DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L 01

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : I -Data Mining & Data Preprocessing **Date of Lecture: 20.08.2021**

Topic of Lecture : Data Mining–Concepts

Introduction :

- Data mining is a process of extracting and discovering patterns in large data sets.
- Data mining (Knowledge Discovery from Data).
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Database
- Schema
- Meta data
- Data set

Detailed Content of the Lecture:

- Knowledge Discovery in Databases (KDD), knowledge extraction, data/pattern analysis, business intelligence, etc.

Advantages of Data Mining :

1. Increasing revenue.
2. Understanding customer segments and preferences.
3. Acquiring new customers.
4. Improving cross-selling and up-selling.
5. Detecting fraud.
6. Identifying credit risks.
7. Monitoring operational performance.

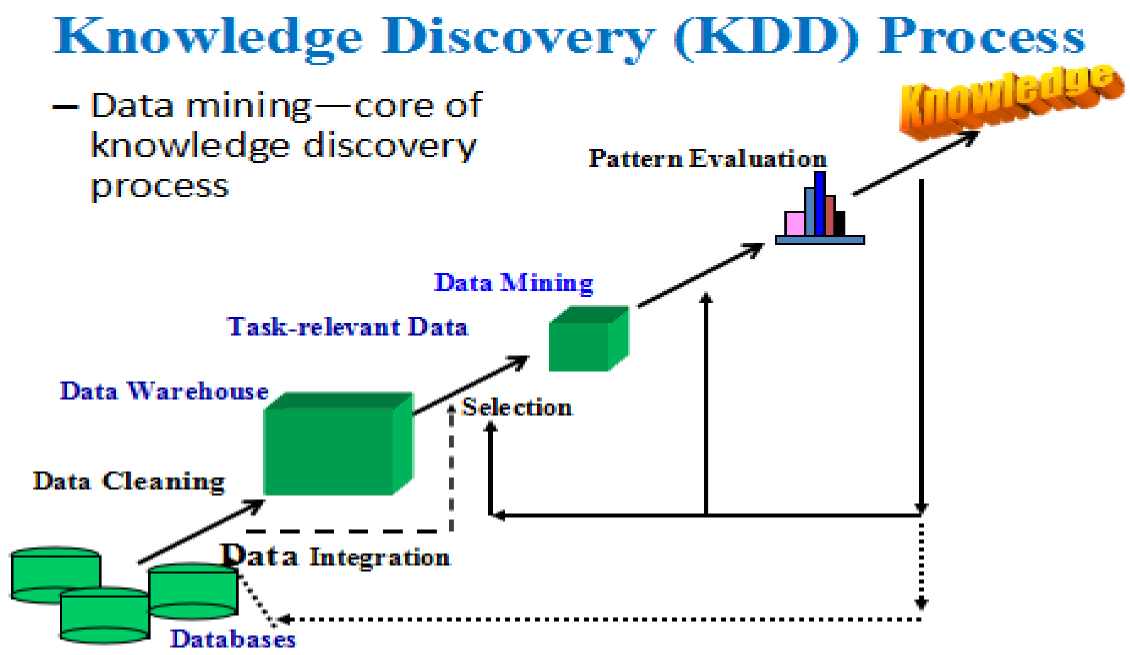
Knowledge Discovery (KDD) Process

KDD Process Steps:

1. Goal-Setting and Application Understanding.
2. Data Selection and Integration.
3. Data Cleaning and Pre processing.
4. Data Transformation.
5. Data Mining.
6. Pattern Evaluation/Interpretation.
7. Knowledge Discovery and Use.

KDD Process: Several Key Elements :

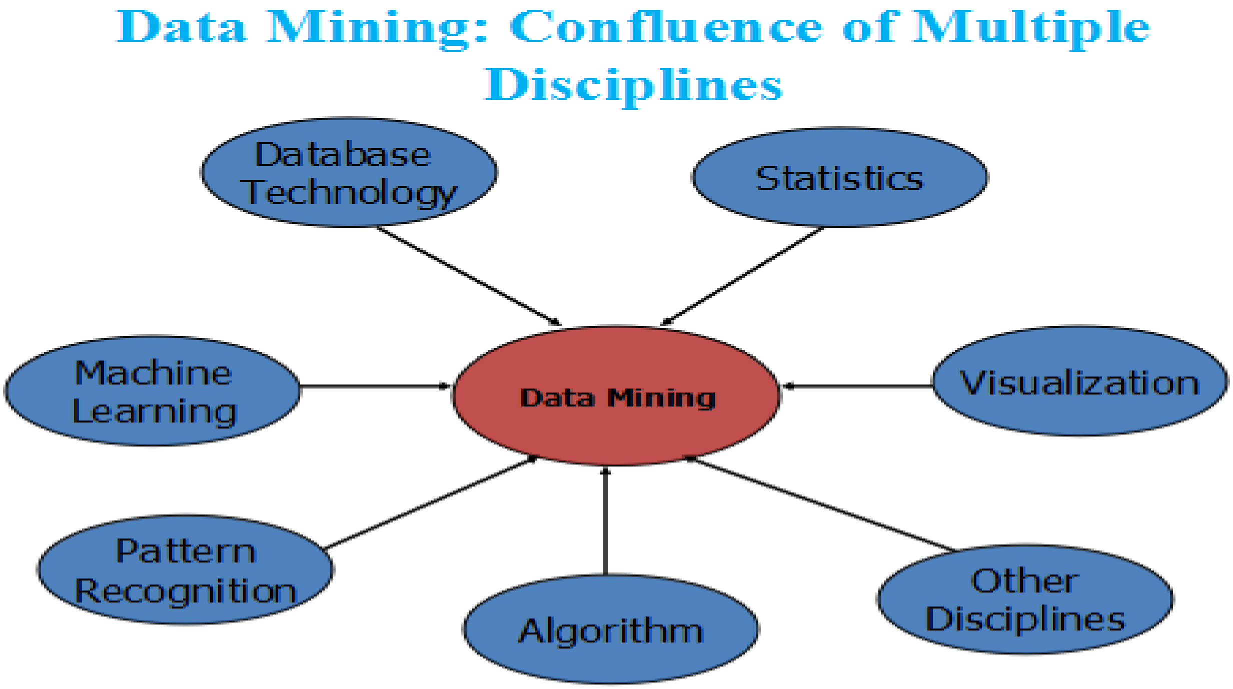
- Learning the application domain
- Creating a target data set
- Data cleaning and pre processing
- Data reduction and transformation



Knowledge Discovery (KDD) Process

– Data mining—core of knowledge discovery process

Data Mining: Confluence of Multiple Disciplines :



Data Mining: Confluence of Multiple Disciplines

Video Content / Details of website for further learning (if any):
https://docs.microsoft.com/en-us/analysis-services/data-mining/data-mining-concepts? view = asallproducts- allversions

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 1 - 8)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L 02

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : I -Data Mining & Data Preprocessing Date of Lecture: 21.08.2021

Topic of Lecture : DBMS versus Data mining , kinds of Data

Introduction :

- Data mining is a process of extracting and discovering patterns in large data sets.
- Data mining (Knowledge Discovery from Data).
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Database
- Schema
- Meta data
- Data set

Detailed Content of the Lecture:

DBMS	**Data Mining**
The database is the organized collection of data. data are stored in very large databases.	Data mining is analyzing data from different information to discover useful knowledge.
A Database may contain different levels of abstraction in its architecture.	Data mining deals with extracting useful and previously unknown information from raw data.
Typically, the three levels: external, conceptual and internal make up the database architecture.	The data mining process relies on the data compiled in the data warehousing phase in order to detect meaningful patterns.

Data Mining: On What Kinds of Data?

Database-Oriented Data Sets and Applications :

Relational database, data warehouse, transactional database

Advanced Data sets and Advanced Applications :

1. Time-series data, sequence data
2. Object-relational databases
3. Heterogeneous databases
4. Multimedia database
5. Text databases
6. The World-Wide Web

Kinds of Data in Data Mining :

1. Relational databases
2. Data warehouses
3. Transactional Databases
4. Advanced database systems
5. Object-relational
6. Spacial and Temporal □
7. Time-series □
8. Multimedia Data Mining
9. Text Mining□
10. Web Mining

□

- Creating a target data set
- Data cleaning and pre processing
- Data reduction and transformation
- Choosing functions of data mining
- summarization, classification, association, clustering
- Choosing the mining algorithm(s)
- Pattern evaluation and knowledge presentation
- Use of discovered knowledge

Video Content / Details of website for further learning (if any):
https://vspages.com/dbms-vs-data-mining-1878/
https://www.includehelp.com/basics/data-types-in-data-mining.aspx

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 9 - 12)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS **L 03**

MCA **II / III**

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : I -Data Mining & Data Preprocessing Date of Lecture: 24.08.2021

Topic of Lecture : Applications of Data Mining

Introduction :

- Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field.
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Database
- Schema
- Data Mining
- Data set

Detailed Content of the Lecture:

Data Mining Applications :

Here is the list of areas where data mining is widely used

- **Financial Data Analysis**
- **Retail Industry**
- **Telecommunication Industry**
- **Biological Data Analysis**
- **Other Scientific Applications**
- **Intrusion Detection**

1. Financial Data Analysis :

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining.
Design and construction of data warehouses for multidimensional data analysis and data mining.
Loan payment prediction and customer credit policy analysis.
Classification and clustering of customers for targeted marketing

2.Retail Industry :

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services.
Design and Construction of data warehouses based on the benefits of data mining.
Multidimensional analysis of sales, customers, products, time and region.
Customer Retention.
Product recommendation and cross-referencing of items.

3. Telecommunication Industry :

The telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc.
Multidimensional Analysis of Telecommunication data.
Identification of unusual patterns.
Multidimensional association and sequential patterns analysis.
Mobile Telecommunication services.

4. Biological Data Analysis :

Biological data mining is a very important part of Bio informatics.
Semantic integration of heterogeneous databases.
Alignment, indexing, similarity search and comparative analysis.
Discovery of structural patterns and analysis of genetic networks.
Association and path analysis.
Visualization tools in genetic data analysis.

5. Other Scientific Applications :

Following are the applications of data mining in the field of Scientific Applications −
Data Warehouses and Data Preprocessing.
Graph-based mining.
Visualization and domain specific knowledge.

6. Intrusion Detection :

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources.
Development of data mining algorithm for intrusion detection.
Association and correlation analysis, aggregation to help select and build discriminating attributes.
Analysis of Stream data.
Distributed data mining.
Visualization and query tools.

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. The major issues are −

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

Video Content / Details of website for further learning (if any):
https://bigdata-madesimple.com/14-useful-applications-of-data-mining/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 15 - 18)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 04

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : I -Data Mining & Data Preprocessing Date of Lecture: 25.08.2021

Topic of Lecture : Issues and Challenges in Data Mining

Introduction :

- 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.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Database Technology
- Schema
- Meta data
- Data Mining

Detailed Content of the Lecture:

The major issues are −

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

1.Mining Methodology and User Interaction :

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.
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.
Incorporation of background knowledge − 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.
Presentation and visualization of data mining results − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations.
Handling noisy or incomplete data − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities.
Pattern evaluation − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

2. Performance Issues :

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.

3. 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.

