UNIT 7 DATA MINING – AN INTRODUCTION

7.0 INTRODUCTION

We are in an age often referred to as the information age in which we have been collecting huge amounts of data. Thanks to sophisticated technologies such as computers, satellites, etc. With the invention of computers and mass digital storage media, we started collecting and storing all sorts of data. Unfortunately, these huge data stored on different storage media became profuse. This chaotic situation of storing the data has led to the creation of database management systems (DBMS). The boosting of database management systems also contributes to massive gathering of data of all types. Today, we have a lot of information than we can handle: from satellite pictures, text reports to business transactions and scientific data and military intelligence. Information retrieval is simply not enough anymore for decision-making. Opposed with huge collections of data, we have generated new requirements for helping us to make better managerial decisions. These requirements may be automatic summarization of data, extraction of the kernel of the information stored, and the discovery of patterns in raw data.

So, Now that we have gone through the introduction part, let's understand what really the concept of data mining is!

7.1 WHAT IS DATA MINING?

Data mining is the process by which the patterns can be found by filtering the data based on a condition of data sorting in order to find something important. At a micro-scale, mining can be defined as an activity that involves gathering data in one place in a structural way. For example, clubbing an Excel Spreadsheet or creating a summarization of the main points of some text.

In other words, the process of data mining can be better explained and defined in the following manner:

- Data mining finds valuable information hidden in large volumes of data.

- Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data.
- The computer is responsible for finding the patterns by identifying the underlying rules and features in the data.

7.2 PURPOSE OF DATA MINING

The purpose of mining the data can be multi-fold which includes:

Predicting various outcomes;

Modeling target audience;

Collecting the information about the product.

Data mining helps to understand certain aspects of customer behavior. This knowledge allows companies to adapt accordingly and offer the best possible services.

In order to have a thorough understanding of the behavior of the customers, Data mining helps to extract the pattern from the databases The knowledge thus acquired, allows the companies to offer the best possible services.

7.3 HOW DOES DATA MINING WORK?

Step-wise, the operation of data mining consists of the following elements:

Building target datasets by selecting what kind of data you need;

Preprocessing is the groundwork for the subsequent operations. This process is also known as data exploration.

Preparing the data - a creation of the segmenting rules, cleaning data from noise, handling missing values, performing anomaly checks, and other operations. This stage may also include further data exploration.

7.4 WHAT TYPES OF DATA CAN BE MINED?

As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data (Section 1.3.1), data warehouse data (Section 1.3.2), and transactional data (Section 1.3.3). The concepts and techniques presented in this book focus on such data. Data mining can also be applied to other forms of data (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW). We present an overview of such data in Section 1.3.4. Techniques for mining of these kinds of data are briefly introduced in Chapter 13. Indepth treatment is considered an advanced topic. Data mining will certainly continue to embrace new data types as they emerge.

1.7.1. Database Data

A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data. The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access. A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.

Example : A relational database for All Electronics. The fictitious All Electronics store is used to illustrate concepts throughout this book. The company is described by the following relation tables: customer, item, employee, and branch. The headers of the tables described here are shown below :(A header is also called the schema of a relation.) The relation customer consists of a set of attributes describing the customer information, including a unique customer identity number (cust ID), customer name, address, age, occupation, annual income, credit information, and category. Similarly, each of the relations item, employee, and branch consists of a set of attributes describing the properties of these entities. Tables can also be used to represent the relationships between or among multiple entities. In our example, these include purchases (customer purchases items, creating a sales transaction handled by an employee), items sold (lists items sold in a given transaction), and works at (employee works at a branch of Al lElectronics).

Relational data can be accessed by database queries written in a relational query language (e.g., SQL) or with the assistance of graphical user interfaces. A given query is transformed into a set of relational operations, such as join, selection, and projection, and is then optimized for efficient processing. A query allows retrieval of specified subsets of the data. Suppose that your job is to analyze the AllElectronics data. Through the use of relational queries, you can ask things like, “Show me a list of all items that were sold in the last quarter.” Relational languages also use aggregate functions such as sum, avg (average), count, max (maximum), and min (minimum). Using aggregates allows you to ask: “Show me the total sales of the last month, grouped by branch,” or “How many sales transactions occurred in the month of December?” or “Which salesperson had the highest sales?” When mining relational databases, we can go further by searching for trends or data patterns. For example, data mining systems can analyze customer data to predict the credit risk of new customers based on their income, age, and previous credit information. Data mining systems may also detect deviations—that is, items with sales that are far from those expected in comparison with the previous year. Such deviations can then be further investigated. For example, data mining may discover that there has been a change in packaging of an item or a significant increase in price. Relational databases are one of the most commonly available and richest information repositories, and thus they are a major data form in the study of data mining. 1.3.2 Data Warehouses Suppose that All Electronics is a successful international company with branches around the world. Each branch has its own set of databases. The president of All Electronics has asked

you to provide an analysis of the company's sales per item type per branch for the third quarter. This is a difficult task, particularly since the relevant data are spread out over several databases physically located at numerous sites. If AllElectronics had a data warehouse, this task would be easy. A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity). The data are stored to provide information from a historical perspective, such as in the past 6 to 12 months, and are typically summarized. For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store or, summarized to a higher level, for each sales region. A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count.

Example A data cube for All Electronics. The cube has three dimensions: address (with city values Chicago, New York, Toronto, Vancouver), time (with quarter values Q1, Q2, Q3, Q4), and item(with item type values home entertainment, computer, phone, security). The aggregate value stored in each cell of the cube is sales amount (in thousands). For example, the total sales for the first quarter, Q1, for the items related to security systems in Vancouver is $400,000, as stored in cell Vancouver, Q1, security. Additional cubes may be used to store aggregate sums over each dimension, corresponding to the aggregate values obtained using different SQL group-bys (e.g., the total sales amount per city and quarter, or per city and item, or per quarter and item, or per each individual dimension). By providing multidimensional data views and the pre computation of summarized data, data warehouse systems can provide inherent support for OLAP. Online analytical processing operations make use of background knowledge regarding the domain of the data being studied to allow the presentation of data at different levels of abstraction. Such operations accommodate different user viewpoints. Examples of OLAP operations include drill-down and roll-up, which allow the user to view the data at differing degrees of summarization. For instance, we can drill down on sales data summarized by quarter to see data summarized by month. Similarly, we can roll up on sales data summarized by city to view data summarized by country. Although data warehouse tools help support data analysis, additional tools for data mining are often needed for in-depth analysis.

1.7.2. Transactional Data

In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page. A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction. A transactional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on. Example: A transactional database for All Electronics. Transactions can be stored in a table, with one record per transaction. From the relational database point of view, the sales table in the figure is a nested relation because the attribute list of item IDs contains a set of items. Because most relational database systems do not support nested relational structures, the transactional database

is usually either stored in a flat file in a format similar to the table in Figure 1.8 or unfolded into a standard relation in a format similar to the items sold table in Figure 1.5. As an analyst of All Electronics, you may ask, "Which items sold well together?" This kind of market basket data analysis would enable you to bundle groups of items together as a strategy for boosting sales. For example, given the knowledge that printers are commonly purchased together with computers, you could offer certain printers at a steep discount (or even for free) to customers buying selected computers, in the hopes of selling more computers (which are often more expensive than printers). A traditional database system is not able to perform market basket data analysis. Fortunately, data mining on transactional data can do so by mining frequent item sets, that is, sets of items that are frequently sold together.

1.7.3. Other Kinds of Data

Besides relational database data, data warehouse data, and transaction data, there are many other kinds of data that have versatile forms and structures and rather different semantic meanings. Such kinds of data can be seen in many applications: time-related or sequence data (e.g., historical records, stock exchange data, and time-series and biological sequence data), data streams (e.g., video surveillance and sensor data, which are continuously transmitted), spatial data (e.g., maps), engineering design data (e.g., the design of buildings, system components, or integrated circuits), hypertext and multimedia data (including text, image, video, and audio data), graph and networked data (e.g., social and information networks), and the Web (a huge, widely distributed information repository made available by the Internet). These applications bring about new challenges, like how to handle data carrying special structures (e.g., sequences, trees, graphs, and networks) and specific semantics (such as ordering, image, audio and video contents, and connectivity), and how to mine patterns that carry rich structures and semantics. Various kinds of knowledge can be mined from these kinds of data. Regarding temporal data, we can mine banking data for changing trends, which may aid in the scheduling of bank tellers according to the volume of customer traffic. Stock exchange data can be mined to uncover trends that could help you plan investment strategies (e.g., the best time to purchase All Electronics stock). We could mine computer network data streams to detect intrusions based on the anomaly of message flows, which may be discovered by clustering, dynamic construction of stream models or by comparing the current frequent patterns with those at a previous time.

With spatial data, we may look for patterns that describe changes in metropolitan poverty rates based on city distances from major highways. The relationships among a set of spatial objects can be examined in order to discover which subsets of objects are spatially auto correlated or associated. By mining text data, such as literature on data mining from the past ten years, we can identify the evolution of hot topics in the field. By mining user comments on products (which are often submitted as short text messages), we can assess customer sentiments and understand how well a product is embraced by a market.

From multimedia data, we can mine images to identify objects and classify them by assigning semantic labels or tags. By mining video data of a hockey game, we can detect video sequences corresponding to goals. Web mining can help us learn about the distribution of information on the WWW in general, characterize and classify web pages, and uncover web dynamics and the association and other relationships among

different web pages, users, communities, and web-based activities. It is important to keep in mind that, in many applications, multiple types of data are present. For example, in web mining, there often exist text data and multimedia data (e.g., pictures and videos) on web pages, graph data like web graphs, and map data on some web sites. In bioinformatics, genomic sequences, biological networks, and 3-D spatial structures of genomes may coexist for certain biological objects. Mining multiple data sources of complex data often leads to fruitful findings due to the mutual enhancement and consolidation of such multiple sources. On the other hand, it is also challenging because of the difficulties in data cleaning and data integration, as well as the complex interactions among the multiple sources of such data. While such data require sophisticated facilities for efficient storage, retrieval, and updating, they also provide fertile ground and raise challenging research and implementation issues for data mining.

7.5 CLASSIFICATION OF DATA MINING SYSTEMS

The data mining system can be classified according to the following criteria:

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Discipline

Some Other Classification Criteria:

- Classification according to kind of databases mined
- Classification according to kind of knowledge mined
- Classification according to kinds of techniques utilized
- Classification according to applications adapted

Classification according to kind of databases mined

We can classify the data mining system according to kind of databases mined. Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example if we classify the database according to data model then we may have a relational, transactional, object- relational, or data warehouse mining system.

Classification according to kind of knowledge mined

We can classify the data mining system according to kind of knowledge mined. It is means datamining system are classified on the basis of functionalities such as:

- Characterization
- Discrimination
- Association and Correlation Analysis

- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

Classification according to kinds of techniques utilized

We can classify the data mining system according to kind of techniques used. We can describes these techniques according to degree of user interaction involved or the methods of analysis employed.

Classification according to applications adapted

We can classify the data mining system according to application adapted. These applications areas follows:

- Finance
- Telecommunications
- DNA
- Stock Markets
- E-mail.

7.6 DATA MINING VS DATA WAREHOUSING

What is Data warehouse?

A data warehouse is a technique for collecting and managing data from varied sources to provide meaningful business insights. It is a blend of technologies and components which allows the strategic use of data.

Data Warehouse is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users for analysis.

What Is Data Mining?

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.

It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology.

The insights extracted via Data mining can be used for marketing, fraud detection, and scientific discovery, etc.

