ISSN: 2278-5213

REVIEW ARTICLE

Data Mining Techniques and their role in Intrusion Detection Systems

Amit Sharma[1], S.N. Panda[2] and Ashu Gupta[3]
[1,3]School of Information Technology, Apeejay Institute of Management Technical Campus, Jalandhar, Punjab, India; [2]RIMT, Mandigobindgarh, Punjab, India
profamitsharma@yahoo.co.in; + 91 9876052925

Abstract

As more and more crucial data is being loaded on computer servers, the security of government and commercial industrial servers is the main concern. The Intrusion Detection Systems can address these problems but they face challenges in robust and changing environments. Data-Mining based IDS at the same time can provide more accuracy of results and Data mining can automatically find the relationships and similarities in patterns in raw data, and can deliver results that can be either utilised in an automated decision support system or assessed by a human analyst thus enhancing the quality of the intrusion detection process. This review will focus on the data mining techniques and their role in intrusion detection systems.

Keywords: Data-mining, intrusion detection, data warehousing, access patterns, data mining techniques.

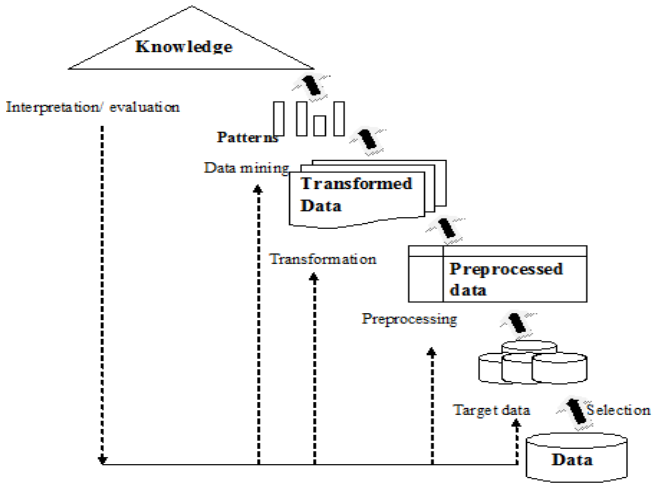
Introduction

Data mining, as the name implies, is about mining data. Just as gold miners sieve through a lot of material to find the precious metal, in data mining we sieve through a lot of data to find useful information (Lee and Stolfo, 1998). Data mining is the process of sorting through large database or data warehouse and extracting knowledge interested by the people. The extracted knowledge may be represented as concept, rule, law and model. The purpose of data mining is to help the decision-maker in order to find potential association between data, found neglected elements which might be very useful for trends and decision- making behaviour. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley *et al*., 1992) and "the science of extracting useful information from large data sets or databases" (Hand *et al*., 2001).

The name is something of an oddity-it should actually have been the name of the product that we mine, for instance, we do not refer to 'gold mining' as 'soil mining'. A distinction used to be made between data mining and Knowledge Discovery in Databases (KDD)-the former being the extraction of information and the latter being processing of that information- but now they are used interchangeably. Data mining is data-driven research that starts with data; it is not empirical research because the users do not start with a premise and then experiment, and it is not theoretical research because it does not start with a theory and uses data to prove it. Due to huge amount of data being created from normal business operations and the competitive business environment, there is a demand to utilise information languishing in modern data repositories to make effective, efficient and prompt decisions. Data mining grew from this demand (Fig. 1).

It applies machine intelligence and statistical tools to extract novel, useful and meaningful patterns in data which are not accessible through data query language. It identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, non-statistician users have the opportunity to identify key attributes of business processes and target opportunities. However, abdicating control of this process from the statistician to the machine may result in false-positives or no useful results at all (Yusufovna, 2008). Data mining sits at the intersection of three disciplines-Databases, Statistics and Machine Intelligence- the differences being that databases are passive as to knowledge discovery whereas data-mining is not; statistics usually makes generalisations from a small number of observations but data mining extracts specific information from a huge number of observations; and machine intelligence usually deals with small- sized data while data mining deals with large volumes of data (Bloedorn, 2001).