Challenges in Data Mining

Some of the Data Mining challenges are :

1. Security and Social Challenges
2. Noisy and Incomplete Data
3. Distributed Data
4. Complex Data
5. Performance
6. Scalability and Efficiency of the Algorithms
7. Improvement of Mining Algorithms
8. Incorporation of Background Knowledge
9. Data Visualization
10. Data Privacy and Security
11. User Interface
12. Mining dependent on Level of Abstraction
13. Integration of Background Knowledge
14. Mining Methodology Challenges

Video Content / Details of website for further learning (if any):

https://www.geeksforgeeks.org/challenges-of-data-mining/

Important Books/Journals for further learning including the page nos.:

Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 36 - 40)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 05

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : I -Data Mining & Data Preprocessing Date of Lecture: 27.08.2021

Topic of Lecture : Need for Data Pre-processing & Data Cleaning

Introduction :

- Data pre-processing is the process of transforming raw data into an understandable format. It is also an important step in data mining.
- Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Prerequisite knowledge for Complete understanding and learning of Topic:

- KDD Process
- Pattern Evaluation
- Meta data
- Data set

Detailed Content of the Lecture:

Incomplete: Lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.
Noisy: Containing Random errors or outliers.
Inconsistent: Containing discrepancies in codes or names.

Need for Data Pre - processing :
Yes Data Pre - Processing is need to check the data quality.

The quality can be checked by the following :
Accuracy: To check whether the data entered is correct or not.
Completeness: To check whether the data is available or not recorded.
Consistency: To check whether the same data is kept in all the places that do or do not match.
Timeliness: The data should be updated correctly.
Believability: The data should be trustable.
Interpretability: The understandability of the data.

1. Data Cleaning

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
Data cleaning tasks :

- Fill in missing values.
- Identify outliers and smooth out noisy data.
- Correct inconsistent data.

(a). Missing Data : Data is not always available

□ E.g., many tuples have no recorded value for several attributes, such as : customer income in sales data

Missing data may be due to :

- inconsistent with other recorded data and thus deleted.
- Data not entered due to misunderstanding.
- Certain data may not be considered important at the time of entry.
- Not register history or changes of the data.

(b). Noisy Data: Random error or variance in a measured variable.

Incorrect attribute values may due to

- Faulty data collection instruments
- Data entry problems
- Data transmission problems
- Technology limitation
- Inconsistency in naming convention

Other data problems which requires data cleaning

- Duplicate records
- Incomplete data
- Inconsistent data

Handle Noisy Data:

1. Binning method:

□ first sort data and partition into (equi-depth) bins, then smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

2. Regression : smooth by fitting the data into regression functions.

3. Clustering : This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters. Finally Detect and remove outliers.

1**. Binning methods for Data Smoothing:**

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into (equi-depth) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

* Smoothing by bin means:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

Video Content / Details of website for further learning (if any):

https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/

Important Books/Journals for further learning including the page nos.:

Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 47- 66)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS **L 06**

MCA **II / III**

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : I -Data Mining & Data Preprocessing Date of Lecture: 28.08.2021

Topic of Lecture : Data Integration

Introduction :

- Data Integration – Integration of multiple databases, data cubes, or files.
- It combines data from multiple sources into a coherent Data store.

Prerequisite knowledge for Complete understanding and learning of Topic:

- KDD Process
- Pattern Evaluation
- Meta data
- Data set

Detailed Content of the Lecture:

- Data Integration is a data pre processing technique that involves combining data from multiple heterogeneous data sources into a coherent data store and provide a unified view of the data.
- These sources may include multiple data cubes, databases, or flat files.

Issues in Data Integration:

1. Schema Integration and Object Matching.
2. Redundancy.
3. Detection and resolution of data value conflicts.

1. Schema Integration and Object Matching: Integrate metadata from different sources. The real-world entities from multiple sources are matched referred to as the entity identification problem.
Entity identification problem: identify real world entities from multiple data sources,
e.g., A.cust-id =B.cust-id

2. Redundancy :

- Redundant data occur often when integration of multiple databases.
- An attribute may be redundant if it can be derived or obtaining from another attribute or set of attributes.
- Inconsistencies in attributes can also cause redundancies in the resulting data set.
- Some redundancies can be detected by correlation analysis.
- The same attribute may have different names in different databases Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality.

3. Detection and resolution of data value conflicts :

- This is the third important issue in data integration.
- Attribute values from different sources may differ for the same real-world entity.
- An attribute in one system may be recorded at a lower level abstraction than the “same” attribute in another.

Linear regression: Y =a + b X□ Two parameters , a and b specify the line and are to be estimated by using the data at hand.

Multiple regression: Y = b0 + b1 X1 + b2 X2.
□ Many nonlinear functions can be transformed into the above.

Log-linear regression: The multi-way table of joint probabilities is approximated by a product of lower-order tables.
□ Probability: p(a, b, c, d) = ab ac ad □

Clustering :

- Partition data set into clusters, and one can store cluster representation only.
- Can be very effective if data is clustered but not if data is “smeared”.
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures.

Incomplete: Lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.
Noisy: Containing Random errors or outliers.
Inconsistent: Containing discrepancies in codes or names.

Video Content / Details of website for further learning (if any):
https://www.omnisci.com/technical-glossary/data-integration

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 67 - 70)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 07

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : I -Data Mining & Data Preprocessing **Date of Lecture: 31.08.2021**

Topic of Lecture : Data Transformation

Introduction :

- Transform the data in appropriate forms suitable for mining process.
- Transforming or consolidating data into mining suitable form is known as Data Transformation.

Prerequisite knowledge for Complete understanding and learning of Topic:

- KDD Process
- Data Cleaning
- Meta data
- Data Integration

Detailed Content of the Lecture:

Transforming or consolidating data into mining suitable form is known as Data Transformation.

This involves following ways:

Smoothing

Aggregation

Generalization

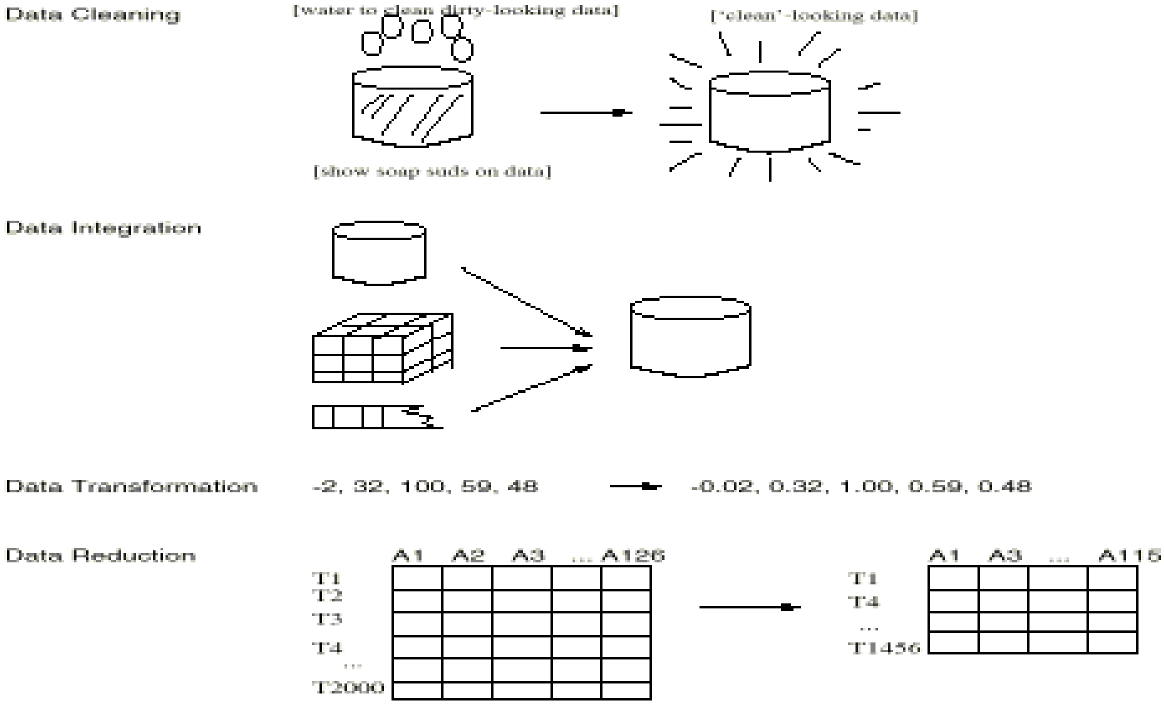
Normalization

Attribute construction

Smoothing: remove noise from data.
Aggregation: summarization, data cube construction.
Generalization: concept hierarchy climbing.
Normalization: It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0). scaled to fall within a small, specified range.

- Min-max normalization
- Z-score normalization
- Normalization by decimal scaling
- Attribute values from different sources may differ for the same real-world entity.
- An attribute in one system may be recorded at a lower level abstraction than the “same” attribute in another.

- **Attribute Selection:** In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
- **Discretization:** This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
- **Concept Hierarchy Generation:** Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.



Video Content / Details of website for further learning (if any):
https://www.stitchdata.com/resources/data-transformation/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 70 - 72)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS | **L 08**

MCA | **II / III**

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : I -Data Mining & Data Preprocessing Date of Lecture: 01.09.2021

Topic of Lecture : Data Reduction

Introduction :

- Data reduction techniques are applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of base data.”
- We uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

Prerequisite knowledge for Complete understanding and learning of Topic:

- KDD Process
- Data Cleaning
- Data Integration
- Data Transformation

Detailed Content of the Lecture:

Data Reduction - Strategies

- Data cube aggregation
- Dimensionality Reduction
- Data Compression
- Numerosity Reduction
- Data Discretization and Concept Hierarchy Generation

1. **Data cube aggregation :** Aggregation operation is applied to data for the construction of the data cube.

The lowest level of a data cube

□ The aggregated data for an individual entity of interest

□ e.g., a customer in a phone calling data warehouse.

Multiple levels of aggregation in data cubes

□ Further reduce the size of data to deal with it.

Reference appropriate levels

□ Use the smallest representation which is enough to solve the task.

2. Dimensionality Reduction :

This reduce the size of data by encoding mechanisms.

- It can be lossy or lossless.
- If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction.

The two effective methods of dimensionality reduction are:
1. Wavelet Transforms
2. PCA (Principal Component Analysis).

3. **Data Compression :** Data compression is the process of encoding, restructuring or otherwise modifying data in order to reduce its size.
It can be lossy or lossless.

- **Lossless Compression: Not Occuring Data Losses.**
- **Lossy Compression: Occurs the Data Losses.**

4. Numerosity Reduction :

This enable to store the model of data instead of whole data,

For example:

- **Regression Models.**
- **Linear regression**
- **Multiple regression**
- **Log-linear regression**

- An attribute may be redundant if it can be derived or obtaining from another attribute or set of attributes.
- Inconsistencies in attributes can also cause redundancies in the resulting data set.
- Some redundancies can be detected by correlation analysis.
- The same attribute may have different names in different databases Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality.

Video Content / Details of website for further learning (if any):
https://searchdatabackup.techtarget.com/definition/data-reduction

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 72 - 85)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 09

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : I -Data Mining & Data Preprocessing **Date of Lecture: 03.09.2021**

Topic of Lecture : Data Discretization and Concept Hierarchy Generation.

Introduction :

- This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
- Divide the range of a continuous attribute into intervals.
- Some classification algorithms only accept categorical attributes.
- Reduce data size by discretization.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Detailed Content of the Lecture:

- Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.
- Interval labels can then be used to replace actual data values.

Three types of attributes:

Nominal — values from an unordered set.
Ordinal — values from an ordered set.
Continuous — It have a Real Numbers.

- Divide the range of a continuous attribute into intervals.
- Some classification algorithms only accept categorical attributes.
- Reduce data size by discretization.

Concept Hierarchy Generation : Attributes are converted from lower level to higher level in hierarchy.

For Example- The attribute “city” can be converted to “country”.

Reduce the data by collecting and replacing low level concepts.
such as numeric values for the (attribute -age) by higher level concepts (such as young, middle-aged, or old).

- If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction.

Data cube aggregation : Aggregation operation is applied to data for the construction of the data cube. The lowest level of a data cube

The aggregated data for an individual entity of interest

 e.g., a customer in a phone calling data warehouse.

Multiple levels of aggregation in data cubes

 Further reduce the size of data to deal with it.