Data Mining Vs Data Warehouse: Key Differences

Data Mining	Data Warehouse
Data mining is the process of analyzing unknown patterns of data.	A data warehouse is database system which is designed for analytical instead of transactional work.
Data mining is a method of comparing large amounts of data to finding right patterns.	Data warehousing is a method of centralizing data from different sources into one common repository.
Data mining is usually done by business users with the assistance of engineers.	Data warehousing is a process which needs to occur before any data mining can take place.
Data mining is the considered as a process of extracting data from large data sets.	On the other hand, Data warehousing is the process of pooling all relevant data together.
One of the most important benefits of data mining techniques is the detection and identification of errors in the system.	One of the pros of Data Warehouse is its ability to update consistently. That's why it is ideal for the business owner who wants the best and latest features.
Data mining helps to create suggestive patterns of important factors. Like the buying habits of customers, products, sales. So that, companies can make the necessary adjustments in operation and production.	Data Warehouse adds an extra value to operational business systems like CRM systems when the warehouse is integrated.
The Data mining techniques are never 100% accurate and may cause serious consequences in certain conditions.	In the data warehouse, there is great chance that the data which was required for analysis by the organization may not be integrated into the warehouse. It can easily lead to loss of information.
The information gathered based on Data Mining by organizations can be misused against a group of people.	Data warehouses are created for a huge IT project. Therefore, it involves high maintenance system which can impact the revenue of medium to small-scale organizations.
After successful initial queries, users may	Data Warehouse is complicated to

ask more complicated queries which would increase the workload.	implement and maintain.
Organisations can benefit from this analytical tool by equipping pertinent and usable knowledge-based information.	Data warehouse stores a large amount of historical data which helps users to analyze different time periods and trends for making future predictions.
Organisations need to spend lots of their resources for training and Implementation purpose. Moreover, data mining tools work in different manners due to different algorithms employed in their design.	In Data warehouse, data is pooled from multiple sources. The data needs to be cleaned and transformed. This could be a challenge.
The data mining methods are cost-effective and efficient compares to other statistical data applications.	Data warehouse's responsibility is to simplify every type of business data. Most of the work that will be done on user's part is inputting the raw data.
Another critical benefit of data mining techniques is the identification of errors which can lead to losses. Generated data could be used to detect a drop-in sale.	Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
Data mining helps to generate actionable strategies built on data insights.	Once you input any information into Data warehouse system, you will unlikely to lose track of this data again. You need to conduct a quick search, helps you to find the right statistic information.

Why use Data Warehouse?

Some most Important reasons for using Data warehouse are:

- Integrates many sources of data and helps to decrease stress on a production system.
- Optimized Data for reading access and consecutive disk scans.
- Data Warehouse helps to protect Data from the source system upgrades.
- Allows users to perform master Data Management.
- Improve data quality in source systems.

Why use Data mining?

Some most important reasons for using Data mining are:

- Establish relevance and relationships amongst data. Use this information to generate profitable insights
- Business can mak informed decisions quickly
- Helps to find out unusual shopping patterns in grocery stores.
- Optimize website business by providing customize offers to each visitor.
- Helps to measure customer's response rates in business marketing.
- Creating and maintaining new customer groups for marketing purposes.
- Predict customer defections, like which customers are more likely to switch to another supplier in the nearest future.
- Differentiate between profitable and unprofitable customers.
- Identify all kind of suspicious behavior, as part of a fraud detection process.

7.7 DATA MINING TOOLS

Data Mining techniques are derived that utilize the domain knowledge from statistical analysis, artificial intelligence, and database systems in order to analyze data in a proper manner in view of different dimensions and perspectives.

Data Mining tools discover patterns or trends from the collection of large sets of data and transforming data into useful information for making decisions.

Ex: Frameworks such as Rstudio or Tableau that allows us to perform different types of data mining analysis.

These tools allow us to implement algorithms like classification and clustering on the dataset and visualize the results.

The Market for Data Mining tool is shining: as per the latest report from Report Linker noted that the market would top **$1 billion** in sales by **2023**, up from **$ 591** million in **2018**

These are the most popular data mining tools:

1. Orange Data Mining:

Developed at the bioinformatics laboratory at the faculty of computer and information science, Ljubljana University, Slovenia, Orange is a machine learning and data mining software suite. It supports data visualization and is a software-based on components written in Python.

The components of Orange are called "widgets." These widgets range from data visualization and preprocessing to assessing the algorithms and are used for predictive modeling.

Widgets deliver much important functionalities such as:

- Displaying data table and allowing to select features
- Reading of data.
- Training predictors and comparison of learning algorithms
- Data element visualization, etc.

Besides, Orange provides a more interactive and enjoyable atmosphere

Why Orange?

Data coming to orange is formatted quickly to the desired pattern, and the widgets can be easily transferred wherever and whenever needed. Orange is quite interesting to users. Orange allows its users to make smarter decisions in a short time by rapidly comparing and analyzing the data. Data mining can be performed via visual programming or Python scripting. Many analyses are feasible through its visual programming interface(drag and drop connected with widgets)and many visual tools tend to be supported such as bar charts, scatterplots, trees, dendrograms, and heat maps. A substantial amount of widgets(more than 100) tend to be supported.

The instrument has machine learning components, add-ons for bioinformatics and text mining, and it is packed with features for data analytics. This is also used as a python library.

Orange comprises of canvas interface onto which the user places widgets and creates a data analysis workflow. The widget proposes fundamental operations, For example, reading the data, showing a data table, selecting features, training predictors, comparing learning algorithms, visualizing data elements, etc. Orange comes with multiple regression and classification algorithms.

Orange can read documents in native and other data formats. Orange is dedicated to machine learning techniques for classification or supervised data mining. There are two types of objects used in classification: learner and classifiers. Learners consider class-leveled data and return a classifier. Regression methods are very similar to classification in Orange, and both are designed for supervised data mining and require class-level data. The learning of ensembles combines the predictions of individual models for precision gain. The model can either come from different training data or use different learners on the same sets of data.

2. SAS Data Mining:

SAS stands for Statistical Analysis System. It is a product of the SAS Institute created for analytics and data management. SAS can mine data, change it, manage information from various sources, and analyze statistics. It offers a graphical UI for non-technical users.

SAS data miner allows users to analyze big data and provide accurate insight for timely decision-making purposes. SAS has distributed memory processing architecture that is highly scalable. It is suitable for data mining, optimization, and text mining purposes.

4. Rattle:

Rattle is a data mining tool based on GUI. It uses the R programming language. Rattle combines the power of R with significant data mining features.

5. Rapid Miner:

Rapid Miner is one of the most popular tool for performing predictions. It is written in JAVA programming language. It offers an integrated environment for text mining, deep learning, machine learning, and predictive analysis.

The instrument can be used for a wide range of applications, including company applications, commercial applications, research, education, training, application development, machine learning.

Rapid Miner provides the server on-site as well as in public or private cloud infrastructure. It has a client/server model as its base. A rapid miner comes with template-based frameworks that enable fast delivery with few errors(which are commonly expected in the manual coding writing process)

Oracle BI

Oracle BI is an open source machine learning and data visualization for naïve as well as expert users.

Features:

- Interactive Data Visualization.
- It Offers Interactive data exploration for rapid qualitative analysis with clean visualizations.
- Orange supports hands-on training and visual illustrations of concepts from data science.
- It offers an extensive range of add-ons to data mining from external data sources.

11) KNIME

KNIME is open source software for creating data science applications and services. This Data mining tool helps you to understand data and to design data science workflows.

Features:

- Helps you to build an end to end data science workflows
- Blend data from any source
- Allows you to aggregate, sort, filter, and join data either on your local machine, in-database or in distributed big data environments.
- Build machine learning models for classification, regression, dimension reduction

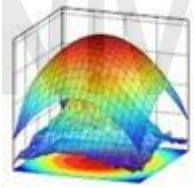
12) Tanagra

Tangra is a free to use data mining tool for study and research purposes. It offers various data mining methods from statistical learning, data analysis, and machine learning.

Features:

- Offers easy to use data mining software for researcher and students
- It allows the user to add their data mining methods.

15) Data Melt



DataMelt is a free to use tool for numeric computation, mathematics, data analysis, and data visualization. This program offers you the simplicity of scripting languages, like Python, Ruby, Groovy with the power of hundreds of Java packages.

Features:

- DataMelt offers statistics, analysis of large data volumes, and scientific visualization.
- You can use it with different programming languages on different operating systems.
- It allows you to create high-quality vector-graphics images (EPS, SVG, PDF, etc.), which can be included in LaTeX and another text processor.
- Data Melt offers the usage of scripting languages, which are significantly faster than the standard Python implemented in C.

16) ELKI:

ELKI is an open source data mining tool written in Java. The tool allows us researching algorithms, with an emphasis on unsupervised methods in cluster analysis and outlier detection.

Features:

- ELKI offers an extensive collection of highly parameterizable algorithms
- It allows easy and fair evaluation and benchmarking of algorithms.
- ELKI provides data index structures such as the R*-tree which enhance the process of Data mining

17) SPMF

SPMF is an open-source data mining library written in Java. It is distributed under the GPL license. It allows you to integrate source code with other Java Software.

Features:

- Allows association rule mining
- Supports sequential pattern and sequential rule mining
- Offers High-utility pattern mining,
- Time-series mining.
- Support complex process of Clustering and classification

18) Alteryx

Alteryx is a Business intelligence and analytics solutions for the enterprise. It is a specially designed tool for data analyst and business leaders.

Features:

- Analytics for Midsize Businesses
- It allows for Ad Hoc Analysis.
- Offers fast online Analytical Processing
- Automatic Scheduled Reporting
- Highly customizable Dashboard

19) Enterprise Miner

Enterprise Miner is a SAS software which offers you and cutting-edge algorithms designed to help you solve the most significant challenges and offers the best solutions for your business.

Features:

- Helps you to improve prediction accuracy. Share reliable results
- Easy-to-use GUI and batch processing
- Advanced predictive and descriptive modeling
- Offers Automated scoring
- Automate model deployment and scoring

20) Datawatch

Datawatch Desktop is a Data mining and business intelligence solution. It allows you to focus on real-time data visualization. It offers tools to build and deploy their monitoring and analysis systems without the need to write a single line of code.

Features:

- Drag-and-drop feature allows users to build a customized view of data
- Identify trading anomalies
- Analyze how alternative scenarios will affect performance using historical data

21) Advanced miner

An advanced miner is a useful tool for data processing, analysis, and modeling. Its user-friendly workflow interface allows you to explore various types of data.

Features:

- Extracting and saving data from/to different database systems, files, and data transformations
- Offers various operations on data, like sampling, joining datasets, etc.
- Helps you to build statistical models, variable importance analysis, clustering analysis, etc.
- Easy and effective Models' integration with external IT applications

22) Analytic Solver

Analytic Solver is free to use the point-and-click tool. It allows you to do risk analysis and prescriptive analytics in your browser. It offers full-power Data mining jobs.

Features:

- Helps you to incorporate uncertainty and solve with simulation optimization, stochastic programming, and robust optimization.
- Allows you to define the Monte Carlo simulation model using Excel formulas

23) PolyAnalyst

PolyAnalyst is the Data mining and analytical tool for extracting actionable knowledge hidden and actual structured of the data.

Features:

- Helps you to access data from various sources and merge data from different sources
- You can select from a broad selection of statistical and machine-learning algorithms.

- Offers you to create stuffing report which can be summarized and communicate your insight

24) Civis

Civis empowering you to make informed decisions with data scientist and decision market in mind. It allows your team to collaborate efficiently and find solutions faster.

Features:

- Offers architecture, products, and processes which helps you to protect your data
- You can configure with a library of data ingestion and ETL modules.
- Write code in a script, offers multiple scripts or jobs into a workflow, and define a workflow to run on a schedule.
- Allows you to turn your analysis and models into applications that run on a flexible, production-level infrastructure

25) Viscovery:

Viscovery is a workflow-oriented software suite. It is based on self-organizing maps and multivariate statistics for explorative data mining and predictive modeling. The system excels in intuitive user-guidance, mature implementation.

