Major Research Challenges in Data Mining

AnnanNaidu Paidi,
Assistant Professor, CSE Department,
Centurion University, Orissa, India.

Abstract: The Government, Corporate and industrial communities are faced with an ever increasing number of databases. These databases need not only be managed, but also explored. Such tremendous amount of data, in the order of tera-to peta-bytes, has fundamentally changed science and engineering, transforming many disciplines from data-poor to increasingly data-rich, and calling for new, data-intensive methods to conduct research in data mining.

This Paper presents the major research challenges in data mining with a focus on the following issues: Design classifiers to handle ultra-high dimensional classification problem, Mining data streams in extremely large database, Mining complex knowledge from complex data, Mining across multiple heterogeneous data sources: Multi database and multi relational mining, Mining Non-Relational data, Automate Data cleaning, Privacy preserving data mining.

Key words: *Data mining, automated data cleaning, data mining process, data mining research challenges.*

1. INTRODUCTION

With the rapid development of computer and information technology in the last several decades, an enormous amount of data in science and engineering has been and will continuously be generated in massive scale, either being stored in massive storage devices or owing into and out of the system in the form of data streams.

Data mining attempts to implement basic processes that facilitate the extraction of meaningful information and knowledge from unstructured data. Data mining[8] extracts patterns, changes, associations and anomalies from large data sets. The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data.

The two "high-level" primary goals of data mining, in practice, are *prediction* and *description*.

1. **Prediction** involves using some variables or fields in the database to predict unknown or future values of other variables of interest.
2. **Description** focuses on finding human-interpretable patterns describing the data.

A) Steps in Data Mining

The following steps are usually followed in data mining. These steps are iterative, with the process moving backward whenever needed.

1. Develop an understanding of the application, relevant prior knowledge, and the end user’s goals.
2. Create a target data set to be used for discovery.
3. Clean and pre-process data (including handling missing data fields, noise in the data, accounting for time series and known changes).
4. Reduce the number of variables and find invariant representations of data if possible.
5. Choose the data mining task (classification, regression, clustering, etc.)
6. Choose the data mining algorithm.
7. Search for patterns of interest (this is the actual data mining).

8. Interpret the pattern mined. If necessary, iterate through any of steps 1 through 7.
9. Consolidate knowledge discovered and prepare a report.

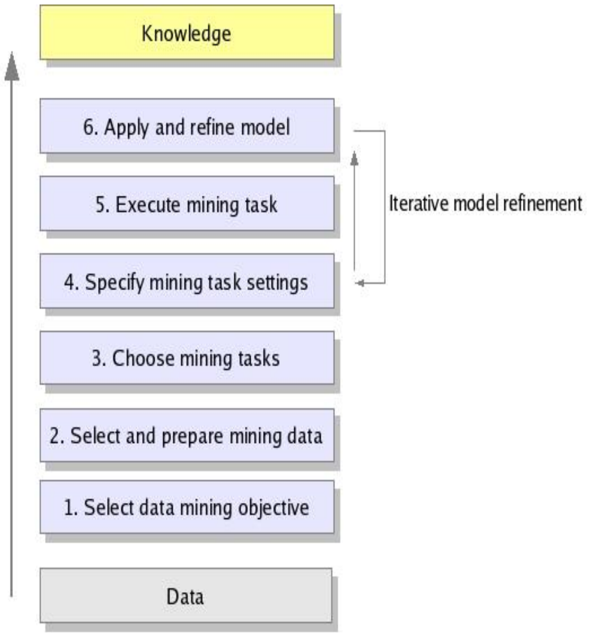
Knowledge
6. Apply and refine model
5. Execute mining task
4. Specify mining task settings
3. Choose mining tasks
2. Select and prepare mining data
1. Select data mining objective
Data
Iterative model refinement

Figure1: Data mining Steps

B) Data Mining Process

Data Mining is an iterative process that uses a variety of data analysis tools to discover patterns and relationships in data. Data mining is an interactive and iterative process involving data pre-processing[9], search for patterns, knowledge evaluation, and possible refinement of the process based on input from domain experts or feedback from one of the steps. The pre-processing of the data is a time-consuming, but critical, first step in the data mining process. It is often domain and application dependent; however, several techniques developed in the context of one application or domain can be applied to other applications and domains as well. The pattern recognition step is usually independent of the domain or application.

