Information Security Issues in Big Data and Data Mining

M.Gomathi#1, V.Sujatha#2

[1]*Master of Computer Applications, S.A. Engineering college, Chennai-77.*
gomathimbca@gmail.com
[2]*Asst Prof., Department of Computer Applications, S.A. Engineering College, Chennai-77.*
sujatha@saec.ac.in

Abstract— The growing popularity of data mining technologies bring serious threat to security of individuals sensitive material differentiating the responsibilities of different user with respect to security of sensitive material. Venture data mining appliances frequently engage multifaceted data such as numerous big various data basis, user penchants and trade collision. These kinds of difficulties confuse the analysis of multi channel data as compared to the analysis of single-channel data. We sight the solitude subjects to data mining from a broader viewpoint and believe the different advances that can assist to defend receptive in sequence. The need for deep accepting of customer and customer government duplications through advanced data analytics has been increasingly accepted by the community at large.

Keywords— Multimedia; Multisource; Data Mining; Sensitive Information; Privacy Preserving Data Mining.

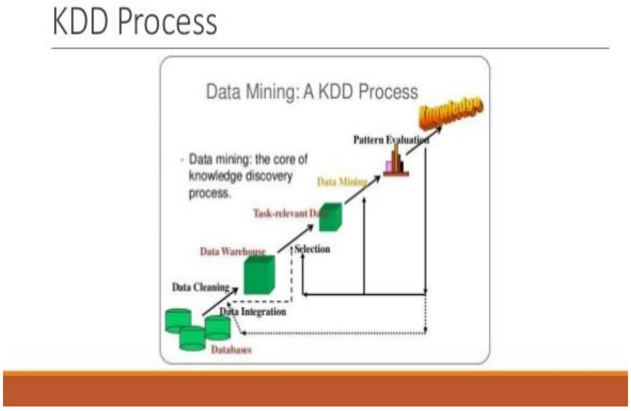
1. Introduction

Data mining has more and more attention in recent years, possibly because of the popularity of the big data model. Data mining is semi-automatic discovery of patterns, associations, alterations, anomalies and statistically important structure and events of data. Sometimes referred to as computer security material technology often some form of computer system. Since it normally paste faster for programs to access data stored locally and data across a cluster and also import attentions which must be made when about big data complications. Social security data mining (SSDM) seeks to discover remarkable patterns and exclusions on social security and social welfare data. The technological revolution has made large economy memory spaces and easy to acquire.

1.1 ` The Process of KDD

The data mining is a group preserved as a meaning for an additional word "Knowledge discovery data" (KDD) which is an routine, probing study and sculpting of big data repositories. KDD is the prearranged procedure of recognize suitable, new, practical, and explicable patterns from big and compound data sets.

The knowledge discovery process (figure 1) is iterative and interactive, consisting of steps. Note that the process is iterative at each step, meaning that moving back to previous steps may be required. The methods have lots of "artistic" features in the intellect that one cannot in attendance one recipe or build a whole nomenclature for the correct options for every step and request kind.

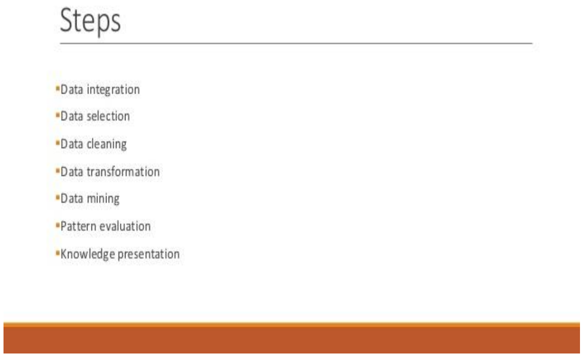

KDD Process

Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.

Fig.1: An overview of the KDD process

Knowledge in an case-to-understand fashion. Truly interesting patterns which represent knowledge in the data mining situation described the whole KDD process involve multi stage options. In this paper we develop a user role based methodology to comportment to review the related studies. In the data mining scenario depicted user represents either somebody or an organization. Detailed discussions will be presented in divisions.


