# Challenges in Data Mining Research

[1]Shard Gupta, [2]Anil kumar, [3]Amit Singh
*College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad*
Shardgupta19@gmail.com
*Chauhananil01@gmail.com*
amit84376@gmail.com

*Abstract*— **In today era industrial communities are faced with an ever increasing number of databases 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: Mining Methodology and User Interaction, Performance Issues, Diverse Data Types Issues,Mining complex knowledge from complex data, Mining across multiple heterogeneous data sources: Multi database and multi relational mining, Mining NonRelational data, Automate Data cleaning, Privacy preserving data mining.**

*Keywords*— **Data mining, data mining process, data mining research challenges**.

## I. INTRODUCTION

Data mining is the process of analysing or discover the hidden patterns from the large sets of data through machine learning, statistics, and database systems. Also called ad knowledge discovery. Data mining powerful new technology mainly used nowadays to assist corporates or companies to focus on the utmost significant information in the data they have possessed about the activities of their valuable customers.

Data mining is a process of extracting interesting knowledge from large amounts of data. That is stored in many data sources. Such as file systems, databases, data warehouses. Also, knowledge used to contributes a lot of benefits to business and individuale .

### *Phases in Data Mining*

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

**a) Define the problem:-** How are those selected business goals translate into specific data mining project goals? The answer to this question will lead to discovering what data sets may be needed and what is in those data sets etc.

**b) Identify reuired data:- Once step 1 is completed, gather required data and understand the data. Are all attributes understood? What is the data quality of those records and attributes? Do some visual inspection of data and do spot checks. This will give you an idea of how much data preparation and pre-processing may be required.**

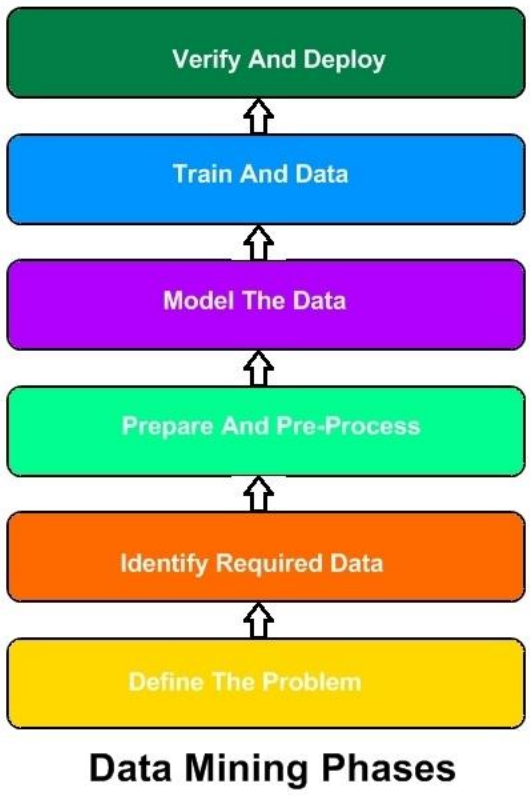
**c) *Prepare and pre-preocess*:-** This is where the grunt work will start. Select required data from the overall collection and go through the process of cleansing and formatting appropriately if necessary. You may realize that you only need partial data sets for the project you or your org has scoped out in step 1. There may be a need for integration of multiple data sources to prepare the final data. Some of these data sources may even be external to complete some attributes of the data.

d) ***Model the data*:-** Actual mining part of data mining will start with this step. Select

appropriate algorithms for the required task and necessary parameters. Look at the data mining techniques article to get an idea of the algorithms. By this time, you would have selected a tool or tools to enhance your productivity. Using those tools, build the model and assess initial results. Given that the end goal of data mining is about predicting, the results at some times may invalidate prior assumptions if the predictions are outside prior hypothesis. Modeling itself may comprise of multiple steps with respect to describing the data as mentioned in data mining techniques article.

e) ***Train and test***:- Evaluate preliminary results and test the model on different sample data sets and review the results. Do these results across different samples correlate? Are there any inconsistencies? Keep iterating until you are satisfied with the consistency of the results.

f) ***Verify and depoly***:- Verify the final model and plan for deployment. Think about the visualizations needed to tell the story. Remember that data mining is as much about story-telling as it is about modelling. Report the findings and operationalize the process.




**Data Mining Phases**

***Data Mining Process***

Here are the 6 essential steps of the data mining process.

1***. Business understanding***

In the business understanding phase:

- First, it is required to understand business objectives clearly and find out what are the business's needs.
- Next, assess the current situation by finding the resources, assumptions, constraints and other important factors which should be considered.
- Then, from the business objectives and current situations, create data mining goals to achieve the business objectives within the current situation.

- Finally, a good data mining plan has to be established to achieve both business and data mining goals. The plan should be as detailed as possible.

**2.** ***Data understanding***

- The data understanding phase starts with initial data collection, which is collected from available data sources, to help get familiar with the data. Some important activities must be performed including data load and data integration in order to make the data collection successfully.
- Next, the "gross" or "surface" properties of acquired data need to be examined carefully and reported.
- Then, the data needs to be explored by tackling the data mining questions, which can be addressed using querying, reporting, and visualization.

**3.** ***Data preparation***

The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set. Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding.

**4.** ***Modeling***

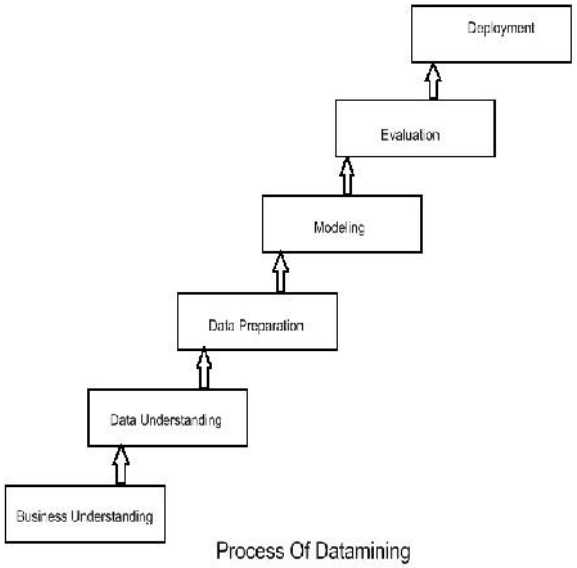
- First, modeling techniques have to be selected to be used for the prepared data set.
- Next, the test scenario must be generated to validate the quality and validity of the model.
- Then, one or more models are created on the prepared data set.
- Finally, models need to be assessed carefully involving stakeholders to make sure that created models are met business initiatives.

**5.** ***Evaluation***

In the evaluation phase, the model results must be evaluated in the context of business objectives in the first phase. In this phase, new business requirements may be raised due to the new patterns that have been discovered in the model results or from other factors. Gaining business understanding is an iterative process in data mining. The go or no-go decision must be made in this step to move to the deployment phase.

**6.** ***Deployment***

The knowledge or information, which is gained through data mining process, needs to be presented in such a way that stakeholders can use it when they want it. Based on the business requirements, the deployment phase could be as simple as creating a report or as complex as a repeatable data mining process across the organization. In the deployment phase, the plans for deployment, maintenance, and monitoring have to be created for implementation and also future supports. From the project point of view, the final report of the project needs to summary the project experiences and review the project to see what need to improved created learned lessons.



Process Of Datamining

## 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 multirelational 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, notdatabases, 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.

.

## CONCLUSION

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

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

## REFERENCES

[1] Major Research Challenges in Data Mining, AnnanNaidu Paidi, International Journal of Trend in Research and Development, Volume 2(4), ISSN 2394-9333 www.ijtrd.com

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

[3] 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.

[4] 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.

[5] 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.

[6] 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

[7] 10 Challenging Problems In Data Mining Research, Qiang Yang, International Journal Of Information Technology & Decision Makingvol. 5, No. 4 (2006) 597–604.

[8] Research Challenges For Data Mining In Science And Engineering, Jiawei Han And Jing Gao.

[9] Data Mining And Visualization**,** Ron Kohavi, National Academy Of Engineering (Nae) Us Frontiers Of Engineering 2000

[10] Research Issues In Data Mining, Sanjeev Kumar,Iasri Library Avenue, New Delhi.

[11] https://digitaltransformationpro.com/data-mining-steps/

[12] https://barnraisersllc.com/2018/10/data-mining-process-essential-steps/