Features:

- An ideal project environment platform for goal-oriented operation
- Dedicated workflows that which allows you to offer focused navigation
- Clear workflow steps with proven default settings
- Workflow branching allowing generation of model variations
- Functions for integrated documentation and annotation
- Multiple handling tools to facilitate usage

7.8 APPLICATIONS OF DATA MINING

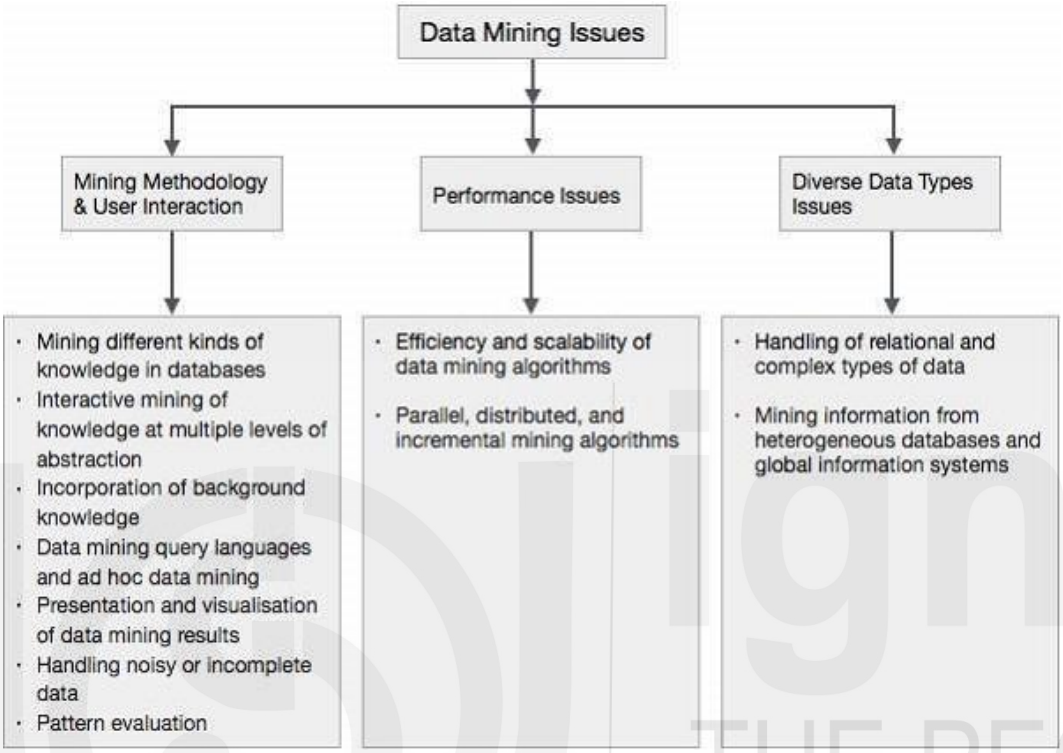
7.9 DATA MINING ISSUES

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.



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 −

- **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.

UNIT 8 DATA PREPROCESSING

8.0 INTRODUCTION

Preprocessing:

Real-world databases are particularly susceptible to noise, incomplete and inconsistent large data sets are popular as one can be a primary, secondary or tertiary source. Low quality data may result in low quality performance, thereby needing additional pre-processing data.

Data Preprocessing Techniques:

- Data cleaning may be used to prevent noise and to correct incoherence of data.
- Data aggregation combines data from multiple channels into a cohesive data store, for example a data center.
- Data transformations may be applied, such as standardization.
- Data reduction will, for example, reduce data size by the compilation, elimination of duplicate components or clustering; it can operate together.

Preprocessing Necessity:

Typical location properties in vast real-world datasets and databases are incomplete, chaotic and unfailing information.
Unfinished data can arise for a number of reasons:

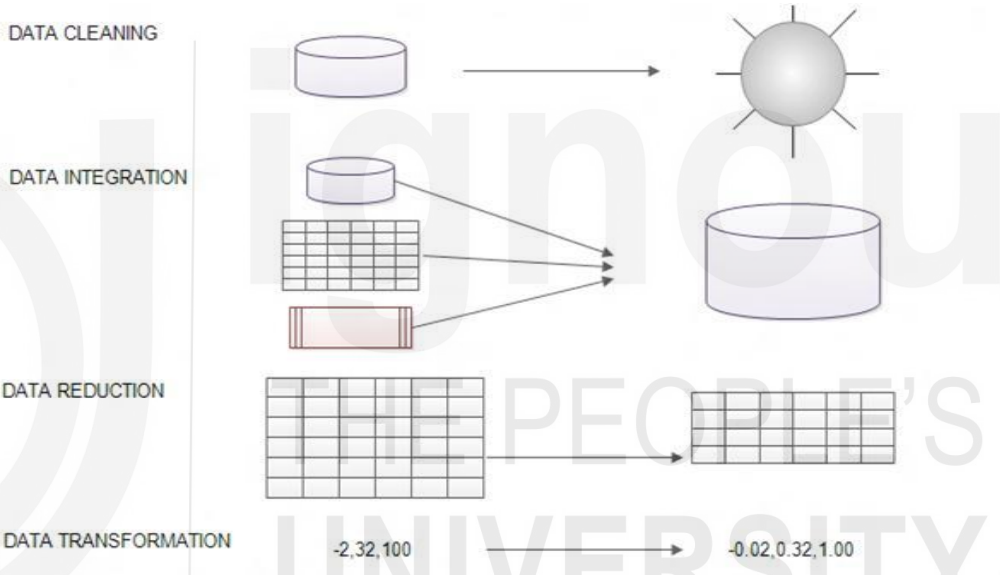
- Important attributes cannot always be usable.
- Valid data cannot be recorded because of misunderstanding or equipment failure.
- Data that disputed other recorded data should have been discarded.
- Missing data may be inferred, especially for tuples with missing values for certain attributes.
- Data collection techniques may be unreliable.
- At the moment of data entry, human or computer errors may have existed.
- Data processing failures can also occur.
- There might be technological disadvantages, such as narrow buffer sized for simultaneous data transfer and usage coordination.
- Data Routines for the cleaning of data are used by filling the missing values.
- Relieve noise effects, detect or remove outliers and address inconsistencies.
- Data integration is the hybrid process with many cubes or archives of databases. However, in multiple databases, those characteristics that define a

specific may have separate titles, which lead to inconsistencies and redundancies.

- ❖ Data transformation is a process approach such as standardizations and consolidation that constitutes additional preprocessing processes that contribute to mining process results.
- ❖ Data reduction obtains a simpler data collection image, which is significantly smaller in duration, but provides the same analytical efficiency (or nearly the same). A multitude of methods for data reduction are in use. This includes:

a) Data Aggregation (e.g., Creating a data cube).
b) Attribute subset selection (By similarity, eliminating unnecessary attributes)
c) Dimensionality Reduction (e.g., using encoding systems such as minimal encoding lengths or wavelets).
d) Numerosity Reduction (e.g., "Replace" details by alternating, smaller representations such as clusters or parametric structures.
e) Generalization (e.g., Data were minimized with the usage of the definition hierarchy)

Data discretion is a method of data reduction that is very useful for the automated generation of idea hierarchies from numerical data.



Forms of Data Preprocessing

8.1 DATA CLEANING

The other problem is that findings in the modern world are inconsistent, noisy and inexact. Software – cleaning (or data cleaning) procedures are intended to complete missing values, smooth out sounds as outliers are identified and data errors are corrected.

8.1.1 Missing Values:

Some tuples have no significance mentioned for a range of attributes, such as consumer income. We're going to fill up the vacant values.
The following methods execute missing values over a number of attributes.

a) Ignore the tuple: This is usually done if the class label is missing (assuming the mining activity requires classification). This solution is not quite good

since the tuple contains several attributes with missing values. It is really unfortunate since the percentage of values missed per variable varies greatly.

b) **Fill in the missing values manually:** This approach is time intensive and might not be feasible due to a wide range of data with many missing values.

c) **Use a global constant to fill in the missing values:** Replace all missing attributes with the same constant, such as a "unknown" or ∞.

d) **Use the attribute mean to fill in the missing value:** Suppose, for instance, that the gross customer income is $71,000. This importance compensates for the loss of revenue.

e) **Use the mean attribute for all samples belonging to the same class as the given tuple:** When the borrower is identified by credit danger, replace the missing number with the average income value in the same credit risk category as the tuple.

f) **Using the most likely meaning to fill in the missing value:** This can be computed using regression, inference-based tools using Bayesian formalism or the introduction of decision tree. For example, in the fixed Decision Tree the use of other market attributes is designed to approximate the missing revenue value.

8.1.2 Noisy Data:

Noisy is a measured random error or variance in the variable. Noise is minimized by techniques of data smoothing.

Binning: Binning uses the "neighborhood" to smooth the storage value of the records. That's the value around it. The sorted values are split into "buckets" and "bins." Since binning methods consult the value community, local smoothing is carried out.

Stored price details (in dollars $): 2, 4, 6, 8, 10, 12, 14, 16, 18.

Example: Partition in bins (Equal-Frequency):

Bin 1: 2, 4, 6
Bin 2: 8, 10, 12
Bin 3: 14, 16, 18

In the above process, the price data is first sorted and then partitioned into equivalent frequency bins of size 3.

Smoothing by bin mean:

Bin 1: 4, 4, 6
Bin 2: 10, 10, 10
Bin 3: 16, 16, 16

The method assumes that any value in a bin is replaced with the mean value of the bin as it is smoothed by bin. In bin 1, for example the mean of definitions 2, 4 and 6 is 4 [(2+4+6)/3].

Bin borders smoothing:

Bin 1: 2, 2, 4

Bin 2: 8, 12, 12
Bin 3: 14, 14, 18

The maximum and minimum values are defined as the boundary values when the bin boundary is smoothed. Each bin value is then replaced by the closest limit value. The larger the diameter, the more smoothing impact is. Bins will also be of similar width when the range of values in each bin remains unchanged.

Example: Delete the noise with smoothing techniques from the following details:
4, 2, 6, 10, 8, 16, 12, 24, 22, 14, 26.

Stored price details (in dollars $):
2, 4, 6, 8, 10, 12, 14, 16, 22, 24, 26
Partition in equal – frequency (equi-depth) bins:

Bin 1: 2, 4, 6, 8
Bin 2: 10, 10, 12, 14
Bin 3: 16, 22, 24, 26

Smoothing by bin mean:

Bin 1: 5, 5, 5, 5
Bin 2: 11, 11, 11, 11
Bin 3: 22, 22, 22, 22

Smoothing by bin boundaries:

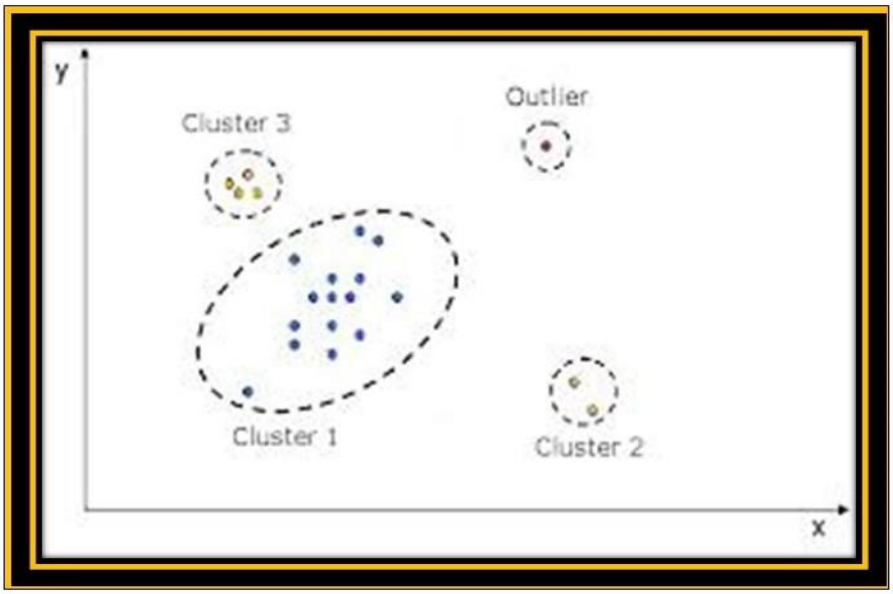
Bin 1: 2, 2, 2, 8
Bin 2: 10, 10, 14, 14
Bin 3: 16, 16, 16, 26

Regression:

By modifying data to perform, for instance for regression, data may be smoothed. Linear regression specifies that two attributes (or variables) be searched for the "right" line so one attribute can be used to predict the other.
Multiple linear regression is a linear regression extension, involving more than two attributes and outcomes that fit onto a multi-dimensional surface.