Steps

- Data integration
- Data selection
- Data cleaning
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation

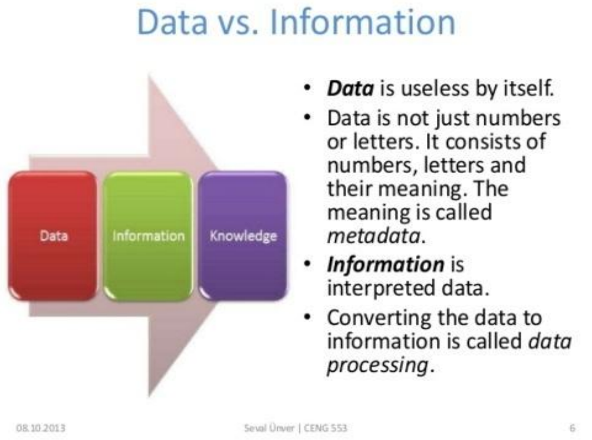
Fig. 2: The KDD Steps

Step1: Data processing effortless operations include data selection, data cleaning (to remove noise) and combination of many sources.

Step2: Data transformation goal is to transform data into forms applicable for the mining task that is to find handy features to characterize the data.

Step3: Data mining is an important process where intellectual methods are employed to abstract data patterns.

Step4: Pattern assessment and appearance. fundamental processes comprise recognized the really outstanding prototypes which distinguish information and presenting the mining.


Data vs. Information

- ***Data*** is useless by itself.
- Data is not just numbers or letters. It consists of numbers, letters and their meaning. The meaning is called *metadata*.
- ***Information*** is interpreted data.
- Converting the data to information is called *data processing*.

08.10.2013 Seval Ünver | CENG 553 6

Data
Information
Knowledge

Fig.3: The data information

1.2 Big Data

Big data is not approximately the dimension of the data it is concerning the worth within the data. Big-data' is comparable to 'Small-data', but superior but having data better therefore necessitates diverse advances systems, apparatus & architectures to resolve innovative troubles and old harms in a improved method.

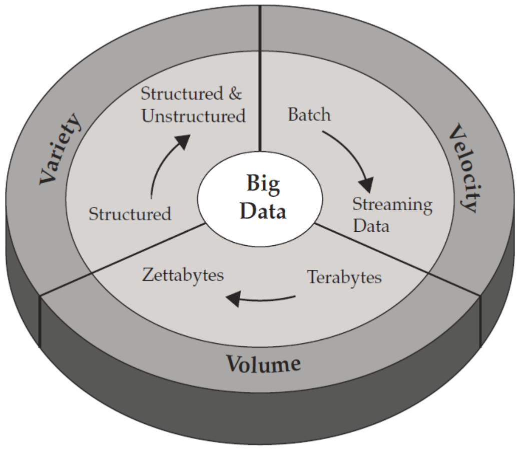


Fig.4: Demonstration of Big Data

1.3 Data Mining Confronts

- Big data mining has a independent platform.
- Big data has its own semantics and information.
- Big data mining algorithm.

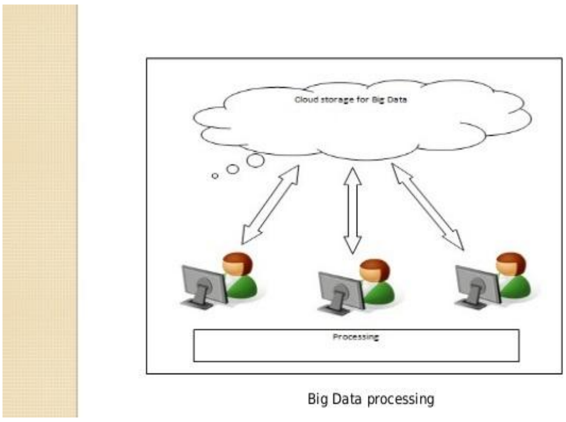


Fig.5: Big data processing

1.4 Information Security in Big Data Privacy and Data Mining

By separating the four different user roles, we preserve the privacy issues in data mining in a just way. All user care about the safekeeping of sensitive information, but each customer role views the security issues from own viewpoint. If the data source considers his data to be very sensitive, that is, the data may disclose some in series that he does not want anyone else to know, the supplier can just refuse to provide such data. Effective contact control are preferred by the data provider, so that he can inhibit his sensitive data from being stolen by the data antenna. Big data mining algorithm. By ``passive" we mean that the data, which are produced by the provider's routine accomplishments. The data provider may even have no awareness of the discovery of his data.