Reference appropriate levels

 Use the smallest representation which is enough to solve the task.

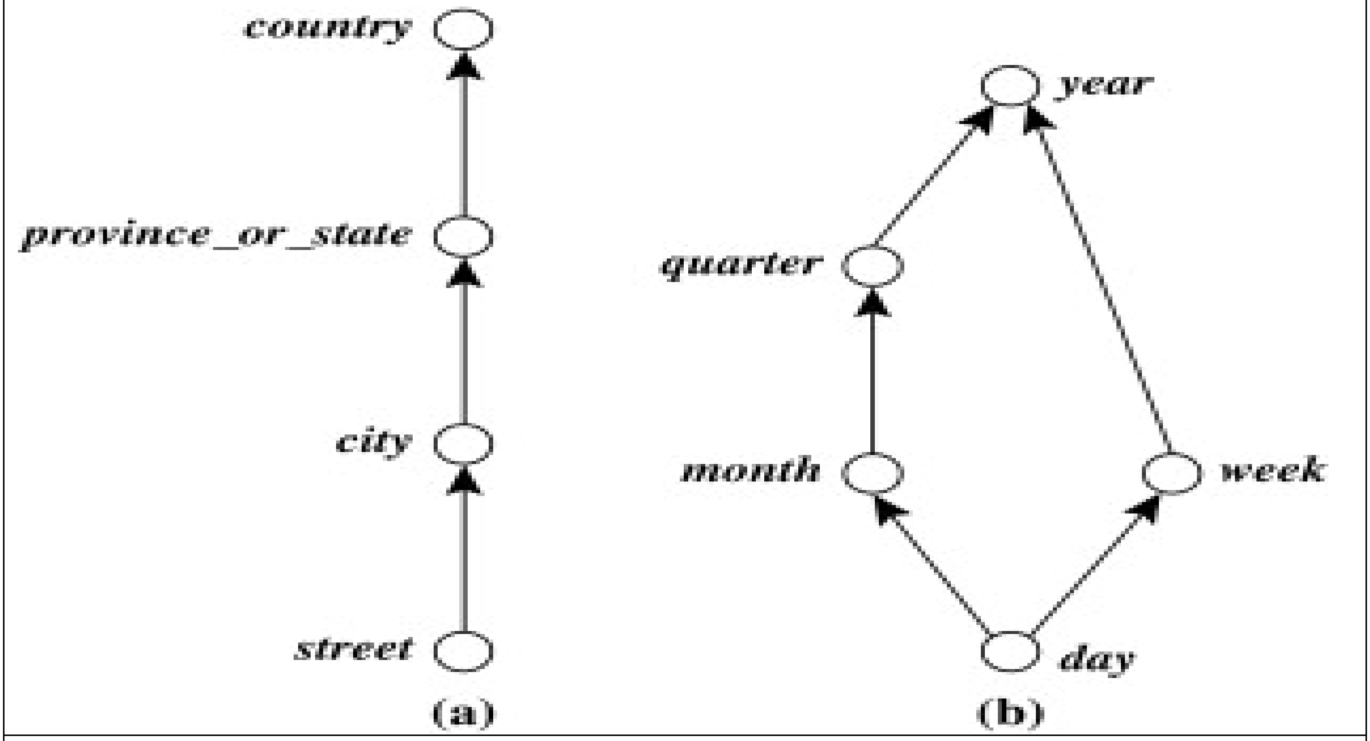
Data Compression : Data compression is the process of encoding, restructuring or otherwise modifying data in order to reduce its size.

It can be lossy or lossless.

Lossless Compression: Not Occuring Data Losses.

Lossy Compression: Occurs the Data Losses.

Example for Concept Hierarchy Generation :



Video Content / Details of website for further learning (if any):

http://www.lastnightstudy.com/Show?id=45/Data-Discretization-and-Concept-Hierarchy-Generation

Important Books/Journals for further learning including the page nos.:

Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 86 - 97)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 10

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : II - Association Rule Mining And Classification Basics

Date of Lecture: 04.09.2021

Topic of Lecture : Introduction - Association Rule Mining

Introduction :

Association rules are "if-then" statements, that help to show the probability of relationships between data items, within large data sets in various types of databases. Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.

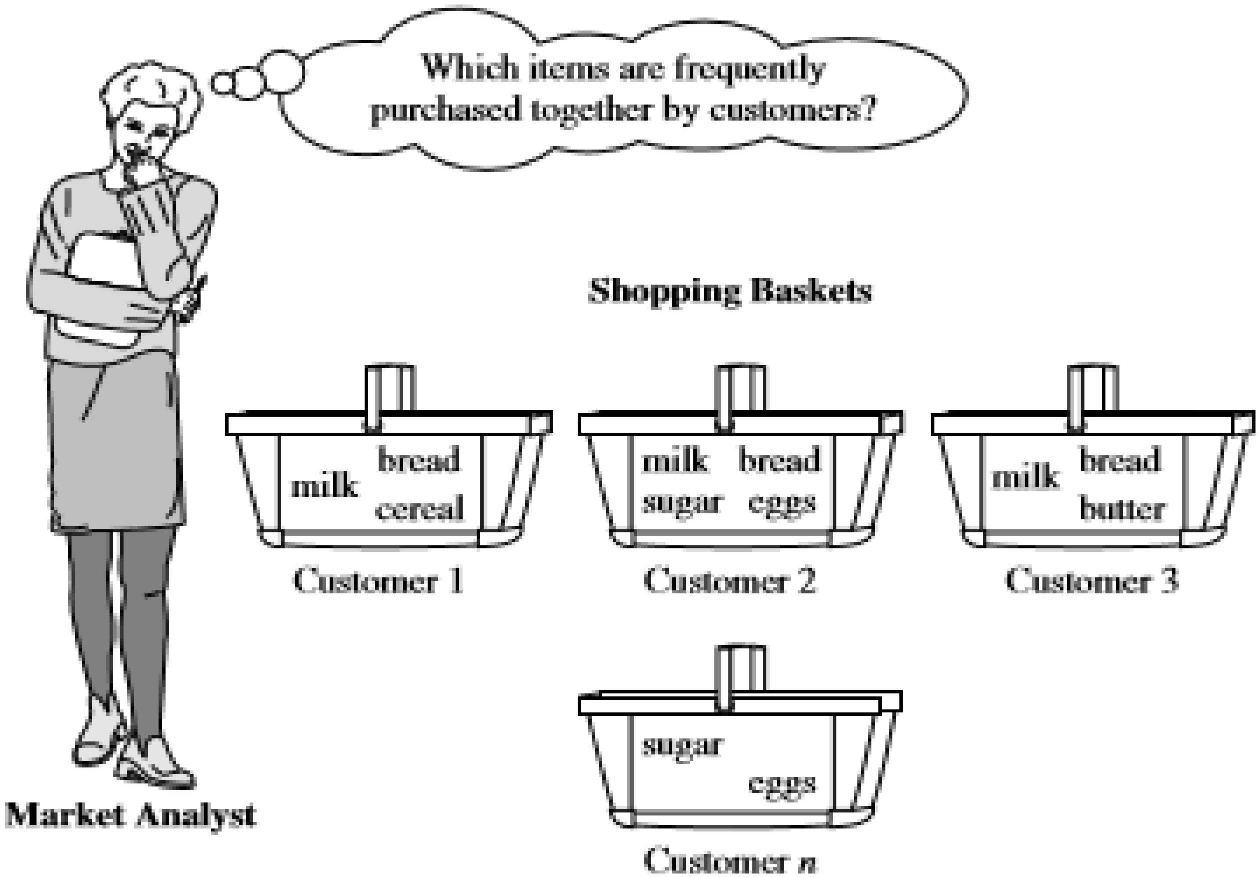
Prerequisite knowledge for Complete understanding and learning of Topic:

- Associations
- Correlations
- Clustering
- Data Compression

Detailed Content of the Lecture:

- Frequent itemset mining leads to the discovery of associations and correlations among item in large transactional or relational datasets.
- With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases.
- The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes such as catalog design, cross-marketing, and customer shopping behavior analysis.
- A typical example of frequent itemset mining is 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 these associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers.
- 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. If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item.
- Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together.

- These patterns can be represented in the form of association rules



- The rank correlation coefficients are used just to establish the type of the relationship, but not to investigate in detail like the Pearson's correlation coefficient.
- They are also used to reduce the calculations and make the results more independent of the non-normality of the distributions considered.
- Association is a concept, but correlation is a measure of association and mathematical tools are provided to measure the magnitude of the correlation.
- Pearson's product moment correlation coefficient establishes the presence of a linear relationship and determines the nature of the relationship (whether they are proportional or inversely proportional)

Video Content / Details of website for further learning (if any):
https://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 228 - 233)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 11

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : II - Association Rule Mining And Classification Basics

Date of Lecture: 07.09.2021

Topic of Lecture : Mining Frequent Itemsets with Candidate Generation

Introduction :

Frequent patterns are patterns (e.g., itemsets, subsequences, or substructures) that appear frequently in a data set.

For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset. frequently in a shopping history database, is a (frequent) pattern.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Associations
- Correlations
- Clustering
- Candidate Generation

Detailed Content of the Lecture:

Mining frequent itemsets with candidate generation

1.Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min sup.

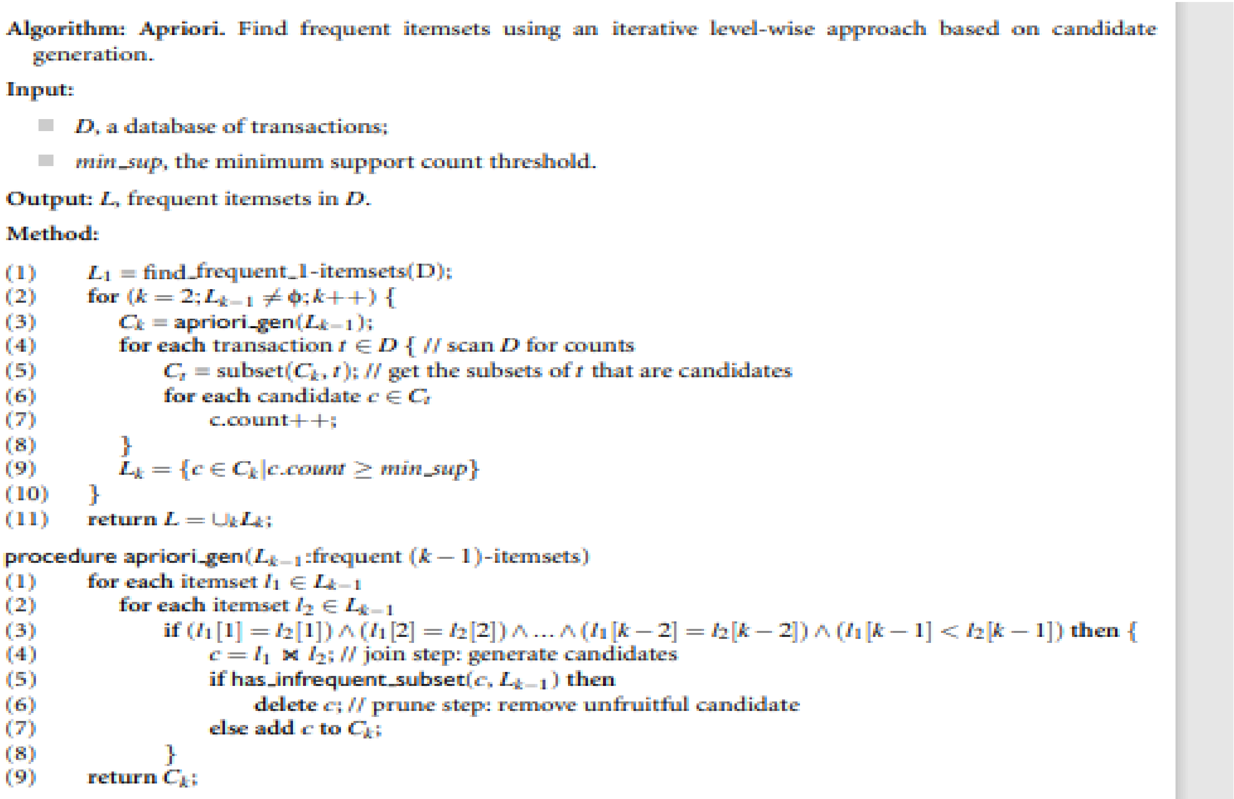
2. Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate Generation

- The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties, as we shall see later. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets.
- Aprioriproperty: All nonempty subsets of a frequent itemset must also be frequent. TheAprioripropertyisbasedonthefollowingobservation.Bydefinition,ifanitemset I does not satisfy the minimum support threshold, min sup, then I is not frequent, that is, P(I) < min sup.
- If an item A is added to the itemset I, then the resulting itemset (i.e., $I \cup A$) cannot occur more frequently thanI.
- Therefore, $I \cup A$ is not frequent either, that is, $P(I \cup A)$ < min sup. This property belongs to a special category of properties called antimonotonicity in the sense that if a set cannot pass a test, all of its supersets will fail the same test as well.
- It is called antimonotonicity because the property is monotonic in the context of failing a test.

