International Journal of Science, Engineering and Technology
An Open Access Journal

A Survey on Fast Distributed Algorithm on Data Mining

[1]Sudha, [2]Poovarasi, [3]Revathi

Abstract

The perception of data mining is to detect the significant data, patterns and trends from magnanimous store of data by using number of algorithms. Intrusion and detection system is the security management system for computers and network that tries to trace the attacks. The fast distributed mining algorithm is applied to the transaction log of the database to recognize all the frequent patterns on the database and also it identifies intermittently any new patterns detected. The framework is cogitating mainly on increasing the accuracy of the patterns detected on the database transactions. It focuses on detecting patterns that were not detected previously by old versions of the algorithms. In addition to, enhancing the performance of the model proposed, especially with high-capacity databases.

Keywords: Association Rule, Data mining, Intrusion and Detection System (IDS), Fast Distributed Mining (FDM).

Introduction

Data mining is a process of discovering and extracting various models, patterns, summaries, and derived values from a given collection of data. Furthermore, it involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. It is ordinarily practiced in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific breakthrough. Finally, it is the process of placing a series of appropriate queries to extract information from large amounts of data in the database

Various data mining's methods and algorithms have been used such as classification tree and support vector machines for intrusion detection, Genetic Algorithms, Neural Networks, and Clustering all these methods helps to provide a good level of security to the systems from external and internal attacks also from new attacks.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining).

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Today, organizations are accumulating vast and growing amounts of data in different formats and different databases which includes operational or transactional data such as sales, cost, inventory, payroll, and accounting, nonoperational data such as industry sales, forecast data, and macro economic data and meta data (data about the data itself) such as logical database design or data dictionary definitions.

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Related Works

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis.

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

In data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

Large-scale information technology has been evolving separate transaction and analytical systems; data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks.

Generally, any of four types of relationships are sought under data mining. They are classes in which stored data is used to locate data in predetermined groups, Clusters in which data items are grouped according to logical relationships or consumer preferences, Associations in which data can be mined to identify associations and sequential patterns in which data is mined to anticipate behavior patterns and trends.

To extract the data following steps are to be performed such as extract, transform, and load transaction data onto the data warehouse system, store and manage the data in a multidimensional database system, provide data access to business analysts and information technology professionals, analyze the data by application software and present the data in a useful format, such as a graph or table.

Different levels of analysis are available such as artificial neural networks, genetic algorithm, and decision trees, nearest neighbor method, rule indication and data visualization.

Artificial neutral network in which non-linear predictive models that learn through training and resemble biological neural networks in structure.

Genetic algorithms in which optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

Decision trees are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

Nearest neighbor method is a technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1).

Rule induction is the extraction of useful if-then rules from data based on statistical significance.

Data visualization is the visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Traditionally, intrusion-detection techniques are classified into four broad categories such as misuse detection, anomaly detection, host based IDS and network based IDS. Misuse detection works by searching for the traces or patterns of well-known attacks, anomaly detection uses a model of normal user or system behavior and flags significant deviations from this model as potentially malicious, host-based IDS analyze host-bound audit sources such as the operating system audit trails, system logs, or application logs and network-based IDS analyze network packets that are captured in a network.

All Intrusion Detection Systems use one of two detection techniques such as signature based and anomaly based approach. Signature based approach monitors the system activities and compares them against a database of signatures from known malicious threats and anomalous Based Approach is based on the normal behavior of the system.

Techniques and Algorithms

Some of the models, techniques and algorithms being used in the existing system are discussed and summarized as follows.

A. *Aprioriall***:** In this algorithm, all the possible sequences in the database are scanned to generate the Candidates Ck. Disadvantages, number of database scan and time taken while the main advantage is the simplicity in the implementation compared to others.

B. *Apriorisome*: In this algorithm, the database or customer's sequences are scanned on two stages. Forward phase that generates Ck for certain lengths depending on the value specified by next function where Ck is generated from Ck-l and backward phase to scan the missing sequences.

