International Journal of Educational Planning & Administration.
Volume 1, Number 1 (2011), pp. 29-35

http://www.ripublication.com/ijepa.htm

Study and Analysis of Data Mining Techniques

[1]Manoj and [2]Jatinder Singh

[1]Ph.D., Research Scholar, Singhania University, India
[2]Dean, Engg., DBIEM, Moga, India

Introduction

Data mining is a rapidly expanding field in many disciplines. It is becoming increasingly necessary to find data mining packages appropriate for a given analysis. Data mining techniques have been advantageously functional in a variety of disciplines including Intrusion detection, manufacturing, process control, fraud detection, marketing, and network management. Over the past few years the data mining techniques are applied to a growing number of research projects to solve a variety of problems in intrusion detection.

Introduction to Data Mining

The relatively new discipline of data mining is most often applied to extraction of useful knowledge from business data however it is also useful in some scientific applications where this more empirical approach complements traditional data analysis. It is an essential ingredient in the more general process of Knowledge Discovery in Databases (KDD). The idea is that by automatically sifting through large quantities of data it should be possible to extract nuggets of knowledge. Data mining has become fashionable not just in computer science (journals & conferences) but particularly in business IT. The emergence is due to the growth in data warehouses and the realization that this mass of operational data has the potential to be exploited as an extension of Business Intelligence.

Objectives of data mining

Data mining techniques are employed with two main objectives:

- To improve our understanding of the relevant factors and their relationships, including the possible discovery of non-obvious features in the data that may suggest better formulations of the physical models.
- To induce models solely from the data so that dynamical simulations might be compared to them and that they may also have utility, offering (short term) predictive power.

Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality data mining technology can generate new business opportunities by providing these capabilities:

- Automated prediction of trends and behaviors.
- Automated discovery of previously unknown patterns.

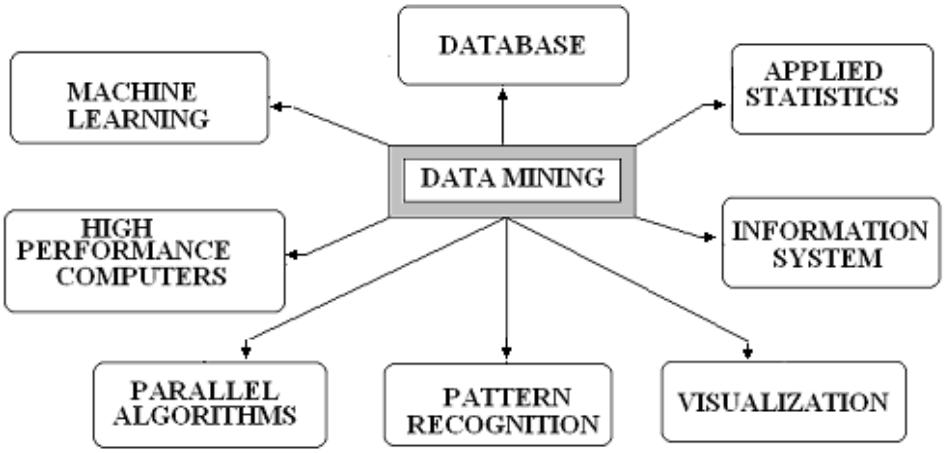

Figure 1: Scope of data mining.

Nature of data mining

'Is data mining as useful in science as in commerce?' is an important question as far as the nature of data mining is to be made clear. Certainly data mining in science has much in common with that for business data. One difference though is that there is a lot of existing scientific theory and knowledge hence there is less chance of knowledge emerging purely from data however empirical results can be valuable in science (especially where it borders on engineering) as in suggesting causality relationships or for modeling complex phenomena. Another difference is that in commerce rules are soft sociological or cultural and assume consistent behavior.

Tools and phases of data mining

To enumerate a precise list of data mining tool characteristics is quite difficult since the tools governing the process of data mining are not standardized. They are not specific and most of the times various approaches and tools result in data mining and it generates families of In spite of the lack of precise standards, we may conclude that data mining is subject to four phases viz.,

- Data preparation
- Data analysis and classification

- Knowledge acquisition
- Prognosis

They are followed one after the other that is in sequence and are clear from the following figure