Clustering:

Clustering Outliers may be detected if the same values are clustered in groups or clusters. Values that go outside the spectrum of clusters can intuitively be considered outliers.



Outliers detected by clustering analysis

Inconsistent Data:

Information stored in the transaction contains inconsistencies. There are inconsistencies where there are conceptual dependence of attributes and incomplete values in data entry. Inconsistencies may be discovered and corrected through manual or information engineering approaches.

8.1.3 Data cleaning as a process:

- **Discrepancy Detection**
- **Data Transformation**

Discrepancy Detection:
The first step in data cleaning is to identify discrepancies. It considers the meta data awareness and examines the following rules in order to identify the distinction.

Specific guidelines:
Each attribute value must be distinct from each other attribute value.

Consecutive Laws:
If there are no missing values between the variable's lowest and highest values and all values must therefore be unique.

Null Rules:
Define the usage of blanks, question marks, special characters or other strings which can specify null status.

Discrepancy Detection Tools:

Data Scrubbing Tools:
Use specific domain details (e.g. postal address awareness and spelling control) to detect errors and correct details.

Data Auditing Tools:

Analyzes data to classify rules and relations and to evaluate data that violate those parameters.

Data Transformation:
This is the second step of the data cleaning process. After identifying inconsistencies, we have to define (sequence of) transformations and execute to correct them.

Data Transformation Tools:

Data Migration Tools:
Requires fundamental transformations such as replacing the 'gender' string with the 'sex' string to be described.

Extraction Transformation Loading (ETL):
Requires users to make specific user experience transformations (GUI).

Disadvantages of Data Cleaning Process:
- Due to missing records, analysts can lose track of feasible information. In situations where incomplete findings and outliers have been dropped, this is quite normal.
- It could contribute to an even bigger issue if automated.
- It takes time.
- The procedure is quite expensive.

8.3.1 Data Integration:

It is a methodology that combines data from various platforms into one device. These sources which include many cubes of records, databases or flat files.
The approach to data integration is officially classified as three times <G, S, M> where,
G Stands for the Global Scheme
S Stands for a Heterogeneous Source of a Schema
M Stands for Mapping between source and Global Schema Queries.
There are two main data integration methods: one is a "solid link approach," while the other is a "loose connection approach"

Tight Coupling:
The data warehouse here is regarded as a retrieval of records.
Data from different sources are combined to one physical location via ETL in this pairing (Extraction, Transformation and Loading phase).

Loose Coupling:
In this case the interface will translate the user query into the source database and send the query back to the source databases in order to get the answer.
And the data are only available in the real source repositories.

Data Integration Issues:
During data migration, there are no problems to consider:
- Schema Integration and Object Matching
- Redundancy
- Detection and Resolution of data value conflicts.

These are explained briefly as follows below:

1. **Schema Integration and Object Matching:**

It can be difficult because in different tables, different individuals can identify the same type. This is linked to the problem of individual identity. Metadata can be used to avoid errors in schema integration. Meta data can also be used for additional data conversation (e.g. where database codes "H" and "S" may be typed in one database, 1 & 2 in a separate database).

2. Redundancy:

That is another serious issue, that if it can be "derived" from another attribute, an attribute (e.g. annual sales) can be redundant. Inconsistencies in the naming of dimensions can also add to the redundancies in the data collection arising from this. A correlation analysis can detect any redundancies, given the two attributes, the test determines how strongly the one attribute indicates the other, based on the data available. The relationship between two attributes, X and Y, can be tested for numerical attributes by calculating the correlation coefficient:

$$\boldsymbol{r_{X,Y}} = \left(\sum(X - \overline{X})(Y - \overline{Y})\right)/(n-1)\sigma_A\sigma_B$$

Where

- n is the number of tuples
- $\overline{X}$ *means the value of* X
- $\overline{Y}$ *means the value of* Y
- σ_A *Standard deviation of* X
- σ_B *Standard deviation of* Y

$$\overline{X} = \frac{\sum X}{n};\ \sigma_X = \sqrt{\sum(X - \overline{X})^2}/(n-1)$$

$\boldsymbol{r_{X,Y}} > 0$ *then* X *and* Y *are positively correlated*

$\boldsymbol{r_{X,Y}} < 0$ *then* X *and* Y *are negatively correlated*

$\boldsymbol{r_{X,Y}} = \mathbf{0}$ *then on correlation between* X *and* Y

Detection and Resolution of Data Value Conflicts:

The third major issue in the integration of data is the detection and resolution of conflicts over data. For instance, the significance of attributes from different sources can vary for the same real person – the world person. This could be connected to changes in image, size or encoding. Room prices in various cities covering not only different currencies, but also different services (such as free breakfast) and taxes for the hotel chain. A lower abstraction level of an attribute can be recorded on one system than the same attribute on another system. Semantic heterogeneity and structure of data face major data integration problems.
A thorough compilation of data from various sources can help to mitigate and avoid the resulting data collection of redundancies and anomalies. This helps to improve the accuracy and speed of the mining process.

8.3.2 Data Transformation:

Data is translated into data mining methods. The data transformation consists of the following steps.

- Smoothing
- Aggregation
- Discretization
- Attribute Construction
- Generalization
- Normalization

Smoothing:

It is a mechanism used to remove data set noise using some algorithms to highlight main data set features. It helps to predict patterns and can be controlled when collecting data to eliminate or reduce any variance or noise.
The concept behind data smoothing is that simple changes that help to predict different trends and patterns can be detected. This helps investors or traders who want to look at a lot of data, which can be difficult to digest for patterns they would not see otherwise.

Aggregation:

The compilation or grouping of data is a way of summer data collection and show. In order to integrate these data sources into the concept of data analysis, the data should be obtained from different data sources. This is a vital step because the accuracy of data analysis depends heavily on the volume and quality of the data used. Reliable high quality and sufficient amounts of data should be collected to achieve the necessary results. Data collection is useful in all aspects including judgments on the financing of commodity, pricing, processes and marketing strategies or business plans. For example, revenue can be aggregated to calculate the monthly and annual total.

Discretization:

It is a way to transform continuous data into a number of short intervals. Many actual data mining activities have ongoing attributes. However, many of the existing data mining systems cannot handle these attributes.
While a data mining task can always handle a continuous attribute, a constant consistency attribute can significantly improve its efficiency by replacements of discrete values.

Attribute Construction:

Where new attributes are created and added to support an attribute set in the mining procedure. This simplifies the original data and improves the efficiency of mining.

Generalization:

Convert low - level data attributes with a hierarchical concept to high-level data attributes. For example, an age in numerical form initially (20, 64) is translated into a categorical sense (Young and Old).
Categorical attributes such as house addresses, for example, may be applied to higher-level definitions such as cities or areas.

Normalization:

All data variables need to be translated into a given set. The standardization approach used is:

- Min – Max Normalization
- Z – Score Normalization
- Decimal Scaling Normalization

Min – Max Normalization:

- Which translates the initial data linearly.

- Suppose min X is the minimum and max X is the maximum limit of the variable.
- We've got the formula

$$v' = \frac{v - minX}{maxX - minX}(\text{new}\_max_X - \text{new}\_min_X) + \text{new}\_min_X$$

- where **v** is the value that you want to plot in the new set.
- v' is the fresh meaning you get when you normalize the old value.

Example: Assume the attribute revenue minimum and maximum values are 1000 and 16000 respectively. Diagram income in the [0.0, 1.0] range. For salaries a value of 14000 is converted into a minimum – maximum standardization.

$$= \frac{14000 - 1000}{16000 - 1000}(1.0 - 0) + 0$$

$$= 0.866$$

Z-Score Normalization:

- The value of a variable (X) in the Z-Score Normalization or Zero-Median Normalization is uniform according to the mean of the X and its standard deviation.
- The X value v of the X attribute is computer-standardized to V.

$$v' = \frac{v - \overline{X}}{\sigma X}$$

Where X and σ are the mean and standard deviations respectively of the X attribute. This method is helpful if the true minimum and maximum attribute X are not certain or if outliers surpass the minimum – maximum normalization.

Example: Suppose that the average and standard deviation in attribute revenue number are respectively 22,000 and 7,000. In the case of Z-Score standardization, the revenue of 42,000 is transferred to Z-Score.

$$\frac{42000 - 22000}{7000} = 2.85$$

Decimal Scaling Normalization:

- Standardizes the variable values by changing their decimal position.
- You can measure the number of points that the decimal point is passed by the absolute maximum value of the X attribute.
- The value of v of the X attribute is computer-standardized to v'.

$$v' = \frac{v}{10^j}$$

J is such an integer that maximum ($|v'|$) < 1.

Example:

- Suppose the observed values differ between the lowest - 486 and the highest - 417.
- The X meaning square is 486. In order to normalize the calculation by decimal scaling, each value must be divided by 1000 so that -486 standardizes –0.486 and 417 to 0.417.

8.4 DATA REDUCTION

Data reduction approaches can be used to simplify the data collection representation, which is much smaller in length and thus maintain accuracy of the original data - the reduced data set mining can be much more efficient, but the analytical output can be the same (or almost the same).

Strategies of Data Reduction:

- Data cube aggregation, which applies aggregation of data during data cube construction.
- Set of sub-set attributes, which may be detected and removed by obsolete, weakly significant or redundant measurements.
- Dimensionality reduction where encoding approaches are used to reduce the data collection size to a minimum.
- Reduction of numbers when alternative, smaller data representation replaces or predicts information.
- Discretization and category hierarchy generation where raw data values for the attributes are replaced by ranges or higher logical levels.

8.4.1 Data cube Aggregation:

Data Cube aggregation where a data cube is created using aggregation operations. Data cubes store multidimensional information aggregated. Each cell stores in a multi-dimensional space a data value corresponding to the data point.
For each attribute concept hierarchies may occur, allowing data to be evaluated at multiple abstraction levels. Data cubes provide quick access to pre-computed summary data; this will support both online analytical analysis and data mining.
The cube is available in three forms:

Based cuboid:

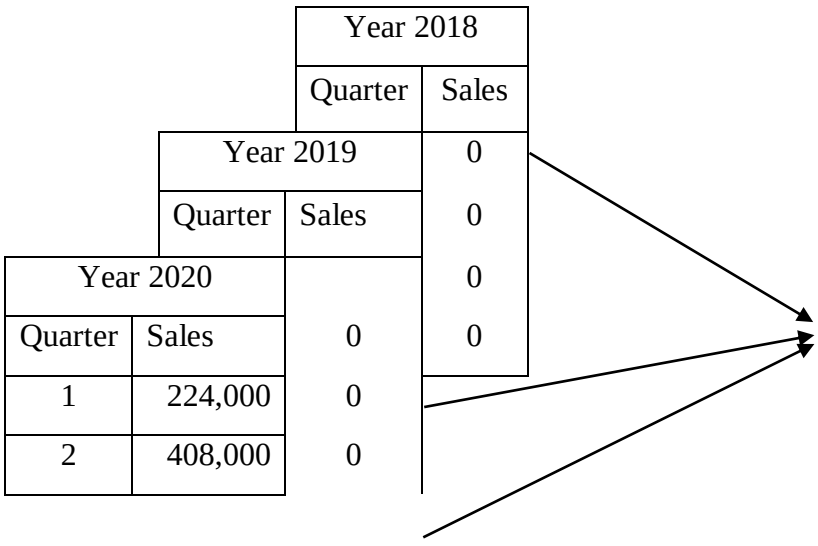
The cube is called the based cuboid, at least in its complexity.