A two-step process is followed, consisting of

1. **join and**
2. **prune actions.**



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

Input:

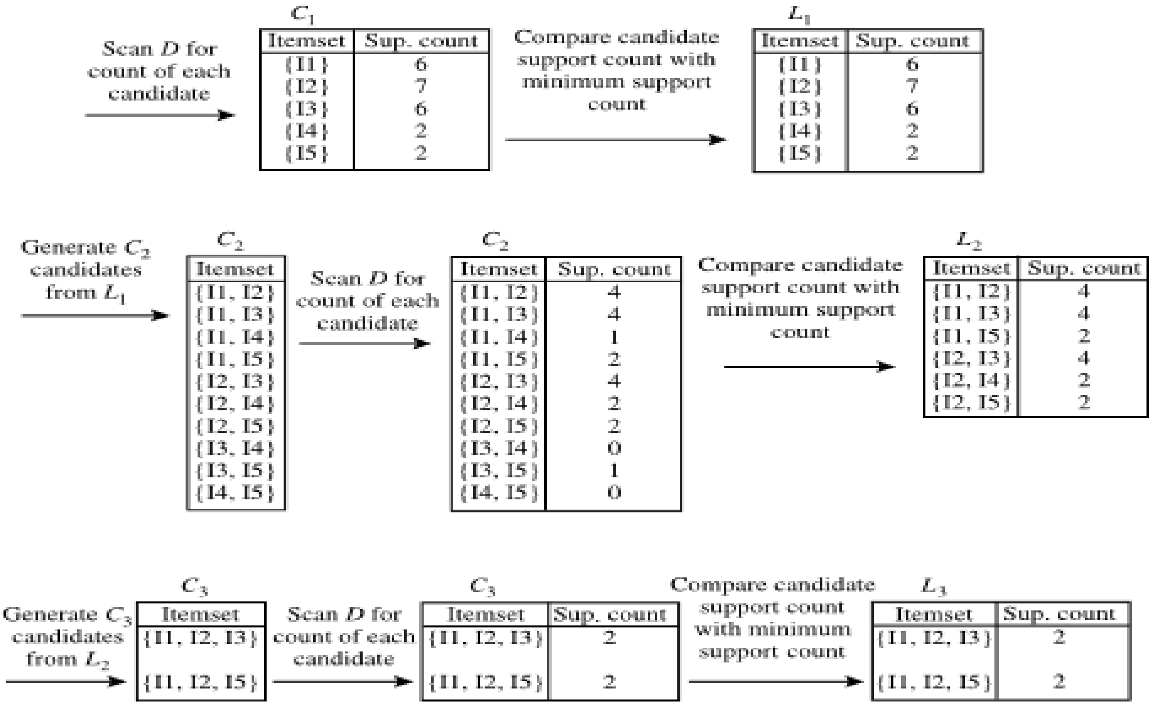
- D, a database of transactions;
- min_sup, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

```
$L_1$ = find_frequent_1-itemsets(D);
for ($k = 2; L_{k-1} \neq \phi; k{+}{+}$) {
    $C_k$ = apriori_gen($L_{k-1}$);
    for each transaction $t \in D$ { // scan $D$ for counts
        $C_t$ = subset($C_k, t$); // get the subsets of $t$ that are candidates
        for each candidate $c \in C_t$
            c.count++;
    }
    $L_k = \{c \in C_k | c.count \geq min\_sup\}$
}
return $L = \cup_k L_k$;

procedure apriori_gen($L_{k-1}$:frequent $(k-1)$-itemsets)
    for each itemset $l_1 \in L_{k-1}$
        for each itemset $l_2 \in L_{k-1}$
            if $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \ldots \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$ then {
                $c = l_1 \bowtie l_2$; // join step: generate candidates
                if has_infrequent_subset($c, L_{k-1}$) then
                    delete $c$; // prune step: remove unfruitful candidate
                else add $c$ to $C_k$;
            }
    return $C_k$;
```


Scan D for count of each candidate
Compare candidate support count with minimum support count
Generate C_2 candidates from L_1
Scan D for count of each candidate
Compare candidate support count with minimum support count
Generate C_3 candidates from L_2
Scan D for count of each candidate
Compare candidate support count with minimum support count

C_1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

L_1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

C_2

Itemset
{I1, I2}
{I1, I3}
{I1, I4}
{I1, I5}
{I2, I3}
{I2, I4}
{I2, I5}
{I3, I4}
{I3, I5}
{I4, I5}

C_2

Itemset	Sup. count
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

L_2

Itemset	Sup. count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

C_3

Itemset
{I1, I2, I3}
{I1, I2, I5}

C_3

Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

L_3

Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

Video Content / Details of website for further learning (if any):
https://dwgeek.com/mining-frequent-itemsets-apriori-algorithm.html/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 234 - 239)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 12

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : II - Association Rule Mining And Classification Basics

Date of Lecture: 08.09.2021

Topic of Lecture : Mining Frequent Itemsets without Candidate Generation

Introduction :

Frequent patterns are patterns (e.g., itemsets, subsequences, or substructures) that appear frequently in a data set.

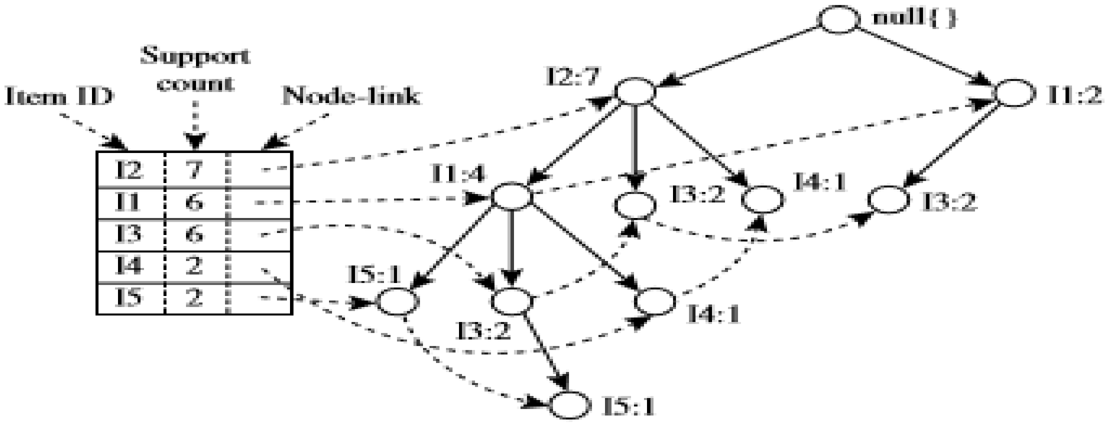
For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset. frequently in a shopping history database, is a (frequent) pattern.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Associations
- Correlations
- Clustering
- Candidate Generation

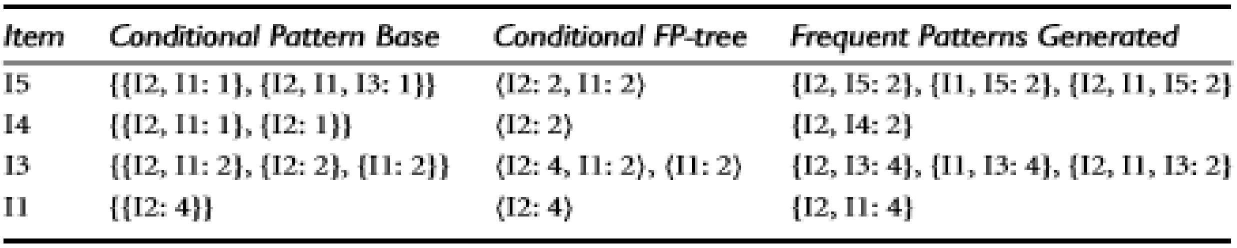
Detailed Content of the Lecture:

A Pattern-Growth Approach for Mining Frequent Itemsets

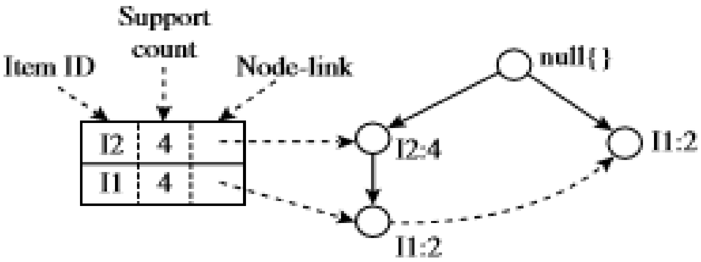


Item ID	Support count	Node-link
I2	7	
I1	6	
I3	6	
I4	2	
I5	2	

- To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links.
- The FP-tree is mined as follows. Start from each frequent length-1 pattern (as an initial suffix pattern), construct its conditional pattern base (a “sub-database,” which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern), then constructits (conditional) FP-tree,and performmining recursively on the tree.
- The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree.



Item	Conditional Pattern Base	Conditional FP-tree	Frequent Patterns Generated
I5	{{I2, I1: 1}, {I2, I1, I3: 1}}	⟨I2: 2, I1: 2⟩	{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}
I4	{{I2, I1: 1}, {I2: 1}}	⟨I2: 2⟩	{I2, I4: 2}
I3	{{I2, I1: 2}, {I2: 2}, {I1: 2}}	⟨I2: 4, I1: 2⟩, ⟨I1: 2⟩	{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}
I1	{{I2: 4}}	⟨I2: 4⟩	{I2, I1: 4}



A Pattern-Growth Approach for Mining Frequent Itemsets

- To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links.
- The FP-tree is mined as follows. Start from each frequent length-1 pattern (as an initial suffix pattern), construct its conditional pattern base (a “sub-database,” which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern), then constructits (conditional) FP-tree,and performmining recursively on the tree.
- The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree.

1.Strong Rules Are Not Necessarily Interesting

- Whether or not a rule is interesting can be assessed either subjectively or objectively.
- Ultimately,only the user can judge if a given rule is interesting,and this judgment,being subjective,may differ from one user to another.
- However,objective interestingness measures, based on the statistics “behind” the data, can be used as one step toward the goal of weeding out uninteresting rules that would otherwise be presented to the user.

2. From Association Analysis to Correlation Analysis

- As we have seen so far, the support and confidence measures are insufficient at filtering out uninteresting association rules.
- To tackle this weakness, a correlation measure can be used to augment the support–confidence framework for association rules.