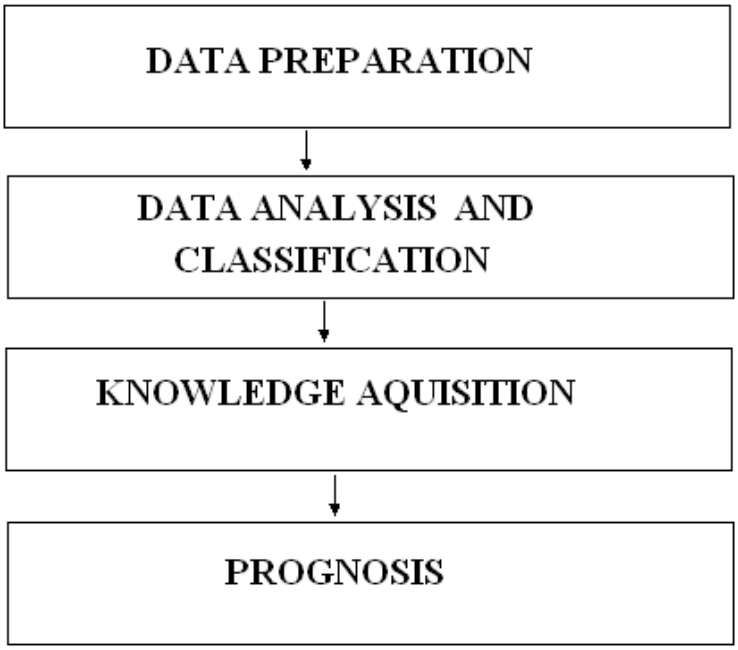

Figure : Tools and phases of data mining.

In the data preparation phase the main data sets to be used by the data mining operation are identified and cleansed of any impurities as data in the data warehouse. As the data warehouses are already integrated and filtered, data warehouse usually is the target set for data mining operation.

Representations

The complete set of findings can be represented in a decision tree, a neural net, a forecasting model or a visual presentation interface that is then used to present future events or results. As data mining technology is still in its infancy many new representation schemes are expected.

Techniques of data Mining

The techniques used for data mining can be broadly categorized into three types as in following figure

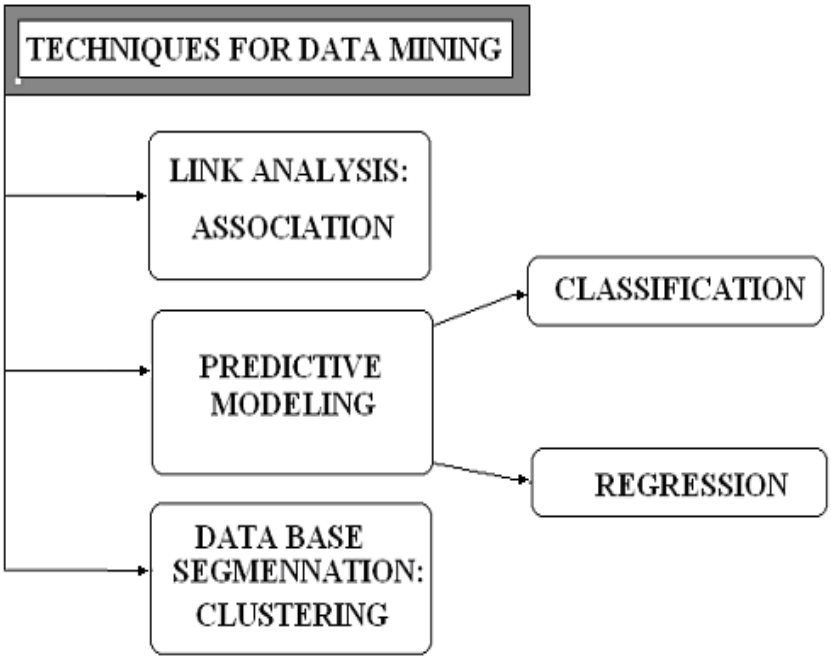

Figure : Techniques for data mining.

Association Rules

This process is usually applied to transactional databases. These are databases where each record represents a transaction, typically some kind of sale. Formally a transaction is defined as follows "given a set of items *I*, each transaction is a subset of the items". For example the items could be all the possible products of a store and the transaction indicates which of these products were bought in a certain occasion by a particular customer. An association rule is defined as a relation: $A \Rightarrow B$, where A and B are both subset of the items. If the rule is a valid one, it indicates that all the records of the database which contain the items of A also contain the items of B. In our example, this could be translated as: "all the customers who bought items #10 and #14 also bought items #19 and #75".

Predictive modeling

It is a kind of supervised learning that uses observations to learn and to predict. It may further be of two types viz., classification and regression.

Database segmentation: clustering

The technique has the focus on making the database segments that is to create clustering. It is also referred to as unsupervised form of learning. It is the technique of forming clusters or segments on the basis of close associations or characteristics. It is

usually performed with intent of capturing gestalt properties of a cluster instead of focusing just on the commonalties.

The process of clustering is somehow similar to classification. Again the purpose is to divide the records of a database in similar homogeneous groups but this time the user does not know the classes before the analysis. The clustering algorithm will have to discover the more natural way to group the records together and then proceed with the grouping. The best application of clustering is with spatial databases. These are databases where each record represents a point in a certain space. The clustering algorithm finds all the points belonging to the same clusters. For example if our database represents the customers of an insurance company they will be clustered according to similar behaviors. An outliner is a customer who has an unusual behavior. This could hide an attempt to defraud the company and could require further investigation. This use of the clustering is called fraud detection.