Lattice of cuboid:

Details of the chart of different abstraction layers are often called cubes.

Apex cuboid:

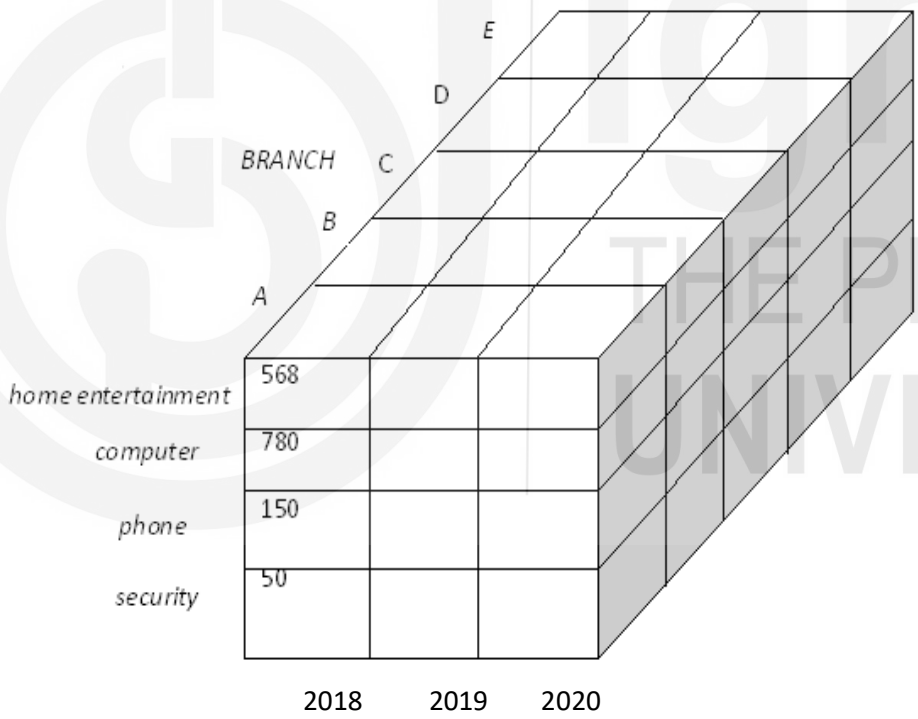
A cube is an apex cuboid in the highest degree of complexity.



3	350,000	0
4	586,000	

Year	Sales
2018	1,568,000
2019	2,356,000
2020	3,594,000

Sales are shown per quarter on the left. The details are aggregated for the annual sales on the right.



A data cube for sales

8.4.2 Attribute Subset Selection:

The selection of a sub-set attribute reduces data by removing redundant attributes or measurements irrelevant. The goal is to find a minimum set of attributes for the selection of a subset attribute. It decreases the number of attributes found in patterns, which promotes comprehension of patterns.

The data set will include a wide range of attributes. Some of them may be outdated or redundant, however. The objective of selecting attributes of sub-sets is to select a minimum number of attributes to lower data collection costs and to reduce loss of

such unwanted attributes. The reduced data collection also enables the discovered pattern to be explained.

Process of Attribute Subset Selection:

The brute force approach can be very expensive, in which data with n attributes can be evaluated by any subset (2^n possible subsets). The simplest way to do this is to use metrics of predictive significance in order to recognise the best or worst attributes. The statistical significance test indicates that the characteristics are separate. This is a greedy approach where the statistically ideal value of the meaning level is determined by 5% and the models are tested time and time again until the p value of all the attributes is less than or equal to the selected meaning level. Attributes with a p-value above the value are discarded. This method is repeated time and time again until all attributes of the data set have a p-value that is less than or equal to their importance. This allows us to obtain reduced data with no irrelevant attributes.

Attribute Subset Selection Methods:

- Stepwise Forward Selection
- Stepwise Backward Elimination
- Combination of Forward Selection and Backward Elimination
- Decision Tree Introduction

All of the above methods are greedy approaches to the selection of attribute subsets.

Stepwise Forward Selection:

The process starts with a minimum number of empty attributes. The most significant attributes with minimal p-value are selected and added to the minimum list. One attribute in each iteration is applied to the reduced collection.

Stepwise Backward Elimination:

All attributes are considered here in the initial collection of attributes. In each iteration, one attribute is omitted from the set of attributes of p significance.

Combination of Forward Selection and Backward Elimination:

Phase by step, sorting and reverse exclusion are combined so that the necessary attributes are selected more effectively. This is the most common way to collect attributes.

Decision Tree Introduction:

The introduction to the Decision Tree outlines a process flow diagram in which each internal node denotes an attribute test, every branch defines the possible test results and each leaf denotes the class forecast. For each node, the algorithm selects the best attribute to divide the data into different classes. From the data given. A tree is established. The tree contains a collection of attributes from the decreased sub-set attribute. The threshold measurement is used as a stop criterion.

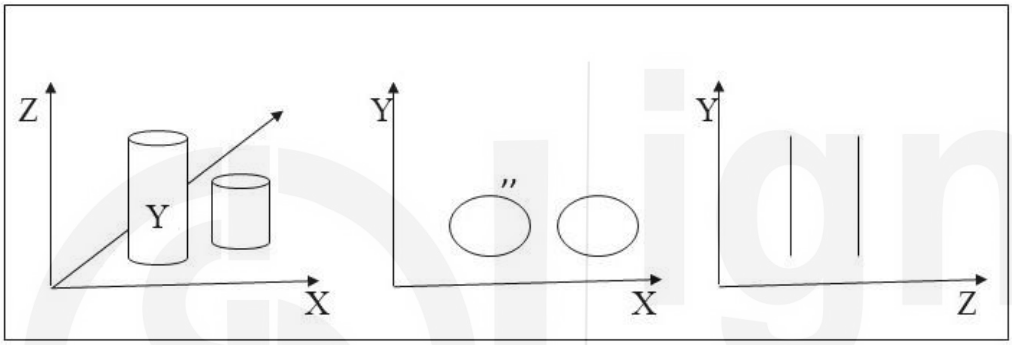
8.4.3 Dimensionality Reduction:

The final definition of machine learning problems is sometimes underlying too many factors. These factors are primarily variables known as features. The higher the number of features, the harder the training set can be imagined and worked on. Most

of these features are frequently connected and therefore redundant. This is where algorithms of dimensional reduction come into play. The reduction in dimension is the method by obtaining a collection of main variables for reducing the number of random variables. It can be divided into feature selection and feature extraction.

Why is reduction in dimensionality significant in Machine Learning:

An simple explanation of reduced dimensionality is given by a basic e-mail classification problem, where we have to classify whether mail is spam or not. There is a wide range of features, including whether the e-mail has a general title, e-mail content, whether an e-mail uses a template, etc. However, both of these functions will overlap. In other cases, a question of moisture and rainfall classification would fall into only one basic characteristic, as the above two correlates strongly. Therefore, in such a problem, we can reduce the number of features. It is difficult to see a problem in 3-D classification if a 2-D can be mapped into a basic 2-D area with the problem of 1-D in a simple diagram. The following figure shows that a 3-D function space is divided into two 1-D function spaces, and the number of characteristics can be further decreased if they are associated.



Dimensionality Reduction

Components of Dimensionality Reduction:

Two-dimensional reduction components are available:

- Feature Selection
- Feature Extraction

Feature Selection:

Here we try to fine-tune a subset of the original variables or features array to create a smaller subset that can be used for the query. It normally takes three directions:

a) **Filter**
b) **Wrapper**
c) **Embedded**

Feature Extraction:

This limits data to a lower dimension in a high dimensional space, i.e. space with a limited number of dimensions.

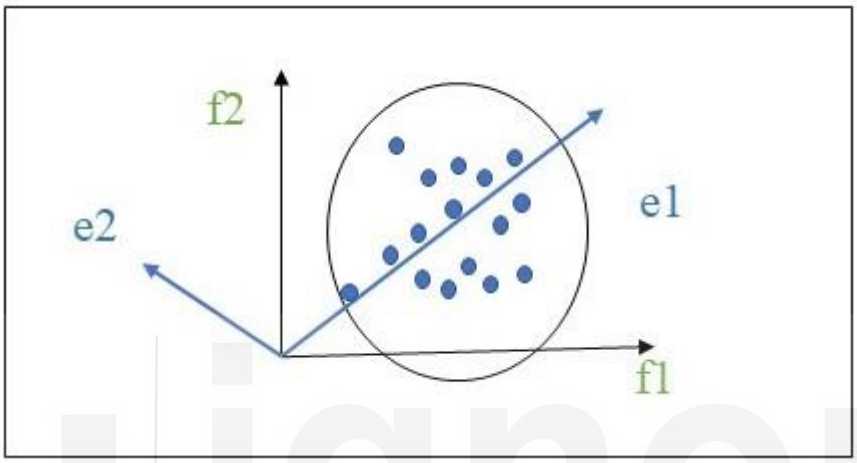
Dimensionality Reduction Methods:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA)

Dimensionality reduction may be linear or not linear depending on the method employed. Below is the primary linear method, referred to as the principal component analysis or PCA.

Principal Component Analysis (PCA):

Karl Pearson has established this method. Works under the condition that the data variance in the lower dimensional space is mapped to a data in a lower dimensional space.



This includes the following steps:

- Build a data covariance matrix
- Calculate the patented vectors in this matrix
- Self – vector equivalent to the largest own values are used to recreate a significant fraction of the variance of the original results.

We then have a reduced number of proprietary vectors, and the method may have lost some data. The big variances should be maintained and the remaining own vectors preserved.

Advantages and Disadvantages of Dimensionality Reduction:

- It helps to compact data and also removes disc space.
- It reduces the time of calculation
- It also helps, if any, to eliminate redundant features.
- This can lead to a certain amount of loss of data.
- PCA continues to recognize ongoing correlations between variables that are often unwanted.
- If mean and covariance are not sufficient to classify datasets, PCA fails.
- We cannot see how many of the primary components for maintaining such thumb laws are applied.

8.4.4 Numerosity Reduction:

It is a technique for data reduction that replaces the original information with a smaller kind of data representation. There are two ways to minimize numbers.

- Parametric Methods
- Non-Parametric Methods

Parametric Methods:

Data is interpreted by any parametric system model. The model is used to simulate data, which means that only data parameters should be processed rather than actual data. To construct these models, regression and log-linear approaches are used.

Regression:

Regression can be a simple linear regression or a multiple linear regression. If there is only one independent variable, the regression model is referred to as simple linear regression, and if there are more than one independent parameter, such models are called multiple linear regression.
The data is modelled on a straight line in linear regression. For example, the random variable y can be modelled as a linear function of another random variable x, with the equation y=a x+b where the line slopes and y – intercepts are determined by a and b (regression coefficients). Y is modelled as a linear function in multiple linear regression, with two or more (independent) prediction variables.

Log-Linear Model:

For a series of discrete attributes based on a smaller subset of dimensional combinations, a log-linear model can be used to predict the probability of each data point in a multidimensional space. This allows the development of higher dimensional data space from lower dimensional attributes.
The regression as well as the log - linear model on sparse data can be used, but their use can be limited.

Non-Parametric Methods:

These techniques are used to store reduced data representations such as histograms, clusters, samples, and aggregation of data cubes.

Histograms:

Histogram is a frequency overview of the data. It uses binning to estimate the distribution of data that is a popular way of reducing data.

Clustering:
Clustering separates data into groups or clusters. This method divides all data into individual clusters. To keep specifics to a minimum. The representation of cluster data is used to bypass actual data. It also helps to locate outliers of data.
Sampling:

Sampling can be used for data reduction, as a wider range of data can be represented by a much smaller random sample (or subset).

Data cube Aggregation:

Data Cube Aggregation means data transfer to a reduced number of measurements from a complex stage. The resulting data collection is smaller without the information required for the study mission.

Parametric Model:

Regression:

The problem with regression is where the output variable is a real or permanent feature, like "wage" or "weight." There are several different approaches, the simplest

being linear regression. It tries to align the data with the right hyperplane travelling across the dots.