Video Content / Details of website for further learning (if any):
https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 242 - 248)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS | **L 13**

MCA | **II / III**

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : II - Association Rule Mining And Classification Basics

Date of Lecture: 14.09.2021

Topic of Lecture : Mining Various Kinds of Association Rules

Introduction :

Frequent patterns are patterns (e.g., itemsets, subsequences, or substructures) that appear frequently in a data set.

Finding frequent patterns plays an essential role in mining associations,correlations,and many other interesting relationships among data.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Associations
- Correlations
- Item Sets
- Candidate Generation

Detailed Content of the Lecture:

Basic patterns:

A frequent pattern may have several alternative forms, including a simple frequent pattern, a closed pattern,or a max-pattern.

To review, a frequent pattern is a pattern (or itemset) that satisfies a minimum support threshold.

A pattern p is a closed pattern if there is no super pattern p0 with the same support as p.

Pattern p is a max-pattern if there exists no frequent super pattern of p.

Frequent patterns can also be mapped into association rules, or other kinds of rules based on interestingness measures.

Sometimes we may also be interested in infrequent or rarepat terns (i.e., patterns that occur rarely but are of critical importance, or negative patterns (i.e., patterns that reveal a negative correlation between items).

Based on the abstraction levels involved in a pattern:

Patterns or association rules may have items or concepts residing at high, low, or multiple abstraction levels.

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

buys(X, “computer”)⇒buys(X, “printer”)
buys(X, “laptop computer”)⇒buys(X, “color laser printer”)

The items bought are referenced at different abstraction levels (e.g., “computer” is a higher-level abstraction of “laptop computer,” and “color laser printer” is a lower-level abstraction of “printer”). We refer to the rule set mined as consisting of multilevel association rules. If, instead, the rules within a given set do not reference items or attributes at different abstraction levels, then the set contains single-level association rules.

Based on the number of dimensions involved in the rule or pattern: If the items or attributes in an association rule or pattern reference only one dimension, it is a single-dimensional association rule/pattern.

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 the calculation of frequent item sets 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.

Using uniform minimum support for all levels (referred to as uniform support): The same minimum support threshold is used when mining at each level of abstraction. For example, in Figure 5.11, a minimum support threshold of 5% is used throughout (e.g., for mining from “computer” down to “laptop computer”). Both “computer” and “laptop computer” are found to be frequent, while “desktop computer” is not.

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 support threshold. An Apriori-like optimization technique can be adopted, based on the knowledge that an ancestor is a superset of its descendants: The search avoids examining itemsets containing any item whose ancestors do not have

Video Content / Details of website for further learning (if any):
https://www.brainkart.com/article/Mining-Various-Kinds-of-Association-Rules_8317/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 248 - 250)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

LECTURE HANDOUTS **L 14**

MCA **II / III**

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : II - Association Rule Mining And Classification Basics

Date of Lecture: 15.09.2021

Topic of Lecture : Mining Multilevel Association Rules

Introduction :

Mining multilevel association rules Suppose we are given the task-relevant set of transactional data in Table for sales in an All Electronics store, showing the items purchased for each transaction.

A frequent item set typically refers to a set of items that often appear together in a transactional data set.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Associations
- Correlations
- Item Sets
- Candidate Generation

Detailed Content of the Lecture:

Mining multiple level association rules from transactional databases.

Items often form hierarchy. Items at the lower level are expected to have lower support.

• Rules regarding itemsets at appropriate levels could be quite useful.

• Transaction database can be encoded based on dimensions and levels

• We can explore shared multi-level mining

• Figure shows the Mining multiple level association rules from transactional databases.

Mining Multi-Level Associations

• A top_down, progressive deepening approach:

– First find high-level strong rules: milk ® bread [20%, 60%].

– Then find their lower-level "weaker" rules: 2% milk and wheat bread [6%, 50%].

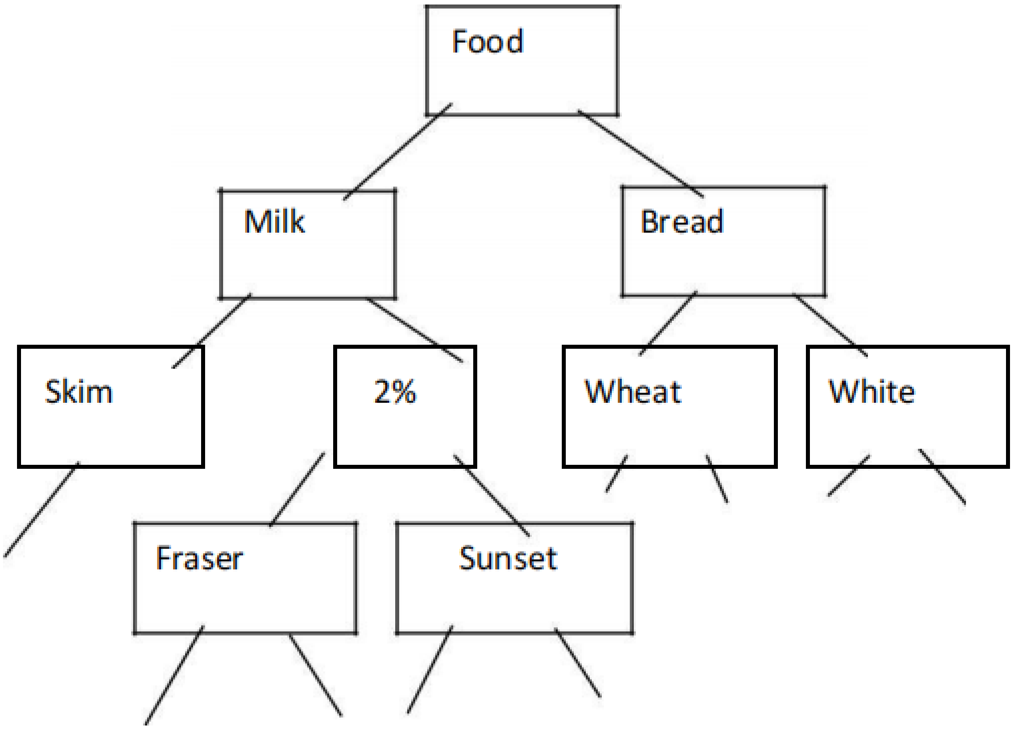
• Variations at mining multiple-level association rules.

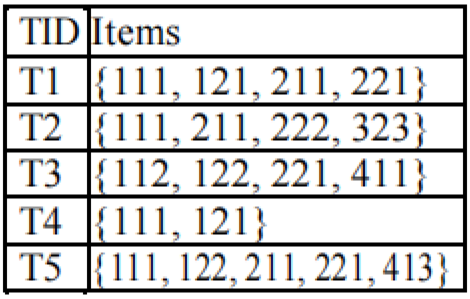
– Level-crossed association rules: 2% milk ® Wonder wheat bread

– Association rules with multiple, alternative hierarchies: 2% milk and Wonder bread

Using reduced minimum support at lower levels (referred to as reduced support):
Each abstraction level has its own minimum support threshold. The deeper the abstraction level,the smaller the corresponding threshold.

Using item or group-based minimum support (referred to as group-based support):
Because users or experts often have insight as to which groups are more important than others,it is sometimes more desirable to setup user-specific,item,or group-based minimal support thresholds when mining multilevel rules.





TID	Items
T1	{111, 121, 211, 221}
T2	{111, 211, 222, 323}
T3	{112, 122, 221, 411}
T4	{111, 121}
T5	{111, 122, 211, 221, 413}

Video Content / Details of website for further learning (if any):
https://www.geeksforgeeks.org/multilevel-association-rule-in-data-mining/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 250 - 253)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 15

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : II - Association Rule Mining And Classification Basics

Date of Lecture: 17.09.2021

Topic of Lecture : Mining Multidimensional Association Rules

Introduction :

Mining Multi-dimensional association rules Suppose we are given the task-relevant set of transactional data in Table for sales in an All Electronics store, showing the items purchased for each transaction.
A frequent item set typically refers to a set of items that often appear together in a transactional data set.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Association
- Frequent itemset
- Correlations
- Candidate Generation

Detailed Content of the Lecture:

- We have studied association rules that imply a single predicate, that is, the predicate buys.
- For instance, in mining our All Electronics database, we may discover the Boolean association rule.

buys(X, “digital camera”)⇒buys(X, “HP printer”)

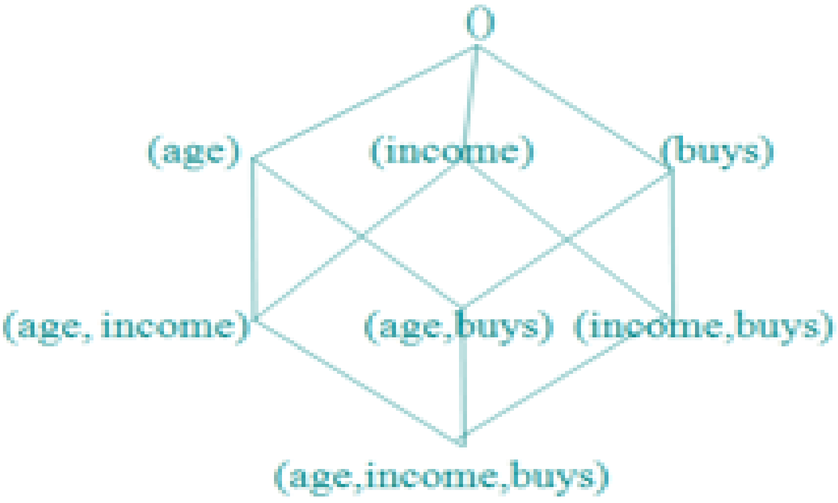
- A single dimensional or intra dimensional association rule because it contains a single distinct predicate (e.g., buys) with multiple occurrences (i.e., the predicate occurs more than once within the rule).
- Such rules are commonly mined from transactional data.
- Instead of considering transactional data only,sales and related information are often linked with relational data or integrated into a data warehouse. Such data stores are multidimensional in nature.
- Additional relational information regarding the customers who purchased the items (e.g., customer age, occupation, credit rating, income, and address) may also be stored. Considering each database attribute or warehouse dimension as a predicate, we can therefore mine association rules containing multiple predicates such as

age(X, “20...29”)∧occupation(X, “student”)⇒buys(X, “laptop”)

- Association rules that involve two or more dimensions or predicates can be referred to as multidimensional association rules.
- 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”)
- In the first approach, quantitative attributes are discretized using predefined concept hierarchies. This discretization occurs before mining.
- For instance, a concept hierarchy for income may be used to replace the original numeric values of this attribute by interval labels such as “0..20K,” “21K..30K,” “31K..40K,” and so on. Here, discretization is static and predetermined. Chapter 3 on data preprocessing gave several techniques for discretizing numeric attributes.
- The discretized numeric attributes, with their interval labels, can then be treated as nominal attributes (where each interval is considered a category). We refer to this as mining multidimensional association rules using static discretization of quantitative attributes.
- In the second approach, quantitative attributes are discretized or clustered into “bins” based on the data distribution.
- These bins may be further combined during the mining process.The discretization process is dynamic and established so as to satisfy somemining criteria such as maximizing the confidence of the rules mined.

Because this strategy treats the numeric attribute values as quantities rather than as pre defined ranges or categories, association rules mined from this approach are also referred to as (dynamic) quantitative associationrules.



Video Content / Details of website for further learning (if any):
https://www.geeksforgeeks.org/data-mining-multidimensional-association-rule/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 254 - 257)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L 16

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : II - Association Rule Mining And Classification Basics

Date of Lecture: 18.09.2021

Topic of Lecture : Constraint Based Association Mining

Introduction :

A data mining process may uncover thousands of rules from a given data set, most of which end up being unrelated or uninteresting to users.Often,users have a good sense of which “direction” of mining may lead to interesting patterns and the “form” of the patterns or rules they want to find.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Association
- Frequent itemset
- Correlations
- Candidate Generation

Detailed Content of the Lecture:

The constraints can include the following:

Knowledge type constraints: These specify the type of knowledge to be mined, such as association, correlation, classification, or clustering.

