DATA MINING: CHALLENGES & FUTURE SCOPE

[1]Neha Mittal , [2]Satvika, [3]Deepika Garg
[1,2,3]M.Tech Scholar ,TITS , Bhiwani
[3]deepugarg22@gmail.com

ABSTRACT : From the last two decades data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. All most every field of human life has become data-intensive, which made the data mining as an essential component. This paper traces some of the major challenges in the field of data mining such as mining complex knowledge from complex data, sequential and time series etc. It reviews approaches adopted for this problem and it identifies challenges and points out future directions in this relatively new field.

Keywords - Data mining, i.i.d.(independent and identically distributed), Knowledge discovery, predictive

I. INTRODUCTION

The discovery of knowledge (in the form of rules, trees, frequent patterns etc.) from large volumes of data. The data collected from different applications require proper mechanism of extracting knowledge information from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data.

II. OPPORTUNITIES & CHALLENGES

a)Developing a Unifying Theory of Data Mining

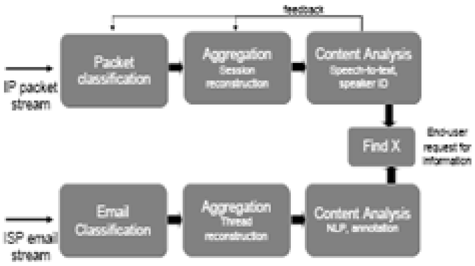
- The current state of the art of data-mining research is too ``ad-hoc“ techniques are designed for individual problems
- no unifying theory
- Needs unifying research
- Exploration vs explanation
- Long standing theoretical issues
- Deep research.
- Knowledge discovery on hidden causes?
- Similar to discovery of Newton’s Law?

b)Scaling Up for High Dimensional Data and High Speed Streams

- Scaling up is needed
- ultra-high dimensional classification problems
- Ultra-high speed data streams
- Streams
- continuous, online process
  e.g. how to monitor network packets for intruders

c)Sequential and Time Series Data

- How to efficiently and accurately cluster, classify and predict the trends ?
- Time series data used for predictions are contaminated by noise.
- Signal processing techniques introduce lags in the filtered data, which reduces accuracy
- Key in source selection, domain knowledge in rules, and optimization methods



d)Mining Complex Knowledge from Complex Data

- Mining graphs.
- Data that are not i.i.d. (independent and identically distributed).
- many objects are not independent of each other, and are not of a single type.
- mine the rich structure of relations among objects.
- E.g. interlinked Web pages, social networks, metabolic networks in the cell.
- Integration of data mining and knowledge inference.
- The biggest gap: unable to relate the results of mining to the real-world decisions they affect - all they can do is hand the results back to the user.

e)Data Mining in a Network Setting

- Community and Social Networks.
- Linked data between emails, Web pages, blogs, citations, sequences and people.
- Static and dynamic structural behaviour.
- Mining in and for Computer Networks.
- Detect anomalies (e.g., sudden traffic spikes due to a DoS (Denial of Service) attacks.
- Need to handle 10Gig Ethernet links
  (a) detect
  (b) trace back
  (c) drop packet

Fig 2 Picture from Matthew Pirretti’s slides, Penn State

An Example of packet streams (data courtesy of NCSA, UIUC)

f)Distributed Data Mining and Mining Multi-agent Data

- Need to correlate the data seen at the various probes (such as in a sensor network)

- Adversary data mining: deliberately manipulate the data to sabotage them (e.g., make them produce false negatives)
- Game theory may be needed for help.

g)Data Mining for Biological and Environmental Problems

- Large scale problems especially so Biological data mining, such as HIV vaccine design DNA, chemical properties, 3D structures, and functional properties need to be fused Environmental data mining.
- Mining for solving the energy crisis

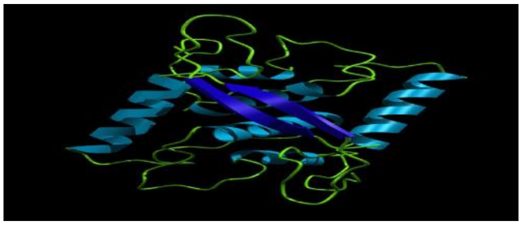


Fig 3 Biological Mining

h)Data-mining-Process Related Problems

- How to automate mining process?
- the composition of data mining operations.
- Data cleaning, with logging capabilities.
- Visualization and mining automation.
- Need a methodology: help users avoid many data mining mistakes.

i)Security, Privacy and Data Integrity

- How to ensure the users privacy while their data are being mined?
- How to do data mining for protection of security and privacy?
- Knowledge integrity assessment.
- Data are intentionally modified from their original version, in order to misinform the recipients or for privacy and security
- Development of measures to evaluate the knowledge integrity of a collection of Data Knowledge and patterns

J)The business case for data mining

Market-leading companies such as Capital One, Amazon.com, Google, and Netflix rely on data mining to drive **significant competitive advantage**. Here are just some of the benefits that are contributing to their bottom line:
using data mining across 140 different organizations. Source: Gartner 2010.