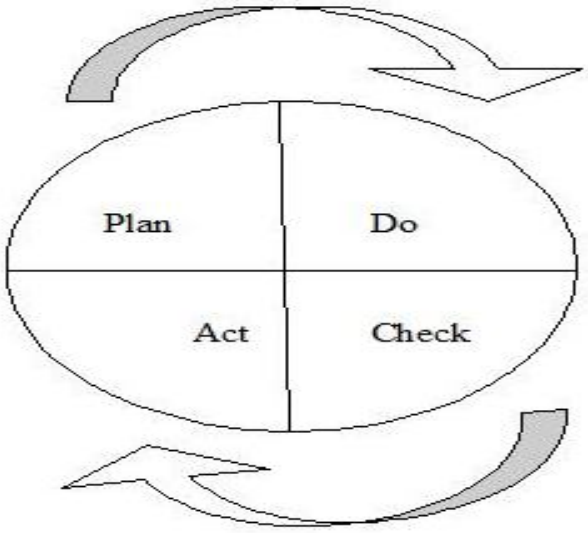
Fig. 1. Basic approach of data mining.



Data mining has been made popular primarily by companies with a strong consumer focus, such as retail, financial, communication and marketing organisations. Data mining enables such 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. Furthermore, it enables them to determine the impact on sales, customer satisfaction and corporate profits. Data mining can automate the process of finding relationships and patterns in raw data, and can deliver results that can be either utilised in an automated decision support system or assessed by a human analyst (Berry and Lino, 1997). Data mining or KDD has been described as the application of the *scientific method* to a database.

The scientific method was first described by Sir Francis Bacon in 1620 and is frequently explained by the graphical method of the Plan- Do- Check- Act (PDCA) cycle (Fig. 2). The PDCA cycle is called the Deming Cycle in honour of W. Edwards Deming, who is credited with revolutionising industrial growth in Japan after the Second World War. To put the Deming cycle succinctly: at the Plan stage the problem is formulated; at the Do stage experimentation is performed; at the Check stage the results are evaluated; and at the Act stage the results are implemented if successful, or go back to the drawing board or problem reformulation if unsuccessful (Rother, 2009).

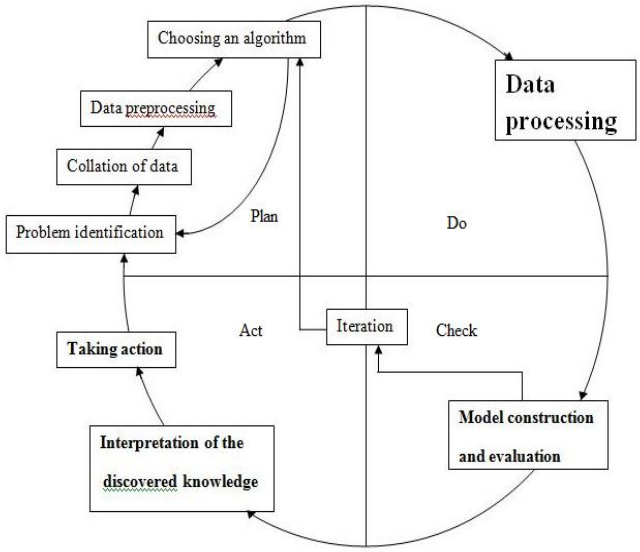
Fig. 2. PDCA cycle.



Data mining follows a logical sequence of steps to build a model from a database. It is cyclic and the sequence is therefore sometimes referred to as data mining life cycle (Anderson, 2011). The number of steps can vary from 4 to 12. The best known are the six steps of the Cross-Industry Standard Process for Data Mining (CRISP- DM), which is the work of a consortium of data mining practitioners.

The knowledge discovery process tends to be highly iterative and interactive. Thus, two smaller PDCA cycles (loops) can be seen within the larger PDCA cycle (Rother, 2009). The eight- step process in the framework of the PDCA cycle which captures all facets of the data mining task is shown in Figure 3.

Fig. 3. Data-mining life cycle.



The major steps are problem identification; gathering and selection of data; data pre-processing for missing, duplicate or erroneous information; selection of appropriate learning algorithms; preparation and processing of data; construction and evaluation of the models; interpretation of the discovered knowledge; and finally, taking action.

The eight steps can be described as:
Problem identification: In the beginning, a focus or definition of a problem is needed that is wished to be addressed with data mining. This will dictate the data mining strategy that is wished to be adopted to meet the expectations. The level of expectation for the solution is also needed to be set, say 80% or 98% satisfaction. The CRISP-DM reference model identifies this step as *business understanding*. Without business understanding and requirements, useful data mining cannot be done.

Collation of data: The problem definition provides the scope of relevant data. If data from a data warehouse are adequate, then the problem of collation does not arise. However, if current data are required they have to be gathered from many sources, which may be a difficult task.