C. *Dynamicsome***:** In this algorithm, the sequences are scanned on four stages. Initialization, where item sets are generated until certain length, which is the step value. Forward phase, where sequence whose length is multiple of the step is generated where Ck is generated on the fly. Intermediate phase, to generate the missing Ck and backward phase, to filter the item set and provide the ones with minimum support.

D. *Prefix span*: Instead of projecting sequence database by considering all possible occurrence the projection is based only on frequent prefixes. Prefix span mines the complete set of patterns and is efficient and runs considerably faster than both Apriori-based GSP algorithm and free span. Among different variations of Prefix span, bi-level projection has better performance at disk-based processing, and pseudo-projection has the best performance when the projected sequence database can fit in main memory. It reduces the number of combination to be examined when the database is large, reduces candidate subsequences and reduces the size of projected databases.

E. *Warfare scenario*: After the detection of a cyber attack on a database system, the intrusion response team of any organization needs to know the damage profile immediately in order to design an appropriate response strategy. Unfortunately obtaining the precise damage status can take up to hour even days. This is because existing approaches to database damage assessment involve significant amount of work including scanning the log file or other auxiliary data structures but this approach concentrates on making an estimated damage profile as soon as possible. This model is based exclusively on apriori knowledge of data relationships mined during normal database operation phase. This knowledge can be used during damage assessment phase for faster damage assessment. The techniques used for warfare scenario are as follows

1. *Intra transaction data relationship miner*: To discover data dependencies that is related to sequence of operation in database transactions.
2. *Inter transaction data relationship miner*: Within a time, one data must be updated after other which are represented by dependency rules.
3. *Damage assessment procedure*: Generate damage evaluation graph by using intra transaction and inter transaction data dependency.

F. *Set based approach*: The set-based approach relaxes the constraints described in Apriori (All/Some), and improves the performance while being more user-oriented and self-adaptive than the probabilistic knowledge representation. The approach can be extended to more set-based mathematical models for further data analysis in order to discover hidden knowledge and patterns with the improved workflow and set based representation.

G. *RBAC*: Information is valuable asset of any organization which is stored in databases. Data in such databases may contain credit card numbers, social security number or personal medical records etc. Failing to protect these databases from intrusions will result in loss of customer's confidence and might even result in lawsuits. Traditional database security mechanism does not design to detect anomalous behavior of database users. There are number of approaches to detect intrusions in network. But they cannot detect intrusions in database. There have been very few ID mechanisms specifically tailored to database systems. We propose transaction level approach to detect malicious behavior in database systems enabled with Role Based Access Control (RBAC) mechanism

H. Bagging: Bagging is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression, reducing variance and avoiding over fitting. Although it is usually applied to decision tree methods, it can be used with any type of method. In Bagging, each classifier is built individually by working with a bootstrap sample of the input data.

I. *Spambase:* It employed a dataset known as SPAMBASE, which features a collection of emails labeled as spam and non-spam for classification purposes. In order to analyze and quantify the anomalies, we have employed the SPAMBASE dataset, which contains 57 data attributes related with the frequency of some words in the email's content. This data set was created in order to improve security software in computer networks, as attacks using spam e-mails can cause losses such as unnecessary time spending, cost increasing, and productivity loss, improper or offensive content.