Dataconstraints: These specify the set of task-relevant data. Dimension/level constraints: These specify the desired dimensions (or attributes) of the data, the abstraction levels, or the level of the concept hierarchies to be used in mining.

Interestingness constraints: These specify thresholds on statistical measures of rule interestingness such as support, confidence, and correlation.

Ruleconstraints: These specify the form of, or conditions on, the rules to be mined. Such constraints may be expressed as meta rules (rule templates),as the maximum or minimumnumberofpredicatesthatcanoccurintheruleantecedentorconsequent, or as relationships among attributes, attribute values, and/or aggregates. These constraints can be specified using a high-level declarative data mining query language and user interface.

Meta rule-guided mining.

- Suppose that as a market analyst for AllElectronics you have access to the data describing customers (e.g., customer age, address, and credit rating) as well as the list of customer transactions.

$$P1(X, Y) \wedge P2(X, W) \Rightarrow buys(X, \text{“office software”})$$

Where P1 andP2 arepredicatevariablesthatareinstantiatedtoattributesfromthegiven database during the mining process, X is a variable representing a customer, and Y and W take on values of the attributes assigned to P1 and P2, respectively. Typically, a user will specify a list of attributes to be considered for instantiation with P1 and P2. Otherwise, a default set may be used.

$$age(X, \text{“30..39”}) \wedge income(X, \text{“41K..60K”}) \Rightarrow buys(X, \text{“office software”})$$

$$P1 \wedge P2 \wedge \cdots \wedge Pl \Rightarrow Q1 \wedge Q2 \wedge \cdots \wedge Qr$$

where Pi (i=1,..., l) and Qj (j=1,..., r) are either instantiated predicates or predicate variables. Let the number of predicates in the metarule be p=l+r.

- To find interdimensional association rules satisfying the template,We need to find all frequent p-predicate sets, Lp.
- We must also have the support or count of the l-predicate subsets of Lp to compute the confidence of rules derived from Lp.

This is a typical case of mining multidimensional association rules. By extending such methods using the constraint-pushing techniques described in the following section,we can derive efficient methods for metarule-guided mining.

Constraint-Based Pattern Generation: Pruning Pattern Space and Pruning Data Space

- Rule constraints specify expected set/subset relationships of the variables in the mined rules, constant initiation of variables, and constraints on aggregate functions and other forms of constraints. Users typically employ their knowledge of the application or data to specify rule constraints for the mining task.
- These rule constraints may be used together with, or as an alternative to, metarule-guided mining. In this section, we examine rule constraints as to how they can be used to make the mining process more efficient.
- Associative classification, where association rules are generated from frequent patterns and used for classification.
- The general idea is that we can search for strong associations between frequent patterns (conjunctions of attribute–value pairs) and class labels. The next is Discriminative frequent pattern–based classification, where frequent patterns serve as combined features, which are considered in addition to single features when building a classification model.

Video Content / Details of website for further learning (if any):
https://www.brainkart.com/article/Constraint-Based-Association-Mining_8319/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 265 - 270)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 17

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : II - Association Rule Mining And Classification Basics

Date of Lecture: 21.09.2021

Topic of Lecture : Classification versus Prediction

Introduction :

Classification is a **data mining function that assigns items in a collection to target categories or classes**. The goal of classification is to accurately predict the target class for each case in the data. It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Clustering
- Association
- Correlations
- Candidate Generation

Detailed Content of the Lecture:

Classification is the problem of identifying to which of a set of categories (sub populations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

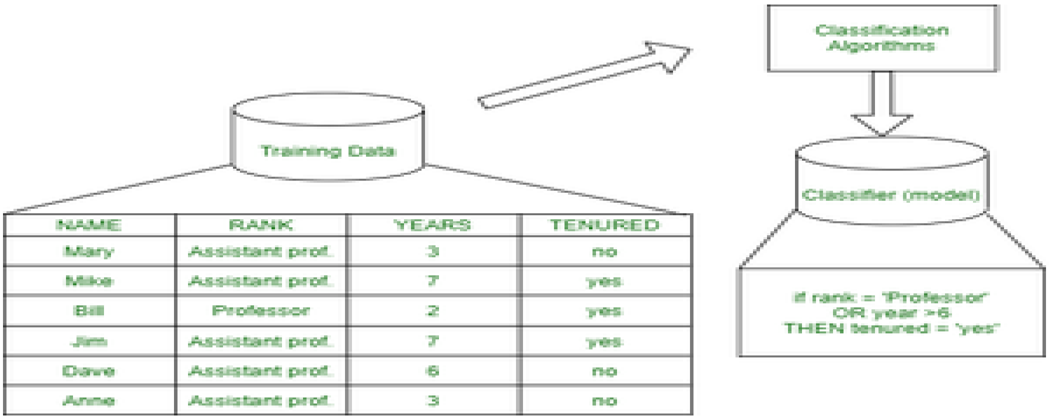
Example: Before starting any project, we need to check its feasibility. In this case, a classifier is required to predict class labels such as ‘Safe’ and ‘Risky’

A bank loans officer needs analysis of her data in order to learn which loan applicants are ―safe‖ and which are ―risky‖ for the bank. A marketing manager at All Electronics needs data analysis to help guess whether a customer with a given profile will buy a new computer.

A medical researcher wants to analyze breast cancer data in order to predict which one of three specific treatments a patient should receive. In each of these examples, the data analysis task is classification, where a model or classifier is constructed to predict categorical labels, such as ―safe‖ or ―risky‖ for the loan application data; ―yes‖ or ―no‖ for the marketing data; or ―treatment A,‖ ―treatment B,‖ or ―treatment C‖ for the medical data.

These categories can be represented by discrete values, where the ordering among values has no meaning. For example, the values 1, 2, and 3 may be used to represent treatments A,B, and C, where there is no ordering implied among this group of treatment regimes.

Learning Step (Training Phase): Construction of Classification Model Different Algorithms are used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.


Classification Algorithms
Training Data
Classifier (model)
if rank = 'Professor' OR year > 6 THEN tenured = 'yes'

NAME	RANK	YEARS	TENURED
Mary	Assistant prof.	3	no
Mike	Assistant prof.	7	yes
Bill	Professor	2	yes
Jim	Assistant prof.	7	yes
Dave	Assistant prof.	6	no
Anne	Assistant prof.	3	no

Issues Regarding Classification and Prediction :

Data Cleaning: This refers to the preprocessing of data in order to remove or reduce noise (by applying smoothing techniques, for example) and the treatment of missing values (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics). Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help reduce confusion during learning.

Relevance analysis: Many of the attributes in the data may be redundant. Correlation analysis can be used to identify whether any two given attributes are statistically related. For example, a strong correlation between attributes A1 and A2 would suggest that one of the two could be removed from further analysis. A database may also contain irrelevant attributes.
Attribute subset selection4 can be used in these cases to find a reduced set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.
It can be used to detect attributes that do not contribute to the classification or prediction task. Including such attributes may otherwise slow down, and possibly mislead, the learning step. Ideally, the time spent on relevance analysis, when added to the time spent on learning from the resulting ―reduced‖ attribute (or feature) subset, should be less than the time that would have been spent on learning from the original set of attributes.

Data transformation and reduction: The data may be transformed by normalization, particularly when neural networks or methods involving distance measurements are used in the learning step. Normalization involves scaling all values for a given attribute so that they fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0. In methods that use distance measurements, for example, this would prevent attributes with initially large ranges (like, say, income) from out weighing attributes with initially smaller ranges (such as binary attributes).

Video Content / Details of website for further learning (if any):
https://www.jigsawacademy.com/blogs/data-science/classification-and-prediction-in-data-mining/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 285 - 288)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 18

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : II - Association Rule Mining And Classification Basics

Date of Lecture: 22.09.2021

Topic of Lecture : Data Preparation for Classification and Prediction

Introduction :

Classification is a **data mining function that assigns items in a collection to target categories or classes**. The goal of classification is to accurately predict the target class for each case in the data. It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Clustering
- Association
- Correlations
- Candidate Generation

Detailed Content of the Lecture:

Binary: Possesses only two values i.e. True or False
Example: Suppose there is a survey evaluating some products. We need to check whether it's useful or not. So, the Customer has to answer it in Yes or No.

Product usefulness: Yes / No
Symmetric: Both values are equally important in all aspects
Asymmetric: When both the values may not be important.

Nominal: When more than two outcomes are possible. It is in Alphabet form rather than being in Integer form.

Example: One needs to choose some material but of different colors. So, the color might be Yellow, Green, Black, Red.

Different Colors: Red, Green, Black, Yellow

- **Ordinal**: Values that must have some meaningful order.
  Example: Suppose there are grade sheets of few students which might contain different grades as per their performance such as A, B, C, D
  Grades: A, B, C, D
- **Continuous**: May have an infinite number of values, it is in float type
  Example: Measuring the weight of few Students in a sequence or orderly manner i.e. 50, 51, 52, 53
  Weight: 50, 51, 52, 53

- **Discrete**: Finite number of values. Example: Marks of a Student in a few subjects: 65, 70, 75, 80, 90

Classifiers can be categorized into two major types:
Discriminative: It is a very basic classifier and determines just one class for each row of data. It tries to model just by depending on the observed data, depends heavily on the quality of data rather than on distributions.

Example: Logistic Regression Acceptance of a student at a University (Test and Grades need to be considered) Suppose there are few students and the Result of them are as follows :

Generative: It models the distribution of individual classes and tries to learn the model that generates the data behind the scenes by estimating assumptions and distributions of the model. Used to predict the unseen data.

Example: Naive Bayes Classifier

Detecting Spam emails by looking at the previous data. Suppose 100 emails and that too divided in 1:4 i.e. Class A: 25%(Spam emails) and Class B: 75%(Non-Spam emails). Now if a user wants to check that if an email contains the word cheap, then that may be termed as Spam.

It seems to be that in Class A(i.e. in 25% of data), 20 out of 25 emails are spam and rest not. And in Class B(i.e. in 75% of data), 70 out of 75 emails are not spam and rest are spam. So, if the email contains the word cheap, what is the probability of it being spam ?? (= 80%)

Classifiers Of Machine Learning:

1. Decision Trees
2. Bayesian Classifiers
3. Neural Networks
4. K-Nearest Neighbour
5. Support Vector Machines
6. Linear Regression
7. Logistic Regression

Video Content / Details of website for further learning (if any):
https://www.upgrad.com/blog/classification-and-prediction-in-data-mining/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 289 - 291)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 19

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : III - Classification And Prediction Techniques

Date of Lecture: 24.09.2021

Topic of Lecture : Classification by Decision Tree

Introduction :

Classification is a **data mining function that assigns items in a collection to target categories or classes**. The goal of classification is to accurately predict the target class for each case in the data. It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Training set
- Machine learning
- Pattern recognition, and statistics

Detailed Content of the Lecture:

Decision Tree Induction

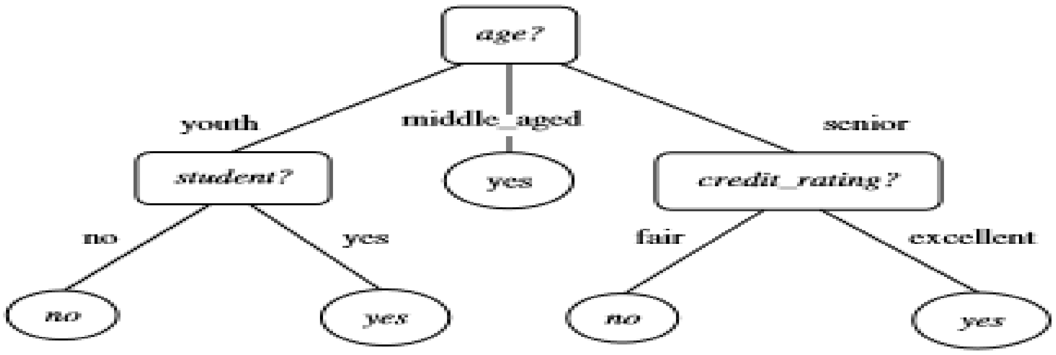
- Decision tree induction is the learning of decision trees from class-labeled training tuples.
- A decision tree is a flow chart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label.
- The topmost node in a tree is the root node.
- It represents the concept buys computer, that is, it predicts whether a customer at AllElectronics is likely to purchase a computer.
- Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals.
- Some decision tree algorithms produce only binary trees (where each internal node branches to exactly two other nodes), whereas others can produce nonbinary trees.

