Data Warehouse of Knowledge: Mining Data for Intelligence

Mrs.K.S.Hemalatha /Mrs.D.Gopika/

Mrs.K.Nithyadevi/Mrs.P.Lakshmi Priya

SCIENTIFIC RESEARCH PUBLICATION

Offices:New Delhi New York St Louis San Francisco AucklandBogotá Caracas Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal San Juan Santiago Singapore Sydney Tokyo Toronto

ISBN : 979-8-20-839577-6
Publications: SCIENTIFIC RESEARCH PUBLICATION

Dedicated to

To the divine,
my family,

and my mentors for their
unwavering support and inspiration.

AUTHOR PROFILE

Mrs.K.S.Hemalatha

Assistant professor in Department of Computer Science at Shri Nehru Maha Vidyalaya College of arts and Science, Coimbatore. With 3 years of experience in the field,she has consistently demonstrated a commitment to excellence in teaching, student development and subject matter expertise. Her area of interest include Data Mining, Data Science and Machine Learning.She actively involved in both curricular and co curricular activities aimed at Multidimensional student devlopment

Mrs.D.Gopika

Assistant Professor, Department of Computer Science at Shri Nehru Maha Vidyalaya College of Arts and Science, Coimbatore. She has several years of experience in undergraduate teaching. Her areas of academic interest include Data Mining, Machine Learning, and Database Systems. She has guided numerous student projects and actively participates in academic and research activities. Her teaching approach emphasizes conceptual clarity, practical understanding, and student-centered learning.

Mrs.K.Nithyadevi

Assistant Professor in Department of Information Technology at Shri Nehru maha vidyalaya college of arts and science, Coimbatore . She has a strong background in Computer Networks with several years of teaching experience at college level.she worked closely with students to address common learning challenges and a passion for presenting concept in a clear structural and student friendly manner and she committed to supporting learners through well organized and application oriented study material

Mrs.P.Lakshmi Priya

Assistant Professor in Department of Computer Technology at Shri Nehru Maha Vidyalaya College of Arts and Science, Coimbatore. Passionate over the field of Computer Science made her to choose an irrepressible profession based on the same. She has a resilient background in Computer Networks and Data Structures with nearly 18 years of teaching experience at college level. She stanched to care the upcoming buds through a structured and application oriented study material. Encouraged the students in a friendly approach and reinforced them with her stretched knowledge to impart the concepts and principles in a well understanding way.

FOREWORD

In the present era of information explosion, vast amounts of data are being generated every moment across diverse domains such as business, healthcare, science, social media, and government sectors. Extracting meaningful patterns and useful knowledge from this massive data has become both a necessity and a challenge. **Data Mining** addresses this challenge by providing techniques and methodologies to discover hidden patterns, relationships, and trends within large datasets, thereby supporting effective decision-making.

This book provides a structured and comprehensive introduction to the fundamental concepts, techniques, and applications of data mining. **Chapter 1** introduces the basic data mining tasks, distinguishes data mining from knowledge discovery in databases (KDD), and discusses key issues, metrics, social implications, and database perspectives. This foundation helps learners understand the scope and relevance of data mining in real-world scenarios.

focuses on core data mining techniques, covering statistical perspectives, similarity measures, decision trees, neural networks, and genetic algorithms. These techniques form the backbone of modern data mining systems and are essential for practical implementation.

It is dedicated to **classification**, presenting various algorithmic approaches such as statistical, distance-based, decision tree-based, neural network-based, and rule-based methods, along with techniques for combining classifiers to improve accuracy. explores **clustering**, a key unsupervised learning task, discussing similarity measures, outlier detection, hierarchical methods, and partitional algorithms. This chapter emphasizes the importance of grouping data objects based on inherent structures within the data.

PREFACE

Data mining has emerged as an essential discipline in the modern data-driven world, enabling the extraction of meaningful knowledge from large and complex datasets. This book is designed to provide a comprehensive understanding of the fundamental concepts, techniques, and applications of data mining.

It covers basic data mining tasks, issues, and metrics, followed by detailed discussions on statistical methods, similarity measures, decision trees, neural networks, genetic algorithms, classification, clustering, and association rule mining. Emphasis is placed on both theoretical foundations and practical approaches, with each chapter supported by practice sets to enhance learning and analytical skills.

The content is structured to meet the academic requirements of undergraduate and postgraduate students while also serving as a useful reference for professionals seeking insight into data mining methodologies and their real-world applications.

The content of this book is organized in a structured manner to facilitate progressive learning. It begins with an introduction to basic data mining tasks, clearly explaining the differences between data mining and knowledge discovery in databases, along with important issues, metrics, and social implications.

The book then explores various data mining techniques from statistical and computational perspectives, including similarity measures, decision trees, neural networks, and genetic algorithms. Special attention is given to classification methods, covering statistical, distance-based, decision tree-based, neural network-based, and rule-based algorithms, as well as techniques for combining multiple classifiers to improve performance.

IMPORTANT TERMS AND ABBREVIATIONS

Data Mining, often referred to as DM, involves uncovering patterns within extensive datasets.

Knowledge Discovery in Databases, or KDD, is a more encompassing process that includes data cleaning, integration, selection, transformation, mining, and interpretation.

ETL, standing for Extract, Transform, Load, outlines the crucial steps for preparing data before mining can occur.

OLAP, or Online Analytical Processing, provides tools for performing multidimensional analyses. Metadata is essential for mining as it offers context by describing the data itself.

A Similarity Measure offers a quantitative method for comparing data objects, with examples like cosine similarity and Euclidean distance.

A Decision Tree, or DT, is a predictive model structured like a tree.

Neural Networks (NN) are computational models that draw inspiration from biological neural structures.

Genetic Algorithms (GA) are an optimization technique rooted in the principles of natural selection.

Support Vector Machines (SVM) are a classification technique that utilizes hyperplanes.

Principal Component Analysis (PCA) is a method used for reducing the dimensionality of data.

Supervised Learning involves training models using labeled data. The Bayesian Classifier is a statistical approach that relies on probability calculations.

KNN, or K-Nearest Neighbor, is a classification method based on distance. Entropy and Information Gain are key metrics employed in decision trees.

Rule-Based Classifiers employ IF-THEN logic to make classifications.

Ensemble Methods combine multiple models, such as Bagging, Boosting, and Random Forests.

Unsupervised Learning, on the other hand, does not use predefined labels and groups data based on inherent similarities.

Hierarchical Clustering constructs nested clusters, either agglomeratively (bottom-up) or divisively (top-down).

Partitional Clustering divides data into distinct, non-overlapping clusters, with K-Means being a common example.

Outliers are data points that significantly deviate from the rest of the dataset.

DBSCAN, a popular clustering algorithm, is known as Density-Based Spatial Clustering of Applications with Noise.

Association Rule Mining aims to discover interesting relationships between variables.

Support quantifies the frequency with which an itemset appears. Confidence measures the likelihood that a consequent will occur given an antecedent.

Lift indicates the strength of an association relative to what would be expected by random chance.

The Apriori Algorithm is a well-established algorithm for mining frequent items.

TABLE OF CONTENTS

CHAPTER 4 Clustering

CHAPTER 5 Association Rule

CHAPTER 1 Basic Data Mining Task

1.1 Basic Data Mining Tasks

The types of patterns that can be mined are the focus of data mining. Two types of functions are involved in data mining, depending on the type of data to be mined:

- Descriptive
- Categorization and Forecasting

The Descriptive Function

The broad characteristics of the data in the database are handled by the descriptive function. The list of descriptive functions is as follows:

- Description of the Class/Concept
- Mining Recurring Patterns
- Mining Associations
- Correlation Mining
- Cluster Mining

Description of the Class/Concept

Class/Concept describes the information to be connected to the concepts or classes. For instance, at a business, the notions of clients include big spenders and budget spenders, and the classes of things for sale include computers and printers. Class/concept descriptions are these kinds of explanations of a class or a notion. The following two methods can be used to derive these descriptions:

- Data characterization is the process of summarizing the data of the class that is being studied. The class that is being studied is referred to as the Target Class.

- The mapping or classifying of a class with a predetermined group or class is known as data discrimination.

Mining Recurring Patterns

Patterns that appear frequently in transactional data are known as frequent patterns. The list of somewhat common patterns is as follows:

A collection of things that commonly appear together, like milk and bread, is referred to as a "frequent item set."

Frequent Subsequence: A series of recurring patterns, such buying a camera, are followed by memory cards.

Frequently Used Substructure: Substructure is a term used to describe several structural forms that can be paired with item-sets or subsequences, such as graphs, trees, or lattices.

Association Mining

In retail sales, associations are utilized to find patterns that are commonly bought together. The process of identifying the relationships between data and establishing association rules is referred to as this procedure.

For instance, a shop creates an association rule that indicates that only 30% of the time are biscuits sold with bread, although milk is sold with bread 70% of the time.

Correlation Mining

Finding intriguing statistical connections between associated-attribute-value pairings or between two item sets to determine if they have a positive, negative, or no effect on one another is a type of extra analysis.

Cluster Mining

A collection of comparable objects is referred to as a cluster. Cluster analysis is the process of grouping things that are very similar to one another but very different from those in other clusters.

Categorization and Forecasting

Finding a model to explain the data classes or concepts is the process of classification. The goal is to be able to forecast the class of objects whose class label is unknown using this model. The study of training data sets served as the foundation for this generated model. The following formats can be used to display the derived model:

- Rules for Classification (IF-THEN)
- Trees of Decisions
- Formulae in Mathematics
- The Neural Network

The following is a list of the functions that these procedures involve:

Classification: It makes predictions about the class of things with unknown class labels. Finding a derived model that characterizes and differentiates data classes or concepts is its goal. The analysis set of training data, or the data object with a known class name, serves as the foundation for the Derived Model.

Prediction: Rather than class labels, it is used to forecast missing or unavailable numerical data values. Prediction is a common application of regression analysis. Based on the data at hand, prediction can also be used to identify distribution trends.

Outlier Analysis: Data objects that deviate from the typical behavior or model of the given data are known as outliers.

Evolution Analysis: For things whose behavior varies over time, evolution analysis describes and models regularities or patterns.

Primitives for Data Mining Tasks

- A data mining query can be used to define a data mining task.
- The system receives this inquiry.
- Data mining task primitives are used to define a data mining query.
- set of pertinent facts for the task to be mined.
- type of information that can be extracted.
- background information to be applied during the process of discovery.
- Thresholds for evaluating patterns and metrics of interest.
- representation to help visualize the patterns found.

set of task-relevant data to be extracted

The user is interested in this section of the database. The following are included in this section:

- Database Features
- Dimensions of interest for data warehouses

Type of information to be extracted

It speaks to the types of tasks that need to be completed. These roles are—

- Characterization
- Discrimination
- Analysis of Association and Correlation
- Prediction of Classification
- Analysis of Clustering Outliers
- Analysis of Evolution

Prior knowledge

Data can be mined at several levels of abstraction thanks to the background knowledge. One of the baseline information that enables data mining at various levels of abstraction, for instance, is concept hierarchies.

Measures of interest and cutoff points for assessing patterns

This is used to assess the patterns found during the knowledge-discovery process. There are various intriguing metrics for various types of knowledge.

Visualization of the identified patterns through representation

This is the format for displaying patterns that have been found. The following are examples of such representations.

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

1.2 Data Mining Versus Knowledge Discovery in Data Bases

The entire process of extracting useful information from massive datasets is known as Knowledge Discovery in Databases (KDD). Selecting pertinent data is the first step, which is followed by preprocessing to clean and arrange it, transformation to get it ready for analysis, data mining to find patterns and relationships, and finally evaluation and interpretation of the findings to produce insightful knowledge. KDD is extensively used in domains such as data visualization, artificial intelligence, statistics, machine learning, and pattern identification.

To guarantee the precision and dependability of the knowledge retrieved, the KDD process is iterative, requiring numerous adjustments. The following steps make up the entire process:

- Data Selection
- Data Cleaning and Preprocessing
- Data Transformation and Reduction
- Data Mining
- Evaluation and Interpretation of Results

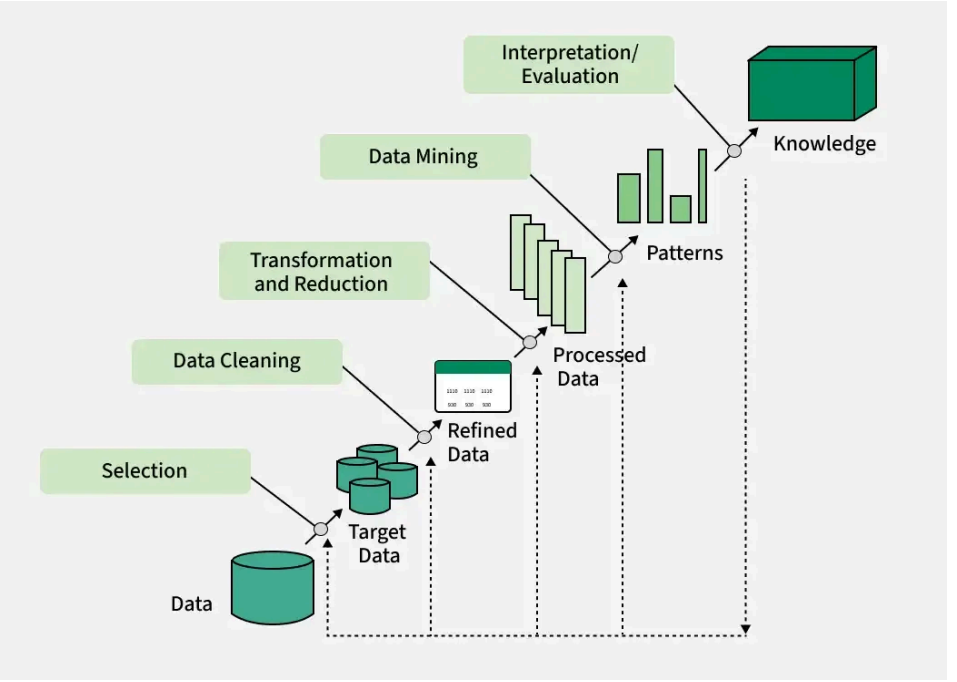

Fig (1.2 KDD)

- **Data Selection**

The first stage of the Knowledge Discovery in Databases (KDD) process is data selection, which involves locating and selecting pertinent data for analysis. In order to extract significant insights, it entails choosing a dataset or concentrating on particular variables, samples, or subsets of data.

- By ensuring that only the most pertinent data are used for analysis, it increases accuracy and efficiency.
- Depending on the objectives of the assignment, it entails choosing the complete dataset or reducing it to certain features or subsets.
- After a thorough grasp of the application domain, data is chosen.

We guarantee that the KDD process produces precise, pertinent, and useful insights by carefully choosing the data.

- **Data Cleaning and Preprocessing**

Data cleaning, which addresses noisy or outlier data, handles missing values, eliminates duplicates, and fixes errors, is crucial to the KDD process in order to guarantee that the dataset is accurate and dependable.

Missing Values: To ensure dataset completeness, missing values are filled in with the mean or most likely value.

Noisy Data: Techniques like binning, regression, or clustering are used to smooth or organize the data in order to reduce noise.

Eliminating Duplicates: To preserve consistency and prevent analytical errors, duplicate records are eliminated.

To improve the quality of the data and increase the efficacy of data mining, data cleaning is essential in KDD.

- **Data Transformation and Reduction**

In KDD, data transformation entails transforming data into a format better suited for analysis.

- Scaling data to a common range to ensure uniformity across variables is known as normalization.
- Discretization is the process of breaking down continuous data into discrete groups for easier analysis.
- Data aggregation is the process of combining several data items (such as averages or totals) to make analysis easier.
- Organizing data into hierarchies for a more comprehensible, higher-level perspective is known as concept hierarchy generation.
- Data reduction preserves important information while making the dataset simpler.
- Dimensionality reduction, such as PCA, is the process of reducing the number of variables while retaining important information.
- Numerosity reduction is the process of reducing data points in order to preserve important patterns through techniques like sampling.
- Data compression is the practice of condensing data to make processing and storage simpler.