Despite all of the benefits, the process of **data mining**, **analytics**, and **prediction** presents a number of challenges to most businesses today:

- **Data Volume -** Most companies have large amounts of data.
- **Data Growth -** Most companies are seeing data growth exploding at exponential rates.
- **Value Lost -** Most know there is value in the data based on what they see competitors doing.
- **Lack of Knowledge -** Most do not have the knowledge of the specialized technologies and techniques that are used in data mining.
- **Lack of Experience -** Most lack experience in applying data mining results to business problems.
- **Budget -** Most do not have the budget for either expensive software and hardware or the skilled data miners to utilize them.

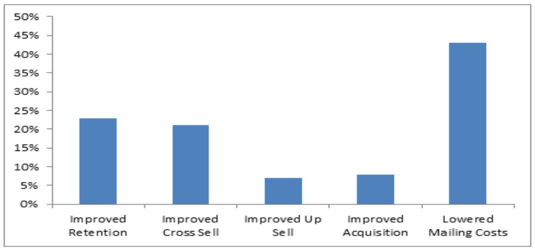
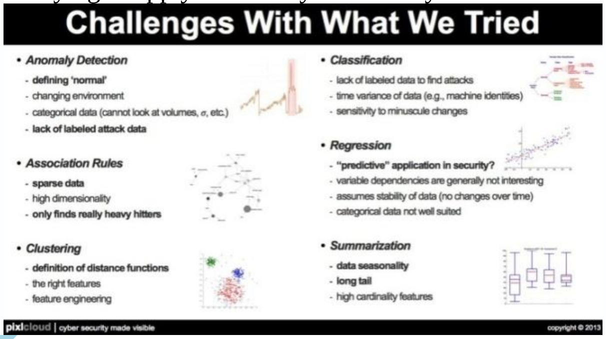


Figure 4 – Indicates frequency of this benefit claimed in project

k)How analytics enables security visualization?

For each of the six areas in data mining, the following slide shows a couple of challenges that one will run into when trying to apply them to cyber security data:



Challenges With What We Tried

- *Anomaly Detection*
  - defining 'normal'
  - changing environment
  - categorical data (cannot look at volumes, σ, etc.)
  - lack of labeled attack data
- *Association Rules*
  - sparse data
  - high dimensionality
  - only finds really heavy hitters
- *Clustering*
  - definition of distance functions
  - the right features
  - feature engineering
- *Classification*
  - lack of labeled data to find attacks
  - time variance of data (e.g., machine identities)
  - sensitivity to minuscule changes
- *Regression*
  - "predictive" application in security?
  - variable dependencies are generally not interesting
  - assumes stability of data (no changes over time)
  - categorical data not well suited
- *Summarization*
  - data seasonality
  - long tail
  - high cardinality features

pixlcloud | cyber security made visible

copyright © 2013

**l)Dealing with Non-static,
Unbalanced and Cost-sensitive Data**

- The UCI datasets are small and not highly unbalanced
- There is much information on costs and benefits, but no overall model of profit and loss
- Data may evolve with a bias introduced by sampling.

III. FUTURE TRENDS

Due to the enormous success of various application areas of data mining, the field of data mining has been establishing itself as the major discipline of computer science and has shown interest potential for the future developments. Ever increasing technology and future application areas are always poses new challenges and opportunities for data mining, the typical future trends of data mining includes:

- Standardization of data mining languages
- Data preprocessing
- Complex objects of data
- Computing resources
- Web mining
- Scientific Computing
- Business data

1. Standardization of data mining languages

There are various data mining tools with different syntaxes. Hence it is to be standardized for making convenient of the users. Data mining applications has to concentrate more in standardization of interaction languages and flexible user interactions.