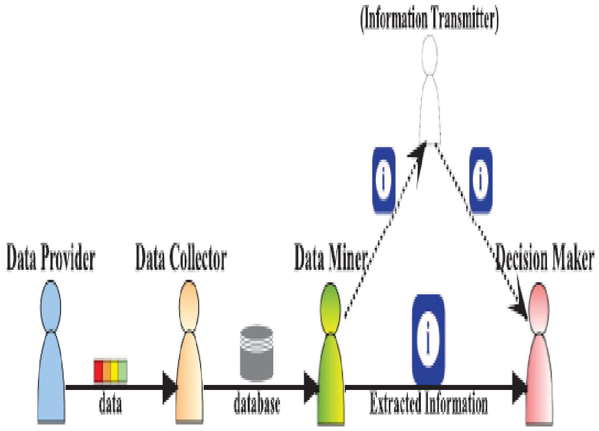


Fig.6: A Simple illustration of the application scenario with data mining at the core.

1.5 Agent Mining The Synergy Of Agents And Data Mining

Agent driven data processing educations issues associated with the removal, collection, and executive of data created by the actions of agents in MAS. Together pastures face dangerous confronts that the other knowledge capacity surrogate**.**

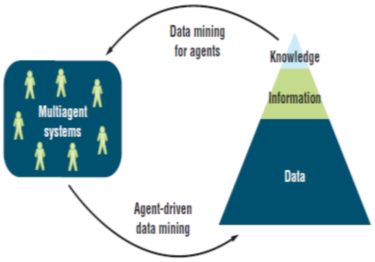


Fig.7: The synergy of agents and data mining

1.6 Processing Of Multichannel Recordings For Data Mining Algorithms

A major sub group of analysing performances deals with data style that the behaviour of an object over time. This type of data materializes dynamic information about how the behaviour of the object changes as time goes by. A chief subgroup of examines measures contracts with data that portray the performance of a thing over time. This kind of data materializes lively in sequence concerning how the actions of the thing adapt as occasion goes by. Hence, analysis of such data could be very useful for education trends about the features of the object.

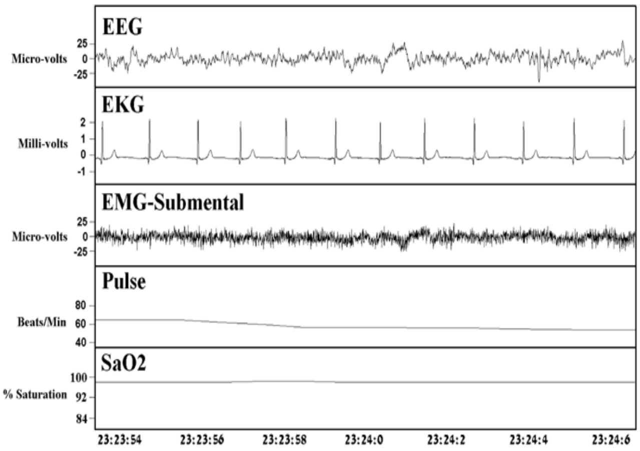


Fig.8: Multichannel data recorded in a sleep laboratory

2. Existing System

With the enhanced structures in recent computer systems gradually larger amounts of data are being accumulated in various fields. Data mining is a powerful tool that enables to achive data mining courses which have become important tools in many domains including business promotion, client association managing and deception finding. Existing data mining tools are not able to run resourcefully on existing system.

3. Proposed System

E-learning is one of the tools used in the knowledge management to share the knowledge among collections. The data mining and proposed new procedures applied for direct or indirect decrimination avoidance separately or both at the same time.