Approaches to Clustering

There are basically four different approaches to the clustering problem:

- Partition clustering
- Grid-based clustering
- Hierarchical clustering
- Density-based clustering

How Data Mining Works

‘How exactly is data mining able to tell you important things that you didn't know or what is going to happen next?’ is important consideration. The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. For instance if you were looking for a sunken galleon on the high seas the first thing you might do is to research the times when treasure had been found by others in the past. You might note that these ships often tend to be found off the coasts and that there are certain characteristics to the ocean currents and certain routes that have likely been taken by the ship’s captains in that era. You note these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand you sail off looking for treasure where your model indicates it most likely might be given a similar situation in the past. Hopefully if you've got a good model you find your treasure.

References

[1] Kimmo Hiltunen, "WLAN Attacks and Risks", White Paper, Ericson, January 2008.

[2] Mazda Salmanian, Julie H. Lefebvre, Steve Leonard and Scott Knight, “Intrusion Detection in 802.11 Wireless Local Area Networks”, Technical Memorandum, Defence R&D Canada & Ottawa , July 2004
[3] H.BELLAAJ, R.KETATA and A.HSINI, "Fuzzy approach for 802.11wireless intrusion detection", in proc. of 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, March 25-29, Tunisia, 2007.
[4] Zonghua Zhang and Hong Shen,” A Brief Observation-Centric Analysis on Anomaly-Based Intrusion Detection”, Springer-Verlag Berlin Heidelberg 2005
[5] Mark Handley and Vern Paxson “Network Intrusion Detection: Evasion, Traffic Normalization and End-to-End Protocol Semantics”,
[6] Martin Rehak, Michal pechoucek, karel Bartos, Martin Grill, Pavel celeda and vojtech krmick “An intrusion detection system for high-speed networks”, national institute of informatics, 2008
[7] John Haggerty, Qi Shi and Madjid Merabti,” STATISTICAL SIGNATURES FOR EARLY DETECTION OF FLOODING DENIAL-OFSERVICE ATTACKS”, Springer Boston, 2006
[8] Giovanni Vigna, Sumit Gwalani, Kavitha Srinivasan, Elizabeth M. Belding-Royer and Richard A. Kemmerer ,” An Intrusion Detection Tool for AODV-based Ad hoc Wireless Networks”, IEEE Computer Society Washington, DC, USA ,2004.
[9] Shukor Abd Razak, Steven Furnell, Nathan Clarke, and Phillip Brooke,” A Two-Tier Intrusion Detection System for Mobile Ad Hoc Networks – A Friend Approach”, Springer-Verlag Berlin Heidelberg 2006
[10] Eduardo Mosqueira-Rey, Amparo Alonso-Betanzos, Belen Baldonedo Del Rio, and Jesus Lago Pineiro, ” A Misuse Detection Agent for Intrusion Detection in a Multi-agent Architecture”. Springer-Verlag Berlin Heidelberg 2007
[11] Magnus Almgren, Ulf Lindqvist, and Erland Jonsson,” A Multi-Sensor Model to Improve Automated Attack Detection”, Springer-Verlag Berlin Heidelberg 2008
[12] Curtis A. Carver, Jr., Jeffrey W. Humphries, and Udo W. Pooch,”Adaptation Techniques for Intrusion Detection and Intrusion Response Systems”,
[13] Naeimeh Laleh and Mohammad Abdollahi Azgomi,” A Taxonomy of Frauds and Fraud Detection Techniques”, Springer-Verlag Berlin Heidelberg 2009.
[14] Jeyanthi Hall Michel Barbeau and Evangelos Kranakis, “Detecting Rouge Devices In Bluetooth Networks Using Radio Frequency Fingerprinting” ,School Of computer Science, Carleton University.
[15] Jeyanthi Hall , Michel Barbeau and Evangelos Kranakis, “Enhancing Intrusion Detection In Wireless Networks Using Radio Frequency Fingerprinting”, School Of computer Science, Carleton University.
[16] M. Charikar, E. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, A. Shelat, “The smallest grammar problem,” IEEE Transactions on Information Theory, vol. 51, Issue 7, pp. 2554-2576, July 2005.
[17] Keesook J. Han and Joseph Giordano, “Intrusion Detection System Modeling,” Proceedings of the HPCMP UGC 2006, IEEE Computer Society, June 2006.

[18] M. Corley, M. Weir, K. Nelson, A. Karam, “Simplified Protocol Capture (SIMPCAP),” Proceedings of the Fifth Annual IEEE Information Assurance Workshop, 2004.

[19] J. C. Kieffer and E.-H. Yang, “Grammar-Based Codes: A New Class of Universal Lossless Source Codes,” IEEE Transactions on Information Theory, vol.46, pp.737-754, 2000.

[20] J. C. Kieffer and E.-H. Yang, “Grammar-Based Lossless Universal Refinement Source Coding," IEEE Transaction on Information Theory, vol. 50, pp.1415-1424, 2004.

[21] J. Dricot and Ph. De Doncker. High-accuracy physical layer model for wireless network simulations in ns-2. In Proc. of the Int. Workshop on Wireless Ad-hoc Networks, IWWAN’04, Oulu, Finland, May-June 2004.