Data pre-processing: If the data comes from a warehouse, no pre-processing of data is usually required because the warehouse data have already been filtered and cleaned and missing values taken care of. However, if the data are collated from a number of sources, including transactional data, they need to be organised and cleaned so that a data mining technique can be

readily applied. The data also have to be consistent. For example, in one database male and female may be represented as M and F, and in another database they may be represented as 1 and 0. Such anomalies have to be removed and any representation has to be made uniform across all sources. The CRISP-DM reference model identifies this step and the previous step as *data understanding*.

Algorithm selection: Nowadays quite a good number of data mining algorithms are available for public use. It is very difficult to say which particular algorithm would best suit a given problem, so a trial- and- error approach is frequently adopted. However, there has been some research on how to select an appropriate algorithm to solve a given data mining problem. In general, parametric algorithms are relatively more suited for the data mining task. These involve choosing the optimal parameters to receive the best solution. The knowledge gained from data collation, pre-processing, and algorithm selection may equip the user to better reformulate the problem or ask the right questions.

Data processing: This task is intended to make the format of the data compatible to the chosen algorithm of the previous step. This may involve data normalisation, data transformation or data integration. Some algorithms cannot work with categorical data, some cannot work with numerical data, and yet others cannot work either unless the values meet certain criteria. Another important part of this task is data splitting, which is about deciding which part of the data are to be used for model building (training data) and which part for model testing (test data). This step is identified as *data preparation* in CRISP-DM.

Model construction and evaluation: Once an algorithm or data mining technique has been chosen, then there is a need to find a software package that automatically develops a model from the training data. Usually a number of models are developed and their performances are compared. The performance indices may be computed with the training data from which a model has been developed or they may come from new cases (the test data), where the model has been applied and the model output is compared with the known values of the output. Model evaluation or testing is an important step for maximising the amount of information that can be drawn from the dataset. If the model performance is unacceptable, then an iterative path is followed for choosing a different data mining algorithm or having a different set of features from the dataset.

Discovering knowledge: This is the stage where the user can see the benefits of the whole process. The user may come up with previously unknown and interesting knowledge about the dataset that will equip him to make more informed decisions.

In discovering knowledge, it should be kept in mind that data mining can discover patterns, but it cannot tell anything about their significance nor can it tell about the causes.

Taking action: Actions are taken based on the discovered knowledge, which could bring rewards. But taking action can simply mean applying the model to new instances. The direct application of this may occur through the inclusion of new instances into the test dataset of the model development phase. The computed output on the test dataset will provide results. This step is identified as *deployment* in CRISP-DM. Broadly, there are two types of knowledge, shallow and deep. Shallow knowledge is simply what makes up a computer's response. If a data query is framed using SQL, the output will constitute shallow knowledge about the data. For example, it might be learnt that Australian Stock Exchange generally follows the lead of Wall Street, but it would not necessarily be known why. Deep knowledge is the underlying reason behind such relationships. Hidden knowledge is the top layer of this deep knowledge, which normally a data mining technique can unveil. Data mining will not give the causes or the significance, but it can point to various associations and links.

Data-Mining Techniques

This section describes the various data mining techniques that are used in the context of intrusion detection.

Correlation analysis: Finding item set model knowledge frequently appeared from given data set for the purpose of excavating the relationship that was hidden in the data.

Feature selection: A subset of features available from the data is selected for the application of a learning algorithm. It is used in machine learning.

Machine learning: It is the study of computer algorithms which automatically improve through experience.

Sequential patterns: It is used to excavate connection between data, time series analysis gains more focus on the relationship of data in times.

Classification: It is a technique of taking each instance of a dataset and assigning it to a particular class. Typical classification techniques are: inductive rule generation, genetic algorithms, fuzzy logic, neural networks and immunological based techniques.

Clustering: It is a technique for statistical data analysis. It is the classification of similar objects into a series of meaningful subset according to certain rules, so that the data in each subset share some common trait.

Deviation analysis: Finding abnormal data from the database.

Forecast: Finding certain laws according to historical data, establishing models and predicting types, characteristics of the future data, etc based on the model.