Table 1 Comparison of Data Mining Techniques

PAPER	ADVANTAGE	DISADVANTAGE
An evaluation of data mining classification models for network intrusion detection	Lowest computation complexity	Datasets with heterogeneous no of traffic.
An internal intrusion detection and protection system by using data mining forensic techniques	Efficiently resist insider attack. Enhance the accuracy of attack detection	Takes long time to identify the user. Processing big data is indeed.
Database intrusion detection using sequential data mining approach	Better detection rate. Identify all frequent patterns.	Repeated patterns. Accuracy of the system.
Detection of intrusive activity in databases combining multiple evidences and belief updates	It gives independent evidences about a transaction's behavior. It facilitates the detection of malicious activity in the database.	System accuracy
Machine learning proposed approach for detecting DB intrusions in RBAC enabled database	Reduce false positive rate Detect malicious behavior in database systems	Correlation among queries
Mining data relationships for DB damage assessment in a post info warfare scenario	Help to deduce the damage profile Very time consuming	If data dependency decreases accuracy of assessment also decreases.
Mining Sequential Patterns	Mine frequent patterns over a large database. Better detection rate	Item categories Repeated patterns
Prefix Span: Mining sequential patterns efficiently by prefix projected pattern	Reduce the candidate subsequence generation Reduces the size of projected database	Time constrain

Set based approach in mining sequential patterns	Improves performance System overhead is minimized More users oriented and self adoptive.	Avoid multiple database scans
Spam intrusion detection in computer network using intelligent techniques	Identify unusual traffic patterns. Techniques based on trees and forest work efficiently.	Recent pattern recognition

Conclusion

This paper presents a fast distributed mining algorithm for different versions of the apriori sequential data mining algorithm. The original and introduced algorithms both were implemented and applied to realistic huge transaction log. In order to increase the accuracy, decrease the rate of false alarms and repeated pattern. We propose alternative protocol (UNIFI-KC) Unifying lists of locally Frequent Itemsets and (THRESHOLD) Secure computation of the t-threshold function for the secure computation of the union of private subsets. We use association rules that help uncover relationships between seemingly unrelated data in a relational database or other information repository.

References

[1] "An evaluation of data mining classification models for network intrusion detection" Chakchai So–In, Member, IEEE, Nutakarn Mongkonchai, Phet Aimtongkham, Kasidit Wijitsopon and Kanokmon Rujirakul

[2] "An internal intrusion detection and protection system by using data mining forensic techniques" Fang-Yie Leu, Kun-Lin Tsai, Member, IEEE, Yi-Ting Hsiao, and Chao-Tung Yang

[3] "Database intrusion detection using sequential data mining approach" Pakinam Elamein Abd Elaziz, Mohamed sobh, Hoda K. Mohamed.

[4] "Detection of intrusive activity in databases combining multiple evidences and belief updates" Suvasini Panigrahi, Shamik Sural, and A. K. Majumdar

[5] "Machine learning proposed approach for detecting DB intrusions in RBAC enabled database" Udai Pratap Rao, G. J. Sahani, Dhiren R. Patel

[6] "Mining data relationships for DB damage assessment in a post info warfare scenario" Yi Hu and Brajendra Panda, Member, IEEE

[7] "Mining Sequential Patterns" Rakesh Agrawal, Ramakrishnan Srikant

[8] "Prefix Span: Mining sequential patterns efficiently by prefix projected pattern" Jian Pei, Jiawei Han, Behzad Mortazavi-As1, Helen Pinto

[9] "Set based approach in mining sequential patterns" Shang Gao, Reda Alhajj, Jon Rokne, Beirut, Lebanon, Jiwen Guan

[10] "Spam intrusion detection in computer network using intelligent techniques" Patricia Bellin Ribeiro, Luis Alexandre da Silva, Kelton Augusto Pontara da Costa

Author's details

[1]Assistant Professor, Computer Science Engineering, Christ college of Engineering and Technology, Pondicherry, India, sudhacse@christcet.edu.in

[2]Student, Computer Science Engineering, Christ college of Engineering and Technology, Pondicherry, India, puvi.murthy14@gmail.com

[3]Student, Computer Science Engineering, Christ college of Engineering and Technology, Pondicherry, India, revathianand6@gmail.com

Copy for Cite this Article- Sudha, Poovarasi and Revathi, "A Survey on Fast Distributed Algorithm on Data Mining", *International Journal of Science, Engineering and Technology,* Volume 4 Issue 2: 2016, pp. 390- 395.