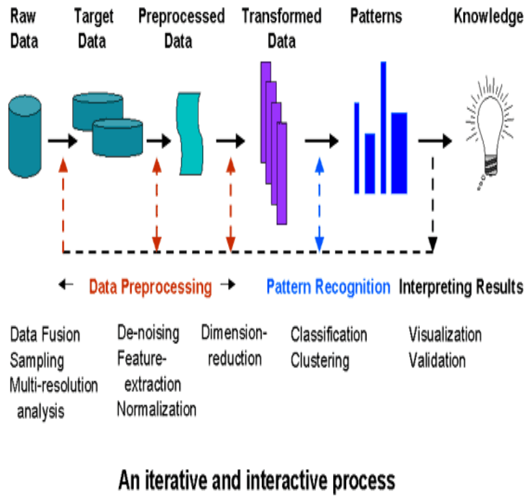

Figure2: Data Mining Process

2. RECENT RESEARCH ACHIEVEMENTS

A) Distributed Data Mining

Distributed computing [2] plays an important role in the Data Mining process for several reasons. First, Data Mining often requires huge amounts of resources in storage space and computation time. To make systems scalable, it is important to develop mechanisms that distribute the work load among several sites in a flexible way. Second, data is often inherently distributed into several databases, making a centralized processing of this data very inefficient and prone to security risks. Distributed Data Mining [5] explores techniques of how to apply Data Mining in a non-centralized way. Distributed data mining can be used in parallel super-computers, P2P networks, and sensor networks. However, different environments have different concerns.

B) Neural Networks

Neural networks [3] have emerged as advanced data mining tools in cases where other techniques may not produce satisfactory predictive models. As the term implies, neural networks have a biologically inspired modelling capability, but are essentially statistical modelling tools. In more practical terms Neural Networks are non-linear statistical data modelling tools[5]. They can be used to model complex relationships between input and output or to find patterns in data.

C) Data Visualisation

In information or data visualization [4], the data usually consists of a large number of records each consisting of a number of variables or dimensions. Each record corresponds to an observation, measurement, transaction, etc. Examples are customer properties, e-commerce transactions, and physical experiments. The number of attributes can differ from data set to data set.
Visualization tools[4] go beyond the standard charts and graphs used in Excel_spreadsheets, displaying data in more sophisticated ways such as dials and gauges, geographic maps, time-series charts, heat maps, tree maps and detailed bar, pie and fever charts. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

D) Fraud Detection Using Outliers

Outlier detection [2] has been used for centuries to detect and, where appropriate, remove anomalous observations from data. Outliers arise due to mechanical faults, changes in system behaviour, fraudulent behaviour, human error, instrument error or simply through natural deviations in populations. Their detection can identify system faults and fraud before they escalate with potentially catastrophic consequences.
An outlier can denote an anomalous object in an image such as a land mine. An outlier may pinpoint an intruder inside a system with malicious intentions so rapid detection is essential. Outlier detection[5] can detect a fault on a factory production line by constantly monitoring specific features of the products and comparing the real-time data with either the features of normal products or those for faults.

E) Large scale optimization

Some data mining algorithms can be expressed as large-scale often non-convex, optimization problems [3]. Recent work has provided distributed methods for large scale continuous and discrete optimization problems, including heuristic search for problems too large to be solved exactly.

F) C-TREND(Cluster-based Temporal Representation of EveNt Data)

C-TREND[6] separates data into user-defined partitions based on time periods and then identifies clusters of the dominant transaction types occurring within each partition. Clusters are then compared to the clusters in adjacent time periods to identify cross-period similarities and, over many time periods, trends are identified. Trends[7] are presented in an output graph that uses nodes to represent dominant transaction types and edges to represent cross-time relationships. It provides the user with the ability to generate graphs from data and adjust the graph parameters.

3. MAJOR RESEARCH CHALLENGES

A) Design classifiers to handle ultra-high dimensional classification problem

One challenge is how to design classifiers to handle ultra-high dimensional classification [2] for text mining and drug safety applications. A new design procedure for a hybrid decision tree classifier which improves the classification efficiency and accuracy for classifying high-dimensional data[5] with a small training sample size.

B) Mining data streams in extremely large database

One important problem is mining data streams in extremely large databases [2](e.g. 100 TB). Satellite and computer network data [3] can easily be of this scale. However, today's data mining technology is still too slow to handle data of this scale. In addition, data mining should be a continuous, online process, rather than an occasional one-shot process. Organizations that can do this will have a decisive advantage over ones that do not. Data streams present a new challenge for data mining researchers.