4. Methodology

4.1 Supervised

- Problem solving.
- Driven by a real business complications and historical data.
- Quality of results helpless on quality of data.

4.2 Unsupervised

- Exploration (aka clustering).
- Relevance often an issue.
- Useful when trying to get an initial accepting of the data.
- Non-obvious pattern can sometimes be popular out of a complete data analysis project.

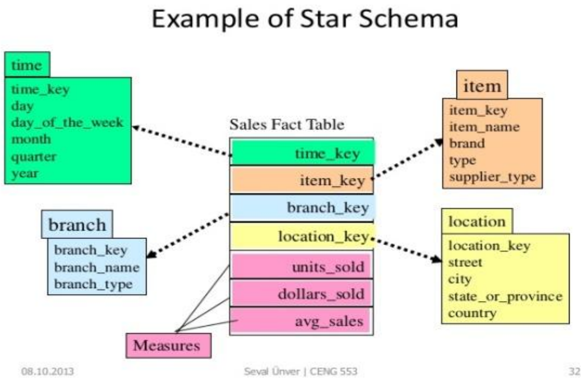


Fig.9: The star schema

4.3 Supervised Algorithm Summary

A. Decision Trees

- Understandable.
- Relatively fast.

- Easy to translate into SQL queries.

B. KNN

- Quick and easy.
- Models incline to be very large.

C. Neural Networks

- Difficult to interpret.
- Can require significant amounts of time to train.

5. Result

The advanced data mining with all these features will help the supervisor to get a clear understanding if the data mining being large where the amount of administrator interval will be saved with the use of feature similar security material. With such feature, security issues related to the malicious tradition of the bandwidth can be traced.

6. Conclusion

Order to be valid data mining, a large amount of quality data is essential. The aim of data mining is attaining rules and comparisons which can be used to expect better results. To be popular on such work is to be dependent on working with database experts and data mining consultants. They need to work together. Work may take longer, but we need to have time and persistence.

7. Future Enhancement

We believe our future work will material security trends and issues. The characteristics defined in NCSA will be used in future for research on security issues and these entities must then be arranged to secure through security mechanisms to provide both hackers and competitors.

References

[1] K. Nissim, S. Vadhan, and D. Xiao, ``Redrawing the boundaries onpurchasing data from privacy-sensitive individuals", in Proc. 5th Conf.Innov. Theoretical Comput. Sci., 2014, pp. 411-422.

[2] M. Kantarcioglu and W. Jiang, ``Incentive compatible privacy-preserving data analysis", IEEE Trans. Knowl. Data Eng., Vol. 25, No. 6, Jun. 2013, pp. 1323_1335.

[3] A. Symeonidis and P. Mitkas, "Agent Intelligence through Data Mining", Springer, Vol.7,2005, pp. 23-25.

[4] P. A. Diamond and P. R. Orszag, "Saving Social Security: A Balanced Approach", Washington, DC: Brookings Institution, Vol.8, 2005, p.86

[5] P. A. Diamond, "Taxation, Incomplete Markets, and Social Security", Cambridge, MA: The MIT Press, Vol.2, 2003, pp.65

[6] B. Mc Namara and K.Wiesenfeld, " Theory of Stochastic Resonance", Phys. Rev. A, vol. 39, 1989, pp. 4854– 4854.

[7] H. Mannila, H. Toivonen and A. I. Verkamo, "Discovering frequentepisodes in sequences", in Proc. Knowledge Discovery and Data Mining (KDD), 1995, pp. 210–212.

[8] Jyothi Mandala and Suneetha Merugula, "Review of Privacy Preserving Data Mining Techniques", Engineering and Scientific International Journal, Volume 2, Issue 1, January - March 2015, pp.1-3.

[9] A.S.Aneeshkumar and Dr.C.Jothi Venkateswaran, "Estimating the Surveillance of Liver Disorder using Classification Algorithms", International Journal of Computer Applications, Volume 57– No.6, November 2012, pp.39-42.