Data-Mining Techniques in IDS

Data mining refers to knowledge discovery in database, these data are usually numerous, incomplete, indistinct and random. Data if as original information can be termed as knowledge but generally, knowledge refers to concepts, rules and restraints. With the increase in computerisation and storage of more and more sensitive data on the data servers, the security of the data servers is a major issue. As the intrusion detection systems are being used for monitoring networked devises where they look for the behaviour patterns of various anomalous and malicious behaviours in the audit data. Making comprehensive IDS requires more time and expertise. On the other hand Data mining based IDS require less expert knowledge and give better performance (Barbara *et al.,* 2001; Noel *et al.,* 2002; Eskin *et al.,* 2002; Markou and Singh, 2003).They can generalize new and unknown attacks in a better way. The methods used for finding knowledge can be mathematical or non-mathematical; it can be deductive or inductive. The available knowledge can be used for optimizing enquiry, manage information, control progress and make intellectual decision. Therefore, data mining can be regarded as a crisscross subject which helps people in application of data from low and simple inquiry to discovery knowledge in data and support decision. In the analysis of intrusion detection system, the data circulating in network has the following characteristics: mass data, even if a small commercial website, the number of data message sent and received are quite impressive and incomplete whose transportation is busy, data message which overweigh network carry will be discarded: noisy, when network is unstable, data information may get changed in the transportation of message (Liu, 2009). It can be seen that these data is in accordance with the feature of data mining, naturally data mining need to be applied to intrusion detection system.

The various detection models need log data as training collection whose accuracy will largely influence intrusion detection system. Because of the density and accuracy of visit network, it is difficult to acquire completely no attack action. In addition, it is also uneasy to log attack behavior. Data mining technology can resolve this problem, in the analysis of general network visit; isolated point is an invasive behavior to reduce the difficulty of acquisition of training data. Intrusion detection system is a passive method in the security field, it monitors information system and sends out warning when it does detect intrusion, but data mining technology can analyze these data when network message is acquired, it can forecast for visit on its own initiative, thus reduce the frequency of matching, and achieve the function of active defense. Data mining technology, for instance, Clustering, Classification, Feature Summary, association rules can be applied in the intrusion detection system.

It has been proved that data mining technology improves the property of intrusion detection system, the processing rate and reduces the rate of misreporting.

Conclusion

With the new and new attack methods, the intrusion detection systems should also be updated from time to time. The basic reason for this is that most of the IDS are constructed using manual coding of the expert knowledge. So changing IDS has been a costly affair. However we can manage this manual encoding by making use of a data mining techniques. With the help of these techniques IDS development and updating becomes an adaptive process. In order to accurately capture the behavior of an intrusive activity, it is essential to record the characteristics of that intrusive activity in an auditing program. These auditing programs can contain features that describe each network connections or host sessions. Data mining can automate the process of finding relationships and patterns in raw data, and can deliver results that can be either utilised in an automated decision support system or assessed by a human analyst thus enhancing the quality of the intrusion detection process.

References

1. Anderson, C. 2011. How PDCA cycles are used. Bizmanualz, p. 20.
2. Barbara, D., Couto, J., Jajodia, S., Popyack, L. And Wu, N. 2001. ADAM: Testbed for exploring the use of data miningin intrusion detection, ACM SIGMOD Record, 30(4): 15-24.
3. Berry, M.J.A. and Lino, G. 1997. Data mining techniques. John Wiley and Sons, Inc.
4. Bloedorn, E. 2001. Data mining for network intrusion detection: How to get started. Technical paper.
5. Eskin, E., Arnold, A., Prerau, M., Portnoy, L. and Stolfo, S. J. 2002. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data, In D. Barbara and S. Jajodia (eds.), Applications of data mining in computer security, Kluwer Academic Publishers, Boston, MA, pp. 78-99.
6. Frawley, W., Piatetsky-Shapiro, G. and Matheus, C.1992. Knowledge discovery in databases: An overview. AI Magazine, pp.213-228.
7. Hand, D., Mannila, H. and Smyth, P. 2001. Principles of data mining. MIT Press, Cambridge, MA.
8. Lee, W. and Stolfo, S. J. 1998. Data mining approaches for intrusion detection, In Proc. of the 7th USENIX Security Symp., San Antonio, TX.USENIX.
9. Liu, W. 2009. Research of data mining in intrusion detection system and the uncertainty of the attack. IEEE.
10. Markou, M. and Singh, S. 2003. Novelty detection: A review, Part 1: Statistical approaches. *Signal Proc*. 8(12): 2481-2497.
11. Noel, S., Wijesekera, D. and Youman, C. 2002. Modern intrusion detection, data mining, and degrees of attack guilt. In D. Barbara and S. Jajodia (eds.), Applications of data mining in computer. Aecurity, Kluwer Academic Publishers, Boston, MA, pp. 2-25.
12. Rother, M. 2009. Toyota Kata, McGraw-Hill,U.S., p. 160.
13. Yusufovna, S.F. 2008. Integrating intrusion detection system and data mining. International Symposium on ubiquitous multimedia computing, IEEE, pp. 256-259.