C) Mining complex knowledge from complex data

One important type of complex knowledge is in the form of graphs[5]. Recent research has touched on the topic of discovering graphs and structured patterns from large data, but clearly, more needs to be done. Another form of complexity is from data that are non-i.i.d. (independent and identically distributed). This problem can occur when mining data from multiple relations. In most domains, the objects of interest are not independent of each other, and are not of a single type. We need data mining systems that can soundly mine the rich structure of relations among objects, such as interlinked Web pages, social networks, metabolic networks in the cell, etc.

D) Mining across multiple heterogeneous data sources: Multi database and multi relational mining

The problem of distributed data mining[2] is very important in network problems. Ina distributed environment (such as a sensor or IP network), one has distributed probes placed at strategic locations within the network. The problem here is to be able to correlate the data seen at the various probes, and discover patterns in the global data seen at all the different probes. There could be different models of distributed data mining here, but one could involve a NOC that collects data from the distributed sites, and another in which all sites are treated equally. The goal here obviously would be to minimize the amount of data shipped between the various sites essentially, to reduce the communication over head. In distributed mining, one problem is how to mine across multiple heterogeneous data sources: multi-database and multi-relational mining [5].

E) Mining Non-Relational data

Yet another important problem is how to mine non-relational data[3]. A great majority of most organizations' data is in *text form*, not databases, and in more complex data formats including Image, Multimedia, and Web data. Thus, there is a need to study data mining methods that go beyond classification and clustering [5]. Some interesting questions include how to perform better automatic summarization of text and how to recognize the movement of objects and people from Web and Wireless data logs in order to discover useful spatial and temporal knowledge.

F) Automate Data cleaning

Data cleaning, also called *data cleansing* or *scrubbing [3]*, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly.

G) Privacy preserving data mining

Privacy preserving data management [2] is an important emerging research area that emerged in response to two important needs: data analysis and ensuring the privacy of the data owners. Privacy preserving data publishing emphasizes the importance of need for privacy threats in data sharing A new approach [5] seeks to protect data without focusing on the infrastructure level, but at element or aggregate data type. This type of pervasive security can be achieved by classifying data and enforcing access control.

4. CONCLUSIONS

Many new problems have emerged and have been solved by data mining researchers. This paper examined a few important research challenges in data mining. There are still several interesting research issues not covered in this short abstract.

Finally, summarize the major challenges

- Design classifiers to handle ultra-high dimensional classification problem
- Mining data streams in extremely large database
- Mining complex knowledge from complex data
- Mining across multiple heterogeneous data sources: Multi database and multi relational mining
- Mining Non-Relational data
- Automate Data cleaning
- Privacy preserving data mining

REFERENCES

1. Online Mining Of Changes From Data Streams:Research Problems And Preliminary Results, Guozhu Dong,Jiawei Han,Laks V.S. Lakshmanan, *Acm Sigmod Mpds* ’03 San Diego, Ca, Usa,2002.

2. 10 Challenging Problems In Data Mining Research, Qiang Yang, International Journal Of Information Technology & Decision Makingvol. 5, No. 4 (2006) 597–604.

3. Research Challenges For Data Mining In Science And Engineering, Jiawei Han And Jing Gao.

4. Data Mining And Visualization**,** Ron Kohavi, National Academy Of Engineering (Nae) Us Frontiers Of Engineering 2000.

5. Research Issues In Data Mining, Sanjeev Kumar,Iasri Library Avenue, New Delhi.

6. Gediminas Adomavicius,” C-Trend: Temporal Cluster Graphs Foridentifying And Visualizing Trends In Multiattribute Transactional Data” Ieee Transactions On Knowledge And Data Engineering, Vol. 20, No. 6, June 2008.

7. Gediminas Adomavicius and Jesse Bockstedt,” C-TREND: A New Technique for Identifying Trends in Transactional Data” Winter Conference on Business Intelligence,2007.

8. M.S. Chen, J. Han, and P.S. Yu. —Data mining: An overview from database perspective, IEEE Transactions on Knowledge and Data Eng., 8(6):866-883, December 1999.

9. Jing He, —Advances in Data Mining: History and Future, Third international Symposium on Information Technology Application, 978-0-7695-3859-4/09 IEEE 2009 DOI 10.1109/IITA.2009.204.