2. Data Preprocessing

To identify useful novel patterns in distributed, large, complex and temporal data, data mining techniques has to evolve in various stages. The present techniques and algorithms of data pre processing stage are not up to the mark compared with its significance in finding out the novel patterns of data. In future, there is a great need of data mining applications with efficient data pre processing techniques.

3. Complex object of data

Data mining is going to penetrate in all fields of human life, the presently available data mining techniques are restricted to mine the traditional forms of data only. In future there is a potentiality for data mining techniques for complex data objects like high dimensional, high speed data streams sequence, noise in the time series, graph, Multi-instance objects, Multi-represented objects and temporal data.

4. Computing Resources

The contemporary developments in high speed connectivity, parallel, distributed, grid and cloud computing has posed new challenges for data mining. The high speed internet connectivity has posed a great demand for novel and efficient data mining techniques to analyze the massive data which is captured of IP packets at high link speeds in order to detect the Denial of Service (DoS) and other types of attacks. Distributed data mining applications demand new alternatives in different fields, such as discovery of universal strategy to configure a distributed data mining, data placement at different locations, scheduling, resource management, and transactional systems etc. New data mining techniques and tools are needed to facilitate seamless integration of various resources in grid based environment. Moreover, grid based data mining has to focus seriously to address the data privacy, security and governance.Cloud computing is a great area to be focused by data mining, as the Cloud computing is penetrating more and more in all ranges of business and scientific computing.

5. Web mining

The development of World Wide Web and its usage grows, it will continue to generate ever more content, structure, and usage data and the value of Web mining will keep increasing. Research needs to be done in developing the right set of Web metrics, and their measurement procedures, extracting process models from usage data, understanding how different parts of the process model impact various Web metrics of interest, how the process models change in response to various changes that are made-changing stimuli to the user, developing Web mining techniques to improve various other aspects of Web services, techniques to recognize known frauds and intrusion detection.

6. Scientific Computing

In recent years data mining has attracted the research in various scientific computing applications, due to its efficient analysis of data, discovering meaningful new correlations, patterns and trends with the help of various tools and techniques. More research has to be done in mining of scientific data in particular approaches for mining astronomical, biological, chemical, and fluid dynamical data analysis. The ubiquitous use of embedded systems in sensing and actuation environments plays major impending developments in scientific computing will require a new class of techniques capable of dynamic data analysis in faulty, distributed framework. The research in data mining requires more attention in ecological and environmental information analysis to utilize our natural environment and resources. Significant data mining research has to be done in molecular biology problems.

7. Business Trends

Business data mining needs more enhancement in the design of data mining techniques to gain significant advantages in today’s competitive global market place (E-Business). The Data mining techniques hold great promises for developing new sets of tools that can be used to provide more privacy for a common man, increasing customer satisfaction, providing best, safe and useful products at reasonable and economical prices, in today’s E-Business environment.

IV.CONCLUSION

Since its conception in the late 1980s, data mining has achieved tremendous success. Many new problems have emerged and have been solved by data mining researchers. The problem of collecting data, which used to be a major concern for most organizations, is almost resolved. However, there is still a lack of timely exchange of important topics in the community as a whole. In this paper we briefly reviewed the various data mining trends from its inception to the future. This review would be helpful to researchers to focus on the various issues of data mining. In future course, we will review the various classification algorithms and significance of evolutionary computing (genetic programming) approach in designing of efficient classification algorithms for data mining. We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before. Everybody is warmly invited to participate in this intrepid journey.

REFERENCES

[1] Heikki, Mannila. 1996. Data mining: machine learning, statistics, and databases, *IEEE Piatetsky-Shapiro*, Gregory. 2000.

[2] The Data-Mining Industry Coming of Age. IEEE Intelligent Systems. Han, J., & Kamber, M. 2001. Data mining: Concepts and techniques .Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press.

[3] Baker, S. and Baker, K. (1998), ``Mine over matter’’, Journal of Business Strategy, Vol. 19 No. 4, pp. 22-7.

[4] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), ``From data mining to knowledge discovery: an overview’’, in Fayyad, U.

[5] Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA.

[6] The Free Dictionary, by Farlex Inc. © 2010, Available at:http://www.thefreedictionary.com/data+mining (accessed on 6th January 2010).

[7] Folorunso O and Ogunde A O, Data Mining as a technique for knowledge management in business process redesign, The Electronic Journal of Knowledge Management, 2 (1) (2004) 33-44, Available online at: www.ejkm.com (accessed on 8th January