Performance Evaluation

In this paper, various data mining algorithms are used for the different detection schemes and the performances of the schemes are described. The following table.1 summarizes the differences in technology intrinsic to IPS and the IDS deployment.

The techniques, algorithms and applicability are briefly described by analyzing the performance of the data mining schemes and systems. Data mining functionality is shown in table.2. It shows the broad function of the data mining schemes.

CONCLUSION

In computer communication area, intrusion and terror relative activities are main

issues. Different data mining schemes and algorithms are used to detect these kinds of issues. In this paper, apriori and genetic algorithms are described and few other algorithms and their applicability were explained. The performance evaluation of the detection system for intrusion and terror relative activities is explained by using the proper data mining algorithm. The techniques used to detect the function of data mining schemes are described.

REFERENCES

1. Navneet Kaur H, "Fuzzy data mining based intrusion detection system using genetic algorithm", *International Journal of Advanced Research in Computer and Communication Engineering*, 2014; **3**(1): 5021-5028.
2. Karen S, Peter M, "Guide to Intrusion Detection and Prevention Systems (IDPS)", Computer Security Resource Center, 2010.
3. Knorr EM, Ng RT, "Algorithms for Mining Distance-Based Outliers in Large Dataset", In Proceeding of the 24th VLDB Conference, New York, USA, 1988.
4. Knorr EM, Ng RT, Tueakov V, "Distance-Based outliers: Algorithms and Applications", In Proc. *VLDB International Journal on Very Large Data Bases*, 2000; **8**(3-4): 237-253.
5. Knorr EM., Ng RT, Zamar RH, "Robust space transformations for distance-based operations", In Proceedings of the seventh ACM SIGKDD International conference on Knowledge discovery and data mining KDD'01, 2001; 126-135.
6. Breunig M, Kriegel HP, Ng RT, Sander J, "LOF: Identifying Density-Based Local Outliers", Proceedings of 2000 ACM SIGMOD International. Conference. on Management of Data, Dallas, TX, SIGMOD'00, 2000, pp. 93-104.
7. Mansur MO, Mohd Noor Md. Sap, "Outlier detection technique in data mining: A Research Perspective", Proceedings of the Postgraduate Annual Research Seminar, 2005.
8. Agrawal R, Srikant R, "Fast algorithms for mining association rules", In Proceedings of the 20 th international conference on very large databases held in Santiago, Chile, 1994; **12-15**; 1-32.
9. Naser Azad, Vahid Ranjbar , Davood Khani , Sara Taheri Moosavi " Information Disclosure by Data Mining Approach" *Indian Journal of Science and Technology*, 2015; **8**(2): 212-216.