Algorithm: Generate decision tree. Generate a decision tree from the training tuples of data partition, D.

Input:

Data partition, D, which is a set of training tuples and their associated class labels; attribute list, the set of candidate attributes; Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split-point or splitting subset.

Output: A decision tree.



Attribute Selection Measures

- An attribute selection measure is a heuristic for selecting the splitting criterion that “best”separates a given data partition, D,of class-labeled training tuples into individual classes.
- If we were to split D into smaller partitions according to the outcomes of the splitting criterion, ideally each partition would be pure (i.e., all the tuples that fall into a given partition would belong to the same class).
- Conceptually, the “best” splitting criterion is the one that most closely results in such a scenario. Attribute selection measures are also known as splitting rules because they determine how the tuples at a given node are to be split.

Information Gain

- ID3 uses information gain as its attribute selection measure.
- This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or “information content” of messages.
- Let node N represent or hold the tuples of partition D.
- The attribute with the highest information gain is chosen as the splitting attribute for node N.

- We can compute Gain(income) = 0.029 bits, Gain(student) = 0.151 bits, and Gain(credit rating) = 0.048 bits. Because age has the highest information gain among the attributes, it is selected as the splitting attribute.
- Node N is labeled with age, and branches are grown for each of the attribute’s values. The tuples are then partitioned accordingly, Notice that the tuples falling into the partition for age = middle aged all belong to the same class. Because they all belong to class “yes,” a leaf should therefore be created at the end of this branch and labeled with “yes.”

Video Content / Details of website for further learning (if any):
https://www.saedsayad.com/decision_tree.htm#:~:text=Decision%20Tree%20%2D%20Classification,decision%20nodes%20and%20leaf%20nodes

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 291 - 299)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

LECTURE HANDOUTS **L 20**

MCA **II / III**

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : III - Classification And Prediction Techniques

Date of Lecture: 28.09.2021

Topic of Lecture : Bayesian Classification

Introduction :

Bayesian classification is based on Bayes' theorem, described next. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naıve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers.Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Neural network
- Bayesian classifier
- Statistics

Detailed Content of the Lecture:

Bayes' Theorem

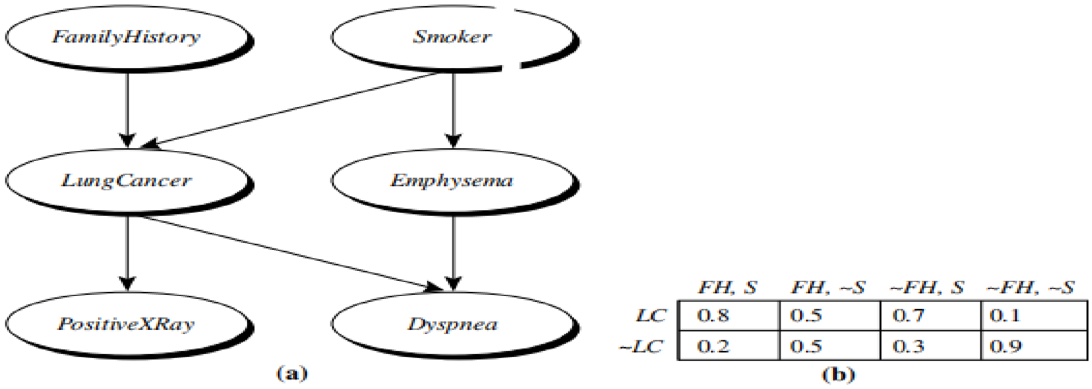
Let X be a data tuple. In Bayesian terms, X is considered "evidence." As usual, it is described by measurements made on a set of n attributes.

Let H be some hypothesis such as that the data tuple X belongs to a specified class C. For classification problems, we want to determine P(H|X), the probability that the hypothesis H holds given the "evidence" or observed data tuple X.

In other words, we are looking for the probability that tuple X belongs to class C, given that we know the attribute description of X.

- P(H|X) is the posterior probability, or a posteriori probability, of H conditionedon X.
- In contrast,P(H) is the prior probability,or apriori probability,of H.
- P(X|H) is the posterior probability of X conditioned on H.
- P(X) is the prior probability of X.

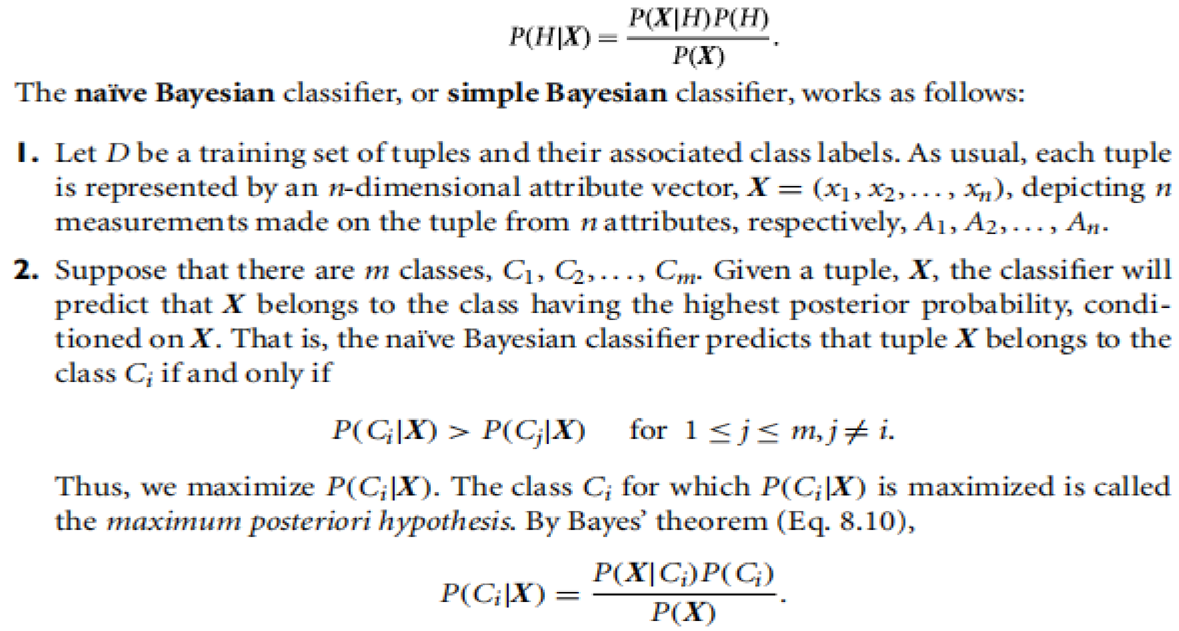
Bayes' theorem is useful in that it provides a way of calculating the posterior probability, P(H|X), from P(H), P(X|H), and P(X). Bayes' theorem is



(a)

	FH, S	FH, ~S	~FH, S	~FH, ~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

(b)



$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}.$$

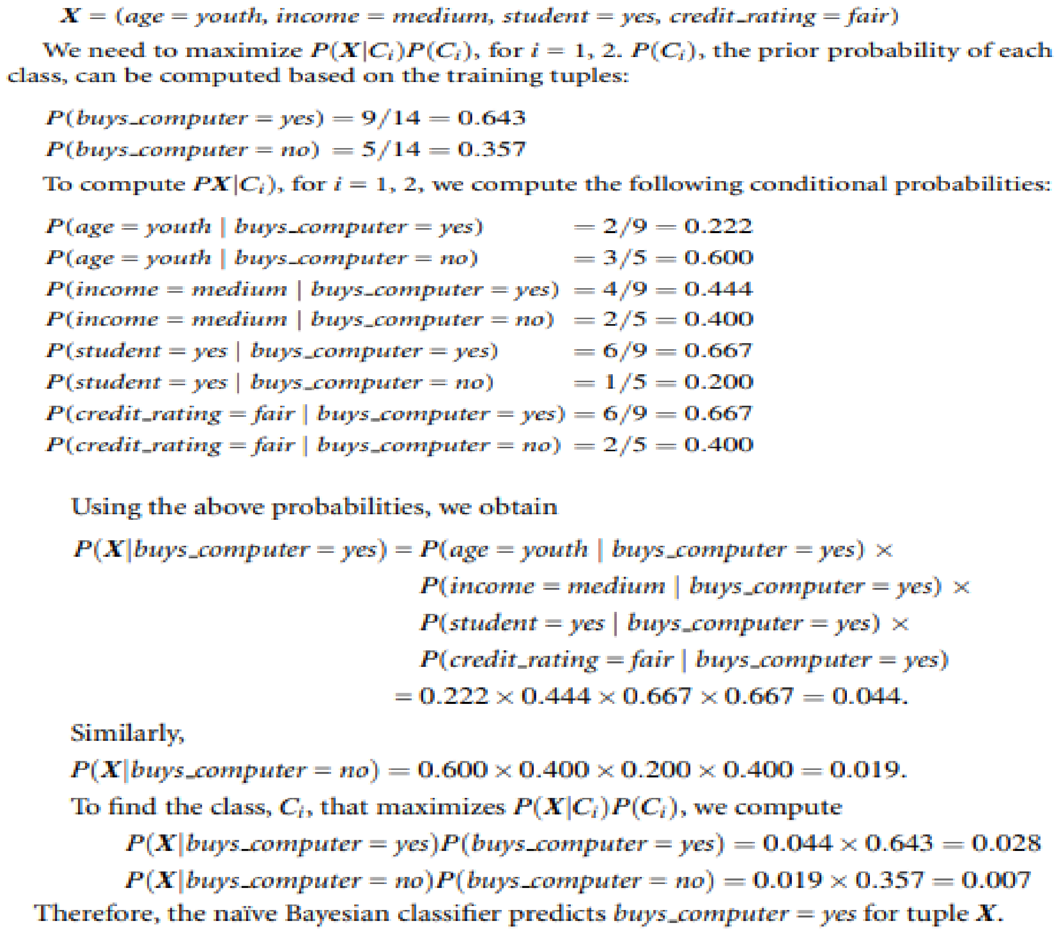
The **naïve Bayesian** classifier, or **simple Bayesian** classifier, works as follows:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, $\mathbf{X} = (x_1, x_2, \ldots, x_n)$, depicting n measurements made on the tuple from n attributes, respectively, $A_1, A_2, \ldots, A_n$.
2. Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given a tuple, $\mathbf{X}$, the classifier will predict that $\mathbf{X}$ belongs to the class having the highest posterior probability, conditioned on $\mathbf{X}$. That is, the naïve Bayesian classifier predicts that tuple $\mathbf{X}$ belongs to the class C_i if and only if

$$P(C_i|\mathbf{X}) > P(C_j|\mathbf{X}) \quad \text{for } 1 \leq j \leq m, j \neq i.$$

Thus, we maximize $P(C_i|\mathbf{X})$. The class C_i for which $P(C_i|\mathbf{X})$ is maximized is called the *maximum posteriori hypothesis*. By Bayes' theorem (Eq. 8.10),

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}.$$



$\mathbf{X} = (age = youth, income = medium, student = yes, credit\_rating = fair)$

We need to maximize $P(\mathbf{X}|C_i)P(C_i)$, for $i = 1, 2$. $P(C_i)$, the prior probability of each class, can be computed based on the training tuples:

$P(buys\_computer = yes) = 9/14 = 0.643$
$P(buys\_computer = no) = 5/14 = 0.357$

To compute $P\mathbf{X}|C_i)$, for $i = 1, 2$, we compute the following conditional probabilities:

$$\begin{aligned}
&P(age = youth \mid buys\_computer = yes) &&= 2/9 = 0.222\\
&P(age = youth \mid buys\_computer = no) &&= 3/5 = 0.600\\
&P(income = medium \mid buys\_computer = yes) &&= 4/9 = 0.444\\
&P(income = medium \mid buys\_computer = no) &&= 2/5 = 0.400\\
&P(student = yes \mid buys\_computer = yes) &&= 6/9 = 0.667\\
&P(student = yes \mid buys\_computer = no) &&= 1/5 = 0.200\\
&P(credit\_rating = fair \mid buys\_computer = yes) &&= 6/9 = 0.667\\
&P(credit\_rating = fair \mid buys\_computer = no) &&= 2/5 = 0.400
\end{aligned}$$

Using the above probabilities, we obtain

$$\begin{aligned}
P(\mathbf{X}|buys\_computer = yes) &= P(age = youth \mid buys\_computer = yes) \times\\
&\quad P(income = medium \mid buys\_computer = yes) \times\\
&\quad P(student = yes \mid buys\_computer = yes) \times\\
&\quad P(credit\_rating = fair \mid buys\_computer = yes)\\
&= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044.
\end{aligned}$$

Similarly,

$P(\mathbf{X}|buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$

To find the class, C_i, that maximizes $P(\mathbf{X}|C_i)P(C_i)$, we compute

$$P(\mathbf{X}|buys\_computer = yes)P(buys\_computer = yes) = 0.044 \times 0.643 = 0.028$$
$$P(\mathbf{X}|buys\_computer = no)P(buys\_computer = no) = 0.019 \times 0.357 = 0.007$$

Therefore, the naïve Bayesian classifier predicts $buys\_computer = yes$ for tuple $\mathbf{X}$.

Video Content / Details of website for further learning (if any):
https://www.tutorialspoint.com/data_mining/dm_bayesian_classification.htm

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 310 - 317)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 21

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : III - Classification And Prediction Techniques

Date of Lecture: 29.09.2021

Topic of Lecture : Rule Based Classification

Introduction :
Rule Based Classification is implemented using IF-THEN Rules for Classification.
It represent the knowledge in the form of IF-THEN rules.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Training set
- Machine learning
- Decision Tree

Detailed Content of the Lecture:

Rule-Based Classification
Using IF-THEN Rules for Classification

- Rules are a good way of representing information or bits of knowledge. A rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form

IF condition THEN conclusion.

An example is rule R1, R1:

IF age=youth AND student=yes THEN buys computer=yes.

The "IF" part (or left side) of a rule is known as the rule antecedent or precondition. The "THEN" part (or right side) is the rule consequent.

- In the rule antecedent, the condition consists of one or more attribute tests (e.g., age = youth and student = yes) that are logically ANDed.
- The rule's consequent contains a class prediction (in this case, we are predicting whether a customer will buy a computer).

R1 can also be written as
R1: $(age=youth) \wedge (student=yes) \Rightarrow (buys\ computer=yes)$.

- A rule R can be assessed by its coverage and accuracy.
- Given a tuple, X, from a class labeled dataset,D,let n covers be the number of tuples covered by R;n correct be the number of tuples correctly classified by R; and|D|be the number of tuples in D.
- We can define the coverage and accuracy of R.

$$coverage(R) = \frac{n_{covers}}{|D|}$$

$$accuracy(R) = \frac{n_{correct}}{n_{covers}}.$$

Rule Induction Using a Sequential Covering Algorithm
Algorithm: Sequential covering. Learn a set of IF-THEN rules for classification.

Input:

D, a data set of class-labeled tuples;
Att vals, the set of all attributes and their possible values.

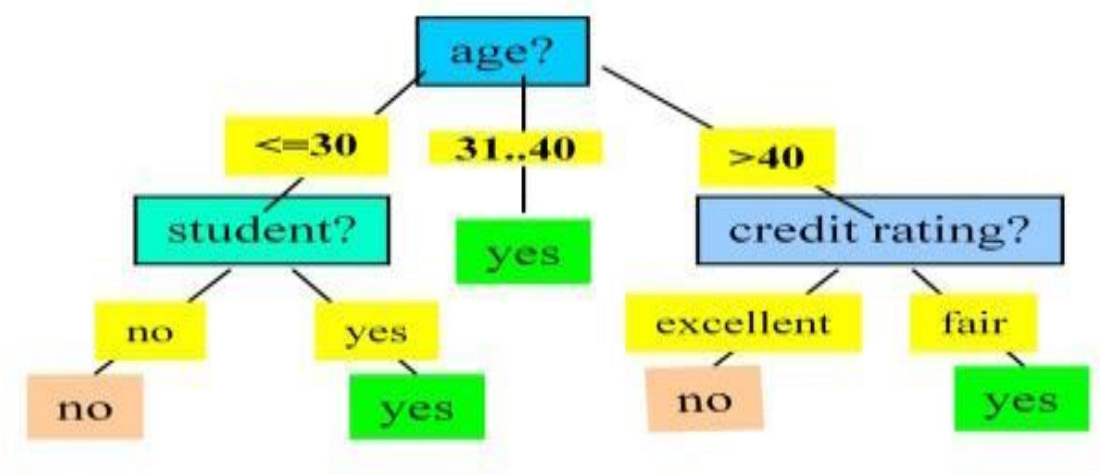
Output: A set of IF-THEN rules.

n_{covers} = # of tuples covered by R

$n_{correct}$ = # of tuples correctly classified by R

coverage(R) = n_{covers} /|D| /* D: training data set */

accuracy(R) = $n_{correct}$ / n_{covers}



R: IF age = youth AND student = yes THEN buys_computer = yes
Rule antecedent/precondition vs. rule consequent
Assessment of a rule: coverage and accuracy
In other words, we are looking for the probability that tuple X belongs to class C, given that we know the attribute description of X.

Video Content / Details of website for further learning (if any):
https://www.tutorialspoint.com/data_mining/dm_rbc.htm

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 318 - 326)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 22

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : III - Classification And Prediction Techniques

Date of Lecture: 01.10.2021

Topic of Lecture : Bayesian Belief Networks

Introduction :

Bayesian Classification is based on Bayes' theorem, described next.
A belief network is defined by two components—a directed acyclic graph and a set of conditional probability tables.

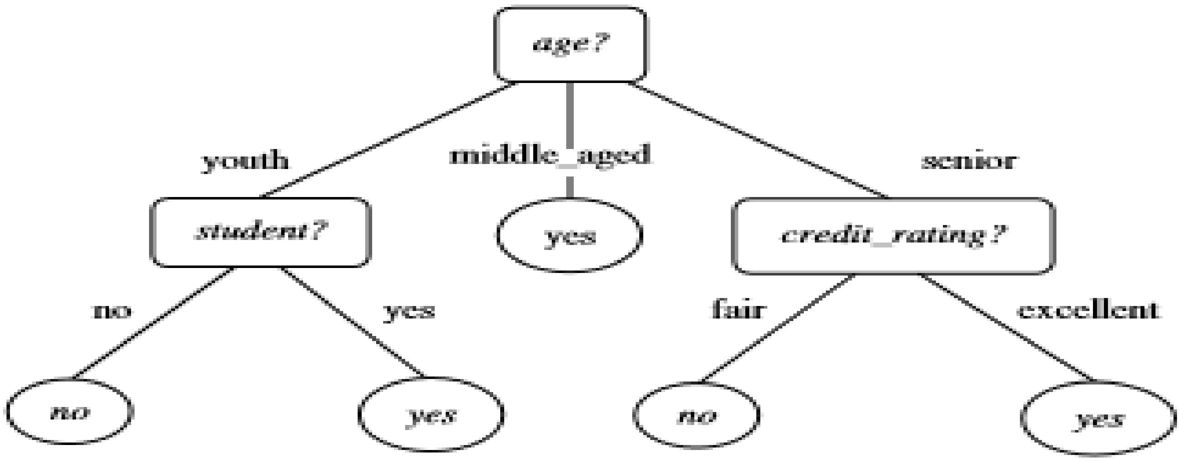
Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Training set
- Bayesian Classification
- Pattern recognition, and statistics

Detailed Content of the Lecture:

Each node in the directed acyclic graph represents a random variable. The variables may be discrete or continuous-valued. They may correspond to actual attributes given in the data or to ―hidden variables‖ believed to form a relationship (e.g., in the case of medical data, a hidden variable may indicate a syndrome,representing a number of symptoms that, together, characterize a specific disease).

Each arc represents a probabilistic dependence.If an arcis drawn from a node Y to a node Z, then Y is a parentor immediate predecessor of Z, and Z is a descendant of Y. Each variable is conditionally independent of its non descendants in the graph,
given its parents.



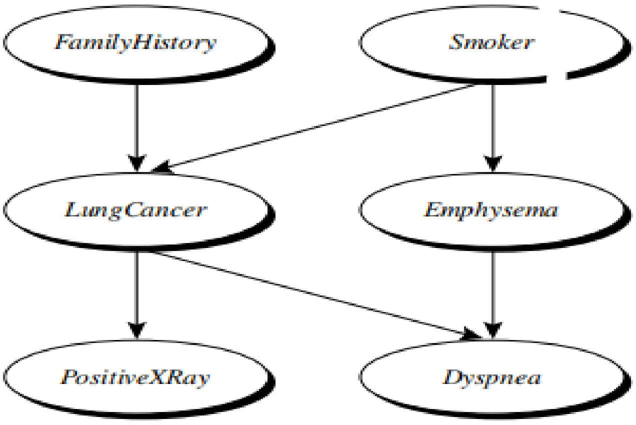
Attribute Selection Measures

- An attribute selection measure is a heuristic for selecting the splitting criterion that “best”separates a given data partition, D,of class-labeled training tuples into individual classes.
- If we were to split D into smaller partitions according to the outcomes of the splitting criterion, ideally each partition would be pure (i.e., all the tuples that fall into a given partition would belong to the same class).
- Conceptually, the “best” splitting criterion is the one that most closely results in such a scenario. Attribute selection measures are also known as splitting rules because they determine how the tuples at a given node are to be split.

Information Gain

- ID3 uses information gain as its attribute selection measure.
- This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or “information content” of messages.
- Let node N represent or hold the tuples of partition D.
- The attribute with the highest information gain is chosen as the splitting attribute for node N.

- A belief network has one conditional probability table (CPT) for each variable. The CPT for a variable Y specifies the conditional distribution P(Yj Parents(Y)), where Parents(Y) are the parents of Y. Figure(b) shows a CPT for the variable Lung Cancer. The conditional probability for each known value of Lung Cancer is given for each possible combination of values of its parents.



(a)

	FH, S	FH, ~S	~FH, S	~FH, ~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

(b)

Video Content / Details of website for further learning (if any):
https://towardsdatascience.com/introduction-to-bayesian-belief-networks-c012e3f59f1b

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 315 - 317)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 23

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : III - Classification And Prediction Techniques

Date of Lecture: 05.10.2021

Topic of Lecture : Classification by Back Propagation

Introduction :

Back propagation is a neural network learning algorithm.The neural networks field was originally kindled by psychologists and neuro biologists who sought to develop and test computational analogs of neurons. Neural network learning is also referred to as connectionist learning due to the connections between units.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Neural network
- Back propagation
- Classification
- Learning algorithm

Detailed Content of the Lecture:

