DATA MINING

data pattern evaluation

tutorialspoint

SIMPLYEASYLEARNING

www.tutorialspoint.com

https://www.facebook.com/tutorialspointindia

https://twitter.com/tutorialspoint

About the Tutorial

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data.

The tutorial starts off with a basic overview and the terminologies involved in data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web.

Audience

This tutorial has been prepared for computer science graduates to help them understand the basic-to-advanced concepts related to data mining.

Prerequisites

Before proceeding with this tutorial, you should have an understanding of the basic database concepts such as schema, ER model, Structured Query language and a basic knowledge of Data Warehousing concepts.

Copyright & Disclaimer

All the content and graphics published in this e-book are the property of Tutorials Point (I) Pvt. Ltd. The user of this e-book is prohibited to reuse, retain, copy, distribute or republish any contents or a part of contents of this e-book in any manner without written consent of the publisher.

We strive to update the contents of our website and tutorials as timely and as precisely as possible, however, the contents may contain inaccuracies or errors. Tutorials Point (I) Pvt. Ltd. provides no guarantee regarding the accuracy, timeliness or completeness of our website or its contents including this tutorial. If you discover any errors on our website or in this tutorial, please notify us at contact@tutorialspoint.com

Table of Contents

About the Tutorial i

Audience i

Prerequisites i

Copyright & Disclaimer i

Table of Contents ii

1. DATA MINING – OVERVIEW 1

What is Data Mining? 1

Data Mining Applications 1

Market Analysis and Management 1

Corporate Analysis and Risk Management 2

Fraud Detection 2

2. DATA MINING – TASKS 3

Descriptive Function 3

Classification and Prediction 4

Data Mining Task Primitives 5

3. DATA MINING – ISSUES 7

Mining Methodology and User Interaction Issues 7

Performance Issues 8

Diverse Data Types Issues 8

4. DATA MINING – EVALUATION 9

Data Warehouse 9

Data Warehousing 9

Query-Driven Approach 9

Update-Driven Approach 10

From Data Warehousing (OLAP) to Data Mining (OLAM) 10

Importance of OLAM 11

5. DATA MINING – TERMINOLOGIES 13

Data Mining 13

Data Mining Engine 13

Knowledge Base 13

Knowledge Discovery 13

User Interface 14

Data Integration 14

Data Cleaning 14

Data Selection 14

Clusters 14

Data Transformation 15

6. DATA MINING – KNOWLEDGE DISCOVERY 16

What is Knowledge Discovery? 16

7. DATA MINING – SYSTEMS 17

Data Mining System Classification 17

Integrating a Data Mining System with a DB/DW System 19

8. DATA MINING – QUERY LANGUAGE 20

Syntax for Task-Relevant Data Specification 20

Syntax for Specifying the Kind of Knowledge 20

Syntax for Concept Hierarchy Specification 22

Syntax for Interestingness Measures Specification 23

Syntax for Pattern Presentation and Visualization Specification 23

Full Specification of DMQL 23

Data Mining Languages Standardization 24

9. DATA MINING – CLASSIFICATION AND PREDICTION 25

What is Classification? 25

What is Prediction? 25

How Does Classification Work? 25

Classification and Prediction Issues 27

Comparison of Classification and Prediction Methods 27

10. DATA MINING – DECISION TREE INDUCTION 28

Decision Tree Induction Algorithm 28

Tree Pruning 30

Cost Complexity 30

11. DATA MINING – BAYESIAN CLASSIFICATION 31

Bayes' Theorem 31

Bayesian Belief Network 31

Directed Acyclic Graph 31

Directed Acyclic Graph Representation 32

Conditional Probability Table 32

12. DATA MINING – RULE-BASED CLASSIFICATION 33

IF-THEN Rules 33

Rule Extraction 33

Rule Induction Using Sequential Covering Algorithm 33

Rule Pruning 34

13. DATA MINING – CLASSIFICATION METHODS 35

Genetic Algorithms 35

Rough Set Approach 35

Fuzzy Set Approach 36

14. DATA MINING – CLUSTER ANALYSIS 38

What is Clustering? 38

Applications of Cluster Analysis 38

Requirements of Clustering in Data Mining 38

Clustering Methods 39

15. DATA MINING – MINING TEXT DATA 42

Information Retrieval 42

Basic Measures for Text Retrieval 42

16. DATA MINING – MINING WORLD WIDE WEB 44

Challenges in Web Mining 44

Mining Web Page Layout Structure 44

Vision-based Page Segmentation (VIPS) 44

17. DATA MINING – APPLICATIONS AND TRENDS 46

Data Mining Applications 46

Data Mining System Products 48

Choosing a Data Mining System 48

Trends in Data Mining 49

18. DATA MINING – THEMES 51

Theoretical Foundations of Data Mining 51

Statistical Data Mining 52

Visual Data Mining 53

Audio Data Mining 53

Data Mining and Collaborative Filtering 54

1. DATA MINING – OVERVIEW

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

What is Data Mining?

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications:

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Data Mining Applications

Data mining is highly useful in the following domains:

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid

Market Analysis and Management

Listed below are the various fields of market where data mining is used:

- **Customer Profiling** - Data mining helps determine what kind of people buy what kind of products.

- **Identifying Customer Requirements** - Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
- **Cross Market Analysis** - Data mining performs Association/correlations between product sales.
- **Target Marketing** - Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
- **Determining Customer purchasing pattern** - Data mining helps in determining customer purchasing pattern.
- **Providing Summary Information** - Data mining provides us various multidimensional summary reports.

Corporate Analysis and Risk Management

Data mining is used in the following fields of the Corporate Sector:

- **Finance Planning and Asset Evaluation** - It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
- **Resource Planning** - It involves summarizing and comparing the resources and spending.
- **Competition** - It involves monitoring competitors and market directions.

Fraud Detection

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

2. DATA MINING – TASKS

Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining:

- Descriptive
- Classification and Prediction

Descriptive Function

The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions:

- Class/Concept Description
- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- Mining of Clusters

Class/Concept Description

Class/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways:

- **Data Characterization** - This refers to summarizing data of a class under study. This class under study is called as the Target Class.
- **Data Discrimination** - It refers to the mapping or classification of a class with some predefined group or class.

Mining of Frequent Patterns

Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns:

- **Frequent Item Set** - It refers to a set of items that frequently appear together, for example, milk and bread.
- **Frequent Subsequence**- A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.

- **Frequent Sub Structure** - Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.

Mining of Association

Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules.

For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

Mining of Correlations

It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

Mining of Clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

Classification and Prediction

Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms:

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

The list of functions involved in these processes are as follows:

- **Classification** - It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
- **Prediction** - It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.

- **Outlier Analysis** - Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.
- **Evolution Analysis** - Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.

Data Mining Task Primitives

- We can specify a data mining task in the form of a data mining query.
- This query is input to the system.
- A data mining query is defined in terms of data mining task primitives.

Note: These primitives allow us to communicate in an interactive manner with the data mining system. Here is the list of Data Mining Task Primitives:

- Set of task relevant data to be mined.
- Kind of knowledge to be mined.
- Background knowledge to be used in discovery process.
- Interestingness measures and thresholds for pattern evaluation.
- Representation for visualizing the discovered patterns.

Set of task relevant data to be mined

This is the portion of database in which the user is interested. This portion includes the following:

- Database Attributes
- Data Warehouse dimensions of interest

Kind of knowledge to be mined

It refers to the kind of functions to be performed. These functions are:

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis

- Evolution Analysis

Background knowledge

The background knowledge allows data to be mined at multiple levels of abstraction. For example, the Concept hierarchies are one of the background knowledge that allows data to be mined at multiple levels of abstraction.

Interestingness measures and thresholds for pattern evaluation

This is used to evaluate the patterns that are discovered by the process of knowledge discovery. There are different interesting measures for different kind of knowledge.

Representation for visualizing the discovered patterns

This refers to the form in which discovered patterns are to be displayed. These representations may include the following:

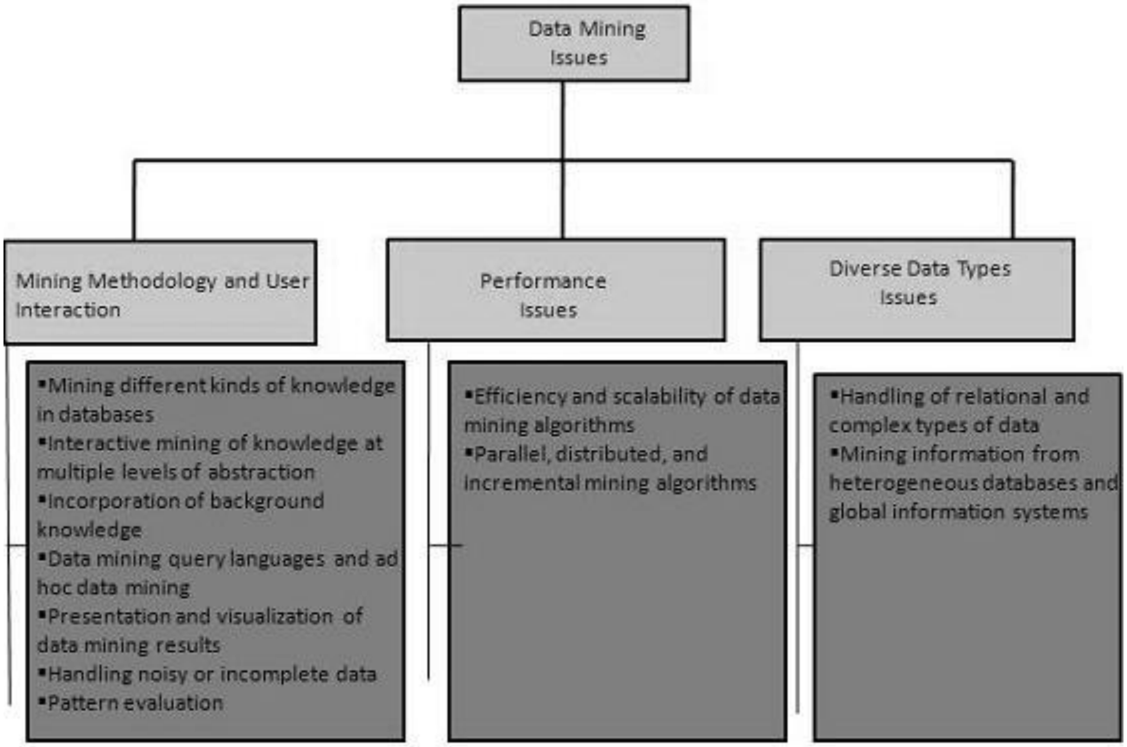
- Rules
- Tables
- Charts
- Graphs
- Decision Trees
- Cubes

3. DATA MINING – ISSUES

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding:

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.



Data Mining Issues
Mining Methodology and User Interaction
Performance Issues
Diverse Data Types Issues
▪Mining different kinds of knowledge in databases
▪Interactive mining of knowledge at multiple levels of abstraction
▪Incorporation of background knowledge
▪Data mining query languages and ad hoc data mining
▪Presentation and visualization of data mining results
▪Handling noisy or incomplete data
▪Pattern evaluation
▪Efficiency and scalability of data mining algorithms
▪Parallel, distributed, and incremental mining algorithms
▪Handling of relational and complex types of data
▪Mining information from heterogeneous databases and global information systems

Mining Methodology and User Interaction Issues

It refers to the following kinds of issues:

- **Mining different kinds of knowledge in databases** - Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.

- **Interactive mining of knowledge at multiple levels of abstraction** - The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

- **Incorporation of background knowledge** - To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.

- **Data mining query languages and ad hoc data mining** - Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

- **Presentation and visualization of data mining results** - Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.

- **Handling noisy or incomplete data** - The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

- **Pattern evaluation** - The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows:

- **Efficiency and scalability of data mining algorithms** - In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

- **Parallel, distributed, and incremental mining algorithms** - The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** - The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.

- **Mining information from heterogeneous databases and global information systems** - The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

4. DATA MINING – EVALUATION

Data Warehouse

A data warehouse exhibits the following characteristics to support the management's decision-making process:

- **Subject Oriented** - Data warehouse is subject oriented because it provides us the information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. The data warehouse does not focus on the ongoing operations, rather it focuses on modelling and analysis of data for decision-making.

- **Integrated** - Data warehouse is constructed by integration of data from heterogeneous sources such as relational databases, flat files etc. This integration enhances the effective analysis of data.

- **Time Variant** - The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from a historical point of view.

- **Non-volatile** - Nonvolatile means the previous data is not removed when new data is added to it. The data warehouse is kept separate from the operational database therefore frequent changes in operational database is not reflected in the data warehouse.

Data Warehousing

Data warehousing is the process of constructing and using the data warehouse. A data warehouse is constructed by integrating the data from multiple heterogeneous sources. It supports analytical reporting, structured and/or ad hoc queries, and decision making.

Data warehousing involves data cleaning, data integration, and data consolidations. To integrate heterogeneous databases, we have the following two approaches:

- Query Driven Approach

- Update Driven Approach

Query-Driven Approach

This is the traditional approach to integrate heterogeneous databases. This approach is used to build wrappers and integrators on top of multiple heterogeneous databases. These integrators are also known as mediators.

Process of Query Driven Approach

1. When a query is issued to a client side, a metadata dictionary translates the query into the queries, appropriate for the individual heterogeneous site involved.
2. Now these queries are mapped and sent to the local query processor.
3. The results from heterogeneous sites are integrated into a global answer set.

Disadvantages

This approach has the following disadvantages:

- The Query Driven Approach needs complex integration and filtering processes.
- It is very inefficient and very expensive for frequent queries.
- This approach is expensive for queries that require aggregations.

Update-Driven Approach

Today's data warehouse systems follow update-driven approach rather than the traditional approach discussed earlier. In the update-driven approach, the information from multiple heterogeneous sources is integrated in advance and stored in a warehouse. This information is available for direct querying and analysis.

Advantages

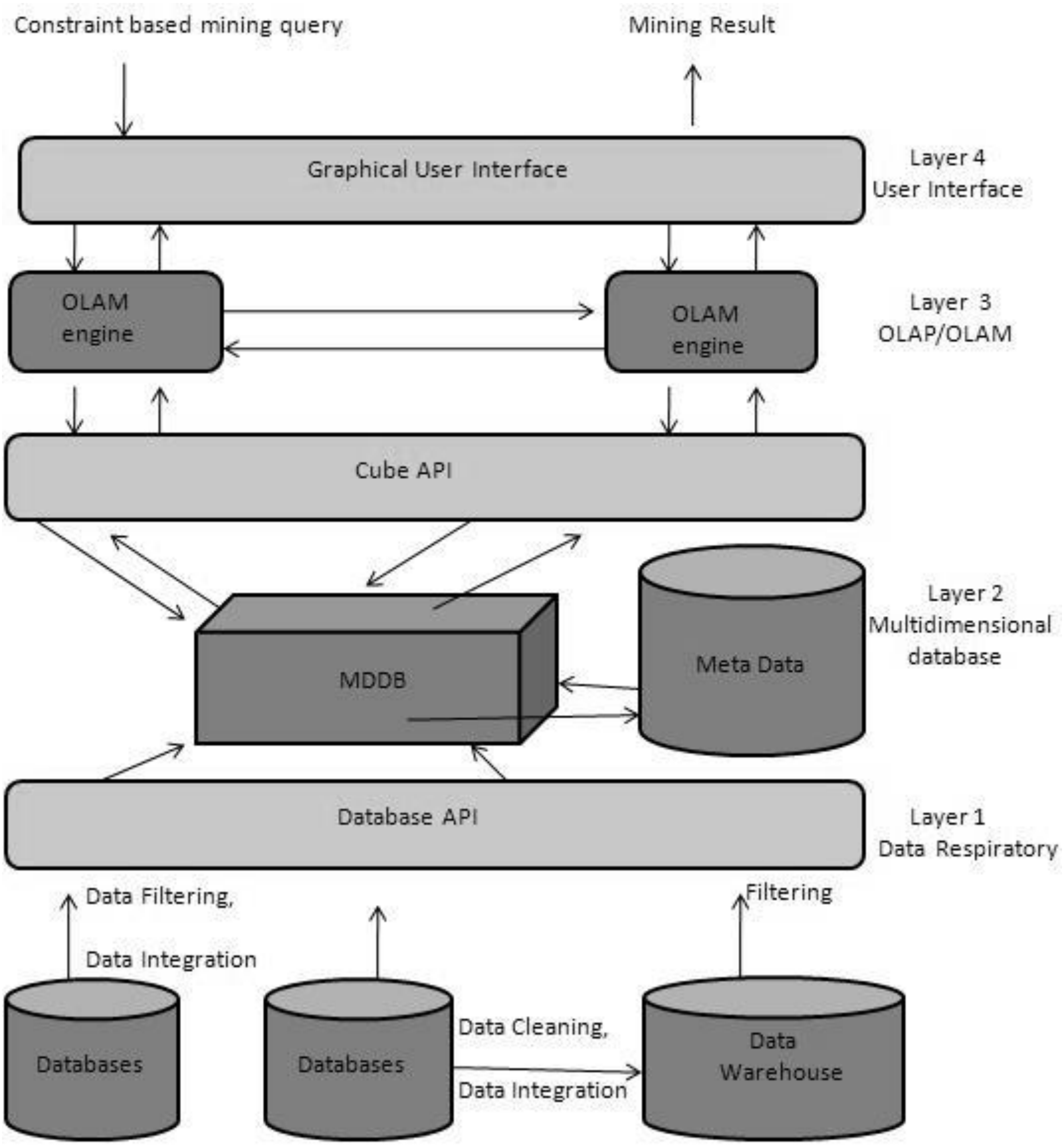
This approach has the following advantages:

- This approach provides high performance.
- The data can be copied, processed, integrated, annotated, summarized and restructured in the semantic data store in advance.

Query processing does not require interface with the processing at local sources.

From Data Warehousing (OLAP) to Data Mining (OLAM)

Online Analytical Mining integrates with Online Analytical Processing with data mining and mining knowledge in multidimensional databases. Here is the diagram that shows the integration of both OLAP and OLAM:



Importance of OLAM

OLAM is important for the following reasons:

- **High quality of data in data warehouses** - The data mining tools are required to work on integrated, consistent, and cleaned data. These steps are very costly in the preprocessing of data. The data warehouses constructed by such preprocessing are valuable sources of high quality data for OLAP and data mining as well.

- **Available information processing infrastructure surrounding data warehouses** - Information processing infrastructure refers to accessing, integration, consolidation, and transformation of multiple heterogeneous databases, web-accessing and service facilities, reporting and OLAP analysis tools.

- **OLAP-based exploratory data analysis** - Exploratory data analysis is required for effective data mining. OLAM provides facility for data mining on various subset of data and at different levels of abstraction.
- **Online selection of data mining functions** - Integrating OLAP with multiple data mining functions and online analytical mining provide users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.