Regression Analysis:

A statistical way to estimate the interaction of dependent variables or criterion variables with one or more separate variables or predictors. The regression analysis explains adjustments for changes in the selected predictors in the parameters. Conditional predictor assumptions - parameters based on which the average value of the dependent variables is given when the individual variables are changed. Three main applications for the regression analysis are predictor strength, effect prediction and trend forecasting.

Types of Regression:

- Linear regression
- Logistic regression
- Polynomial regression
- Stepwise regression
- Ridge regression
- Lasso regression
- Elastic Net regression

Linear Regression:

This Linear Regression is used for statistical software. Linear regression is a linear approach for modelling an interaction with many predictors and explanatory variables between the criterion or scalar answer. Linear regression depends on the conditional distribution of the probability of the result given the predictor values. There is a chance that linear regression would overfit. The formulation of linear regression is:

$$y' = bX + A$$

Logistic Regression:

It shows that the dependent variable is assessed as a dichotomy. Logistic regression measures the parameters of the logistic model and defines a binomial regression. Logistic regression would be used to discuss data with two possible parameters and the association of predictor criteria. The equation of logistic regression is:

$$l = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Polynomial Regression:

It is used for curvilinear information. The least square method is ideal for polynomial regression. The object of regression analysis is to model the expected value of the dependent variable y in relation to the independent variable X. The equation of polynomial regression is:

$$l = \beta_0 + \beta_1 x_1 + €$$

Stepwise Regression:

It is used with predictive models to balance regression models. It is done automatically. The variable is added or removed from the explanatory variables collection for each point. The methods of progressive regression are forward collection, retroactive exclusion and bi-directional elimination. The step-by-step regression is:

$$b_{j.std} = b_j(S_x * S_y^{-1})$$

Ridge Regression:

It is a method for studying the effects of multiple regressions. When there is a multi-linearity, the least square figures are unbiased. A certain degree of bias is added to the regression equations, which removes standard errors. The method of ridge regression is:

$$\beta = (X^T X + \lambda * I)^{-1} X^T y$$

Lasso Regression:

An approach to regression analysis that performs both variable selection and regularisation. It uses soft boundaries. It only selects a subset of covariates to use the final construct. The equation of Lasso regression is:

$$N^{-1} \sum_{i=1}^{N} f(x_i, y_{I,} \alpha, \beta)$$

Elastic Regression:

It is a regularized approach to regression that integrates lasso and ridge penalties linearly. It supports the creation of vector machines, metrics and portfolio optimization. The penalty's job is as follows:

$$|| \beta ||_1 = \sum_{j=1}^{p} |\beta_j$$

Data Discretization and Concept Hierarchy Generation:

Data Discretization methods may be used by dividing the collection of attributes into intervals to minimize the number of values for a given continuous attribute. Interval marks can be used to bypass data values.

Data Discretization features:

- Leads to a concise
- Easy – to – use
- Knowledge – Level representation of mining results

Data Discretization Categories:

- **Supervised Discretization:** Uses details regarding the class
- **Unsupervised Discretization or Splitting:** Does not use class knowledge
- **Top-Down Discretization or Splitting:** Here, the method begins by first finding one or a few points called break points or cut points to separate the whole set of attributes and then performs this recursively over the subsequent intervals.
- **Bottom – up Discretization or merging:** Here, the method starts by treating all continuous quantities as intervals and then applies this process recursively at the following intervals.

Concept Hierarchy:

A hierarchy of the description defines a discretionary attribute to a certain numerical characteristic. It can be used to reduce data by collecting and replacing low-level concepts such as age-attribute numeric values with higher – low, middle and high levels concepts.

Discretization and concept Hierarchy Generation for Numerical data:

- Numerical attributes concept hierarchies may be created automatically based on data discretization.
- Management approaches for qualitative data over definition hierarchy.

- Binning
- Histogram Analysis
- Entropy – Based Discretization
- X^2 – Merging
- Cluster Analysis
- Discretization by Intuitive Partitioning

Binning:

It's a cutting technique that concentrates on the number of bins listed. Session techniques are also used as arbitrary methods to reduce numbers and generate hierarchical descriptions. These methods can be used recursively to generate Hierarchical definitions on the resulting partitions. It does not use class information and is therefore an arbitrary technique that is not tracked. It is adaptable to the number of containers identified by the customer as outliers.

Histogram Analysis:

Like binning, histogram analysis is an unexpected discretization tool when class information is not used. Histograms partition the value for the disjoint range attribute A named buckets. The histogram analysis algorithm can be applied recursively to each partition to automatically produce a multi-level concept hierarchy that ends after a certain number of concept levels have been achieved.

Entropy – Based Discretization:

It is one of the most commonly used discretization measures. Entropical discretization is a manual top-down separation technique. It addresses the knowledge of class distribution in its assessment and determination of divisions. The system selects the A value in order to discern the numerical function, which has the lowest entropy as a divisive point and corrects the resulting intervals to achieve a hierarchical discernment. This discretization forms a meaning hierarchy for A. Let D consist of a list of attributes and a data tuples defined for the label-class attributes. The simple way to distinguish an A attribute in the entropy set is as follows:

1. Each value of A can be seen as a potential interval or dividing point to separate the A spectrum. In other words, a split point for A divides tuples into two sub-sets in D, corresponding with conditions A <split point and A> and produces a binary disc recitation.
2. Suppose we divide the tuples into D with A and some split-points. Ideally, the partitioning of this would lead to an exact tuple classification. For instance, if we had two classes, we expect to have all class C1 tuples on one partition and all class C2 tuples on the other.
3. The method for the description of points of division is applied recursively before such stop thresholds are reached, for example when minimum information is fulfilled. The criterion is less than a small one, e or if the number of intervals is greater than a max interval should be required for all candidate split points.

Interval merging by χ 2 Analysis:

- Chi blends the fastest way to pick the finest.
- Nearby loops and then mix recurrently in larger intervals. The method is regulated using classes.
- The Chi merge form is as follows:
- Initially, each independent value of a numerical attribute A is considered as one interval.
- For each adjacent pair of intervals, C2 monitoring is performed.

- Next intervals are combined with the lowest c2 values, since a couple of low c2 values show that the distributions are similar.
- A predefined stop criterion follows the fusion phase recursively.

Cluster Analysis:

Cluster analysis is a traditional technique of discretion. The A values can be split into clusters or groups to determine the numerical A attribute by a clustering algorithm. Clustering can be used to construct a concept hierarchy for A either by a top-down division technique or by a top-down approach in the form of a hierarchy node in each cluster. Each initial cluster or partition can be further broken down into several subclusters that form a lower hierarchical level. Repeated groups of neighbouring clusters form clusters for the bottom-up approach to fusion.

Discretization by intuitive partitioning:

Although the above discretization methods aid in numerical hierarchy, many users want numerical ranges divided into a very uniform, intuitive and natural interpretation of cycles.
The 3-4-5 rule can be used for a fairly uniform, natural appearance of numerical results. The rules generally divide a given data set by level according to the most important digit range into 3, 4 or 5 periods.
The fact is this:

- When a set of 3,6,7 or 9 is divided into three intervals on the largest digit (3 equal-width intervals for 3,6 and 9; and 3 intervals in the grouping of 2-3-2 for 7)
- If it covers 2, 4 or 8 separate values at the maximum digit, split the spectrum into 4 periods of equivalent duration.
- Split the spectrum into five equal-width increments if the largest digit spans 1,5 or 10 separate values.
- To establish a definition hierarchy for the numerical attribute assigned, the rule can be extended recursively for each interval.
- Real-world knowledge also includes extremely large positive or negative outer values based on minimum and maximum data values, which can distort any top-down discretionary process.

Concept Hierarchy Generation for categorical data:

Discrete data are categorical data. There is a small (but probably large) amount of different values without an order among the category attribute values. Methods for generating categorical data classification hierarchy.

- Description of a partial scheme level attribute by customers or experts. Typical hierarchies or measurements are a list of attributes. By specifying a partial or full plan – the user or expert may easily describe a definition hierarchy in a level order of attributes.
- This classification is basically a manual explanation of the main hierarchy segment by correctly aggregating data for the part of a hierarchy.
- A complete hierarchy of the definition of a simple value list is difficult to describe in a broad database. Therefore, explicit groupings are practical in deciding on a small number of mid-level outcomes.
- The user can define a number of attributes that form a description hierarchy but do not specify its partial order. Structure should try to produce the attribute for a meaningful hierarchy of the meanings immediately.

UNIT 9 MINING FREQUENT PATTERNS AND ASSOCIATIONS

9.1 ASSOCIATION RULE MINING

- Association rule mining is a popular and well researched method for discovering interestingrelations between variables in large databases.
- It is intended to identify strong rules discovered in Databases using different measures of interestingness.
- Based on the concept of strong rules, Rakesh Aggarwal et al. introduced association rules.

9.1.1. Problem Definition

The problem of association rule mining is defined as:

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called *items*.

Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the *database*.

Each transaction in D has a unique transaction ID and contains a subset of the items in I.

A *rule* is defined as an implication of the form $X \Rightarrow Y$

where $X, Y \subseteq I$ and $X \cap Y = \emptyset$.

The sets of items (for short *itemsets*) X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

Example:

To illustrate the concepts, we use a small example from the supermarket domain. The set of items is $I = \{\text{milk}, \text{bread}, \text{butter}, \text{beer}\}$ and a small database containing the items (1 codes presence and 0 absence of an item in a transaction) is shown in the table.

An example rule for the supermarket could be $\{\text{butter}, \text{bread}\} \Rightarrow \{\text{milk}\}$ meaning that if butter and bread are bought, customers also buy milk.

Example database with 4 items and 5 transactions

Transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

9.2 IMPORTANT CONCEPTS OF ASSOCIATION RULE MINING

- The **support** $\text{supp}(X)$ of an item X set is defined as the proportion of transactions in thedata set which contain the item set. In the example database, the item set $\{\text{milk}, \text{bread}, \text{butter}\}$ has a support of $1/5 = 0.2$ since it occurs in 20% of alltransactions (1 out of 5 transactions).

- The **confidence** of a rule is defined

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X).$$

For example, the rule $\{\text{butter}, \text{bread}\} \Rightarrow \{\text{milk}\}$ has a confidence of $0.2/0.2 = 1.0$ in the database, which means that for 100% of the transactions containing butter and bread the rule is correct (100% of the times a customer buys butter and bread, milk is bought as well). Confidence can be interpreted as an estimate of the probability $P(Y|X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

- The ***lift*** of a rule is defined as

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

or the ratio of the observed support to that expected if X and Y were independent. The rule $\{\text{milk}, \text{bread}\} \Rightarrow \{\text{butter}\}$ has a lift of $\frac{0.2}{0.4 \times 0.4} = 1.25$.

The **conviction** of a rule is defined as

$$\text{conv}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)}$$

The rule $\{\text{milk, bread}\} \Rightarrow \{\text{butter}\}$ has a conviction of $\frac{1 - 0.4}{1 - .5} = 1.2$, and can be interpreted as the ratio of the expected frequency that X occurs without Y (that is to say, the frequency that the rule makes an incorrect prediction) if X and Y were independent divided by the observed frequency of incorrect predictions.

9.3 MARKET BASKET ANALYSIS

This process analyzes customer buying habits by finding associations between the different items that customers place in their shopping baskets. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket. Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.

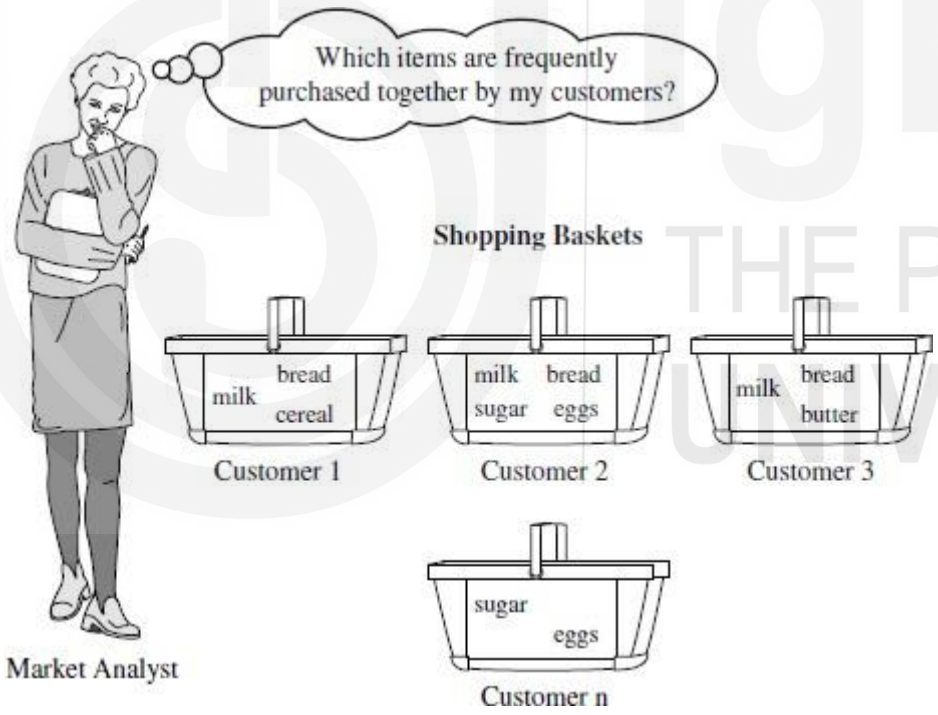


Fig. 9.1. Market Basket Analysis

Example:

If customers who purchase computers also tend to buy anti-virus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items. In an alternative strategy, placing hardware and software at opposite ends of the store may entice customers who purchase such items to pick up other items along the way. For instance, after deciding on an expensive computer, a customer may observe security systems for sale while heading toward the software display to purchase antivirus software and may decide to purchase a home security system as well. Market basket analysis can also help retailers plan

which items to put on sale at reduced prices. If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers *as well as* computers.

9.4 FREQUENT PATTERN MINING

Frequent pattern mining can be classified in various ways, based on the following criteria:

1. Based on the completeness of patterns to be mined:

- We can mine the complete set of frequent item sets, the closed frequent item sets, and themaximal frequent item sets, given a minimum support threshold.
- We can also mine constrained frequent item sets, approximate frequent item sets, near-match frequent item sets, top-k frequent item sets and so on.

2. Based on the levels of abstraction involved in the rule set:

Some methods for association rule mining can find rules at differing levels of abstraction.

For example, suppose that a set of association rules mined includes the following ruleswhere X is a variable representing a customer:

buys(X, ‒computer‖))=>buys(X, ‒HP printer‖) (1)

buys(*X*, ‒*laptop computer*‖)) =>*buys*(*X*, ‒*HP printer*‖) (2)

In rule (1) and (2), the items bought are referenced at different levels of abstraction (e.g.,

‒*computer*‖ is a higher-level abstraction of ‒*laptop computer*‖).

3. Based on the number of data dimensions involved in the rule:

- If the items or attributes in an association rule reference only one dimension, then it is asingle-dimensional association rule.
  buys(X, ‒computer‖))=>buys(X, ‒antivirus software‖)

- If a rule references two or more dimensions, such as the dimensions age, income, and buys,then it is a multidimensional association rule. The following rule is an example of a multidimensional rule:
  age(X, ‒30,31…39‖) ^ income(X, ‒42K,…48K‖))=>buys(X, ‒high resolution TV‖)

4. Based on the types of values handled in the rule:

- If a rule involves associations between the presence or absence of items, it is a Booleanassociation rule.

- If a rule describes associations between quantitative items or attributes, then it is aquantitative association rule.

5. Based on the kinds of rules to be mined:

- Frequent pattern analysis can generate various kinds of rules and other interestingrelationships.
- Association rule mining can generate a large number of rules, many of which areredundant or do not indicate a correlation relationship among item sets.
- The discovered associations can be further analyzed to uncover statistical correlations,leading to correlation rules.

6. Based on the kinds of patterns to be mined:

- Many kinds of frequent patterns can be mined from different kinds of data sets.
- Sequential pattern mining searches for frequent subsequences in a sequence data set, where a sequence records an ordering of events.
- For example, with sequential pattern mining, we can study the order in which items are frequently purchased. For instance, customers may tend to first buy a PC, followed by a digital camera, and then a memory card.
- Structured pattern mining searches for frequent substructures in a structured data set.
- Single items are the simplest form of structure.
- Each element of an itemset may contain a subsequence, a sub tree, and so on.
- Therefore, structured pattern mining can be considered as the most general form of frequent pattern mining.

9.5 EFFICIENT FREQUENT ITEMSET MINING METHODS:

FINDING FREQUENT ITEM SETS USING CANDIDATE GENERATION: THEAPRIORI ALGORITHM

- Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for miningfrequent itemsets for Boolean association rules.
- The name of the algorithm is based on the fact that the algorithm uses *prior knowledge* offrequent itemset properties.
- Apriori employs an iterative approach known as a *level-wise* search, where *k*-item sets are used to explore (*k*+1)-itemsets.
- First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted *L*1.Next, *L*1 is used to find *L*2, the set of frequent 2-itemsets, which is used to find *L*3, and so on, until no more frequent *k*-item sets can be found.

- The finding of each L_k requires one full scan of the database.
- A two-step process is followed in Apriori consisting of join and prune action.

Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

- D, a database of transactions;
- min_sup, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

```
(1)   L1 = find_frequent_1-itemsets(D);
(2)   for (k = 2; Lk−1 ≠ ϕ; k++) {
(3)      Ck = apriori_gen(Lk−1);
(4)      for each transaction t ∈ D { // scan D for counts
(5)         Ct = subset(Ck, t); // get the subsets of t that are candidates
(6)         for each candidate c ∈ Ct
(7)            c.count++;
(8)      }
(9)      Lk = {c ∈ Ck | c.count ≥ min_sup}
(10)  }
(11)  return L = ∪k Lk;

procedure apriori_gen(Lk−1:frequent (k − 1)-itemsets)
(1)   for each itemset l1 ∈ Lk−1
(2)      for each itemset l2 ∈ Lk−1
(3)         if (l1[1] = l2[1]) ∧ (l1[2] = l2[2]) ∧ ... ∧ (l1[k − 2] = l2[k − 2]) ∧ (l1[k − 1] < l2[k − 1]) then {
(4)            c = l1 ⋈ l2; // join step: generate candidates
(5)            if has_infrequent_subset(c, Lk−1) then
(6)               delete c; // prune step: remove unfruitful candidate
(7)            else add c to Ck;
(8)         }
(9)   return Ck;

procedure has_infrequent_subset(c: candidate k-itemset;
          Lk−1: frequent (k − 1)-itemsets); // use prior knowledge
(1)   for each (k − 1)-subset s of c
(2)      if s ∉ Lk−1 then
(3)         return TRUE;
(4)   return FALSE;
```

Fig. 9.2. Pseudo-code for Apriori Algorithm

Example:

TID	List of item IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

There are nine transactions in this database, that is, |D| = 9.

Steps:

1. In the first iteration of the algorithm, each item is a member of the set of candidate1- itemsets, C1. The algorithm simply scans all of the transactions in order to count the number of occurrences of each item.

2. Suppose that the minimum support count required is 2, that is, min sup = 2. The set of frequent 1-itemsets, L1, can then be determined. It consists of the candidate 1-itemsets satisfying minimum support. In our example, all of the candidates in C1 satisfy minimum support.

3. To discover the set of frequent 2-itemsets, L2, the algorithm uses the join L1 on L1to generate a candidate set of 2-itemsets, C2.No candidates are removed fromC2 during the prune step because each subset of the candidates is also frequent.

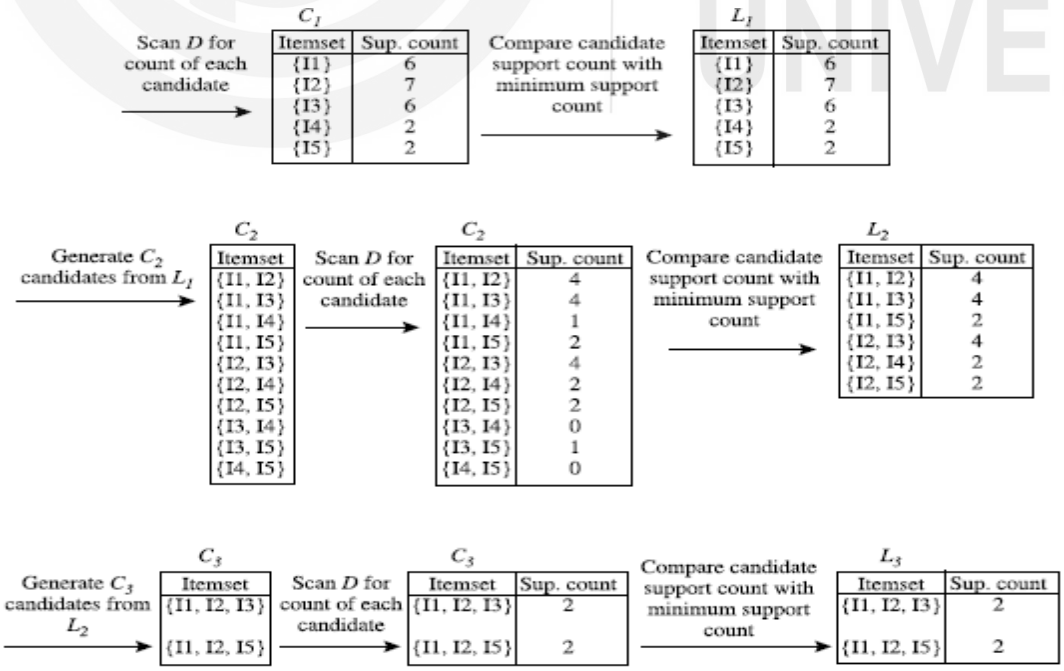
4. Next, the transactions in D are scanned and the support count of each candidate itemsetInC2 isaccumulated.

5. The set of frequent 2-itemsets, L2, is then determined, consisting of those candidate2- item sets in C2 having minimum support.

6. The generation of the set of candidate 3-itemsets,C3, From the join step, we first getC3 =L2x L2 = ({I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4},{I2, I3, I5}, {I2, I4, I5}. Based on the Apriori property that all subsets of a frequent item-set must also be frequent, we can determine that the four latter candidates cannot possibly be frequent.

7. The transactions in D are scanned in order to determine L3, consisting of those candidate3-itemsets in C3 having minimum support.

8. The algorithm uses L3x L3 to generate a candidate set of 4-itemsets, C4.



Generation of candidate itemsets and frequent itemsets, where the minimum support count is 2.

(a) Join: $C_3 = L_2 \bowtie L_2 = \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\} \bowtie \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\} = \{\{I1, I2, I3\}, \{I1, I2, I5\}, \{I1, I3, I5\}, \{I2, I3, I4\}, \{I2, I3, I5\}, \{I2, I4, I5\}\}$.

(b) Prune using the Apriori property: All nonempty subsets of a frequent itemset must also be frequent. Do any of the candidates have a subset that is not frequent?

- The 2-item subsets of {I1, I2, I3} are {I1, I2}, {I1, I3}, and {I2, I3}. All 2-item subsets of {I1, I2, I3} are members of L_2. Therefore, keep {I1, I2, I3} in C_3.
- The 2-item subsets of {I1, I2, I5} are {I1, I2}, {I1, I5}, and {I2, I5}. All 2-item subsets of {I1, I2, I5} are members of L_2. Therefore, keep {I1, I2, I5} in C_3.
- The 2-item subsets of {I1, I3, I5} are {I1, I3}, {I1, I5}, and {I3, I5}. {I3, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I1, I3, I5} from C_3.
- The 2-item subsets of {I2, I3, I4} are {I2, I3}, {I2, I4}, and {I3, I4}. {I3, I4} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I3, I4} from C_3.
- The 2-item subsets of {I2, I3, I5} are {I2, I3}, {I2, I5}, and {I3, I5}. {I3, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I3, I5} from C_3.
- The 2-item subsets of {I2, I4, I5} are {I2, I4}, {I2, I5}, and {I4, I5}. {I4, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I4, I5} from C_3.

(c) Therefore, $C_3 = \{\{I1, I2, I3\}, \{I1, I2, I5\}\}$ after pruning.

Generation and pruning of candidate 3-itemsets, C_3, from L_2 using the Apriori property.

Fig. 9.3. Working of Apriori Algorithm

9.6 GENERATING ASSOCIATION RULES FROM FREQUENT ITEM-SETS

Once the frequent item-sets from transactions in a database *D* have been found, it isstraightforward to generate strong association rules from them.

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

The conditional probability is expressed in terms of itemset support count, where $support\_count(A \cup B)$ is the number of transactions containing the itemsets $A \cup B$, and $support\_count(A)$ is the number of transactions containing the itemset A. Based on this equation, association rules can be generated as follows:

- For each frequent itemset l, generate all nonempty subsets of l.
- For every nonempty subset s of l, output the rule "$s \Rightarrow (l - s)$" if $\frac{support\_count(l)}{support\_count(s)} \geq$ *min_conf*, where *min_conf* is the minimum confidence threshold.

Example:

Generating association rules. Let's try an example based on the transactional data for *AllElectronics* shown in Table 5.1. Suppose the data contain the frequent itemset $l = \{I1, I2, I5\}$. What are the association rules that can be generated from l? The nonempty subsets of l are {I1, I2}, {I1, I5}, {I2, I5}, {I1}, {I2}, and {I5}. The resulting association rules are as shown below, each listed with its confidence:

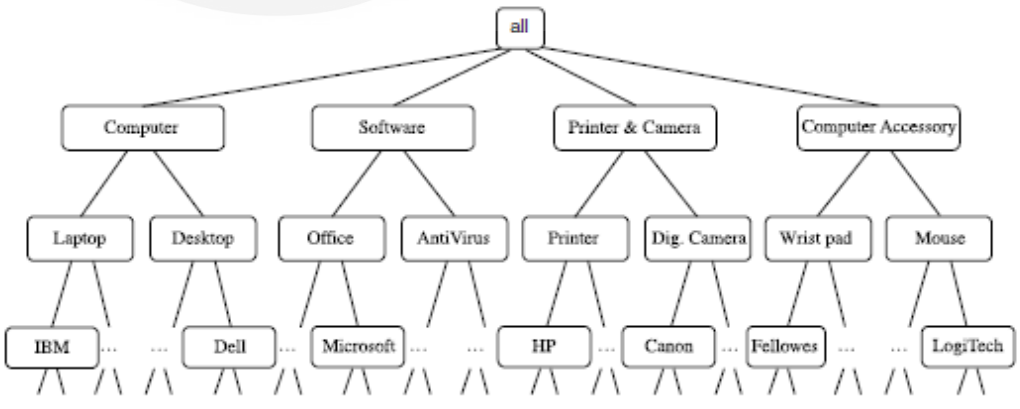
$$\begin{aligned}
&I1 \wedge I2 \Rightarrow I5, && confidence = 2/4 = 50\% \\
&I1 \wedge I5 \Rightarrow I2, && confidence = 2/2 = 100\% \\
&I2 \wedge I5 \Rightarrow I1, && confidence = 2/2 = 100\% \\
&I1 \Rightarrow I2 \wedge I5, && confidence = 2/6 = 33\% \\
&I2 \Rightarrow I1 \wedge I5, && confidence = 2/7 = 29\% \\
&I5 \Rightarrow I1 \wedge I2, && confidence = 2/2 = 100\%
\end{aligned}$$

9.7 MINING MULTILEVEL ASSOCIATION RULES

- For many applications, it is difficult to find strong associations among data items at lowor primitive levels of abstraction due to the sparsity of data at those levels.
- Strong associations discovered at high levels of abstraction may represent commonsenseknowledge.
- Therefore, data mining systems should provide capabilities for mining association rulesat multiple levels of abstraction, with sufficient flexibility for easy traversal among different abstraction spaces.
- Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules.
- Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework.
- In general, a top-down strategy is employed, where counts are accumulated for thecalculation of frequent itemsets at each concept level, starting at the concept level 1 and working downward in the hierarchy toward the more specific concept levels, until no more frequent itemsets can be found.

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher level, more general concepts. Data can be generalized by replacing low-level concepts within the data by their higher-level concepts, or ancestors, from a concept hierarchy.

TID	Items Purchased
T100	IBM-ThinkPad-T40/2373, HP-Photosmart-7660
T200	Microsoft-Office-Professional-2003, Microsoft-Plus!-Digital-Media
T300	Logitech-MX700-Cordless-Mouse, Fellowes-Wrist-Rest
T400	Dell-Dimension-XPS, Canon-PowerShot-S400
T500	IBM-ThinkPad-R40/P4M, Symantec-Norton-Antivirus-2003
...	...



A concept hierarchy for *AllElectronics* computer items.

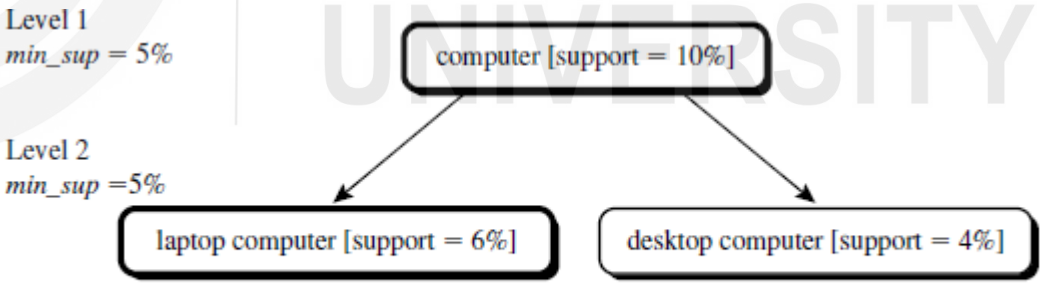
The concept hierarchy has five levels, respectively referred to as levels 0to 4, starting with level0 at the root node for all.

- Here, Level 1 includes computer, software, printer & camera, and computer accessory.
- Level 2 includes laptop computer, desktop computer, office software, antivirus software
- Level 3 includes IBM desktop computer. . . Microsoft office software and so on.
- Level 4 is the most specific abstraction level of this hierarchy.

9.8 APPROACHES FOR MINING MULTILEVEL ASSOCIATION RULES

1. Uniform Minimum Support:

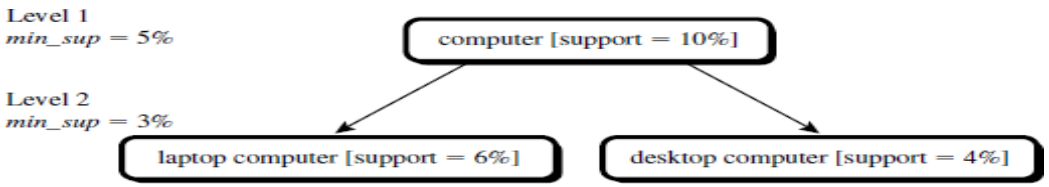
- The same minimum support threshold is used when mining at each level of abstraction.
- When a uniform minimum support threshold is used, the search procedure is simplified.
- The method is also simple in that users are required to specify only one minimum supportthreshold.
- The uniform support approach, however, has some difficulties. It is unlikely that items at lower levels of abstraction will occur as frequently as those at higher levels of abstraction.
- If the minimum support threshold is set too high, it could miss some meaningful associationsoccurring at low abstraction levels. If the threshold is set too low, it may generate many uninteresting associations occurring at high abstraction levels.



Multilevel mining with uniform support.

2. Reduced Minimum Support

- Each level of abstraction has its own minimum support threshold.
- The deeper the level of abstraction, the smaller the corresponding threshold is.
- For example, the minimum support thresholds for levels 1 and 2 are 5% and 3%, respectively.In this way, computer, laptop computer, and desktop computer‖ are all considered frequent.



3. Group-Based Minimum Support:

- Because users or experts often have insight as to which groups are more important than others, it is sometimes more desirable to set up user-specific, item, or group based minimal support thresholds when mining multilevel rules.
- For example, a user could set up the minimum support thresholds based on product price, or on items of interest, such as by setting particularly low support thresholds for laptop computers and flash drives in order to pay particular attention to the association patterns containing items in these categories.

9.9 MINING MULTIDIMENSIONAL ASSOCIATION RULES FROM RELATIONALDATABASES AND DATA WAREHOUSES

- Single dimensional or intra dimensional association rule contains a single distinct predicate (e.g., buys) with multiple occurrences i.e., the predicate occurs more than once within the rule.
  buys(*X, –digital camera‖*)=>*buys*(*X, –HP printer‖*)
- Association rules that involve two or more dimensions or predicates can be referred toas multidimensional association rules.
  age(X, "20…29")^occupation(X, "student")=>buys(X, "laptop")
- Above Rule contains three predicates (age, occupation, and buys), each of which occurs only once in the rule. Hence, we say that it has no repeated predicates.
- Multidimensional association rules with no repeated predicates are called inter dimensional association rules.
- We can also mine multidimensional association rules with repeated predicates, which contain multiple occurrences of some predicates. These rules are called hybrid-dimensional association rules. An example of such a rule is the following, where the predicate buys is repeated:
  age(X, –20…29‖)^buys(X, –laptop‖)=>buys(X, –HP printer‖)

9.10 MINING QUANTITATIVE ASSOCIATION RULES

- Quantitative association rules are multidimensional association rules in which the numeric attributes are *dynamically* discretized during the mining process so as to satisfy some mining criteria, such as maximizing the confidence or compactness of the rules mined.

- In this section, we focus specifically on how to mine quantitative association rules havingtwo quantitative attributes on the left-hand side of the rule and one categorical attribute onthe right-hand side of the rule. That is
*Aquan*1 ^*Aquan*2 =>*Acat*

  Where *Aquan*1 and *Aquan*2 are tests on quantitative attribute interval

  Acat tests a categorical attribute from the task-relevant data.

- Such rules have been referred to as two-dimensional quantitative association rules, because they contain two quantitative dimensions.
- For instance, suppose you are curious about the association relationship between pairs ofquantitative attributes, like customer age and income, and the type of television (such as *high-definition TV,* i.e., *HDTV*) that customers like to buy.
An example of such a 2-D quantitative association rule is

  age(*X*, ‒30…39‖)^*income*(*X*, ‒42*K*…48*K*‖)=>*buys*(*X*, ‒*HDTV*‖)

9.11 FROM ASSOCIATION MINING TO CORRELATION ANALYSIS

- A correlation measure can be used to augment the support-confidence framework forassociation rules. This leads to *correlation rules* of the form
A=>*B* [*support, confidence, correlation*]
- That is, a correlation rule is measured not only by its support and confidence but also by the correlation between itemsets *A* and *B*. There are many different correlation measures from which to choose. In this section, we study various correlation measures to determine which would be good for mining large data sets.
- Lift is a simple correlation measure that is given as follows. The occurrence of itemset*A* is independent of the occurrence of itemset *B* if $P(A \cup B)$ = *P*(*A*)*P*(*B*); otherwise, itemsets *A* and *B* are dependent and correlated as events. This definition can easily be extended to more than two itemsets.

The lift between the occurrence of *A* and *B* can be measured by computing

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}.$$

- If the lift(A,B) is less than 1, then the occurrence of A is negatively correlated with theoccurrence of B.
- If the resulting value is greater than 1, then A and B are positively correlated, meaning thatthe occurrence of one implies the occurrence of the other.
- If the resulting value is equal to 1, then A and B are independent and there is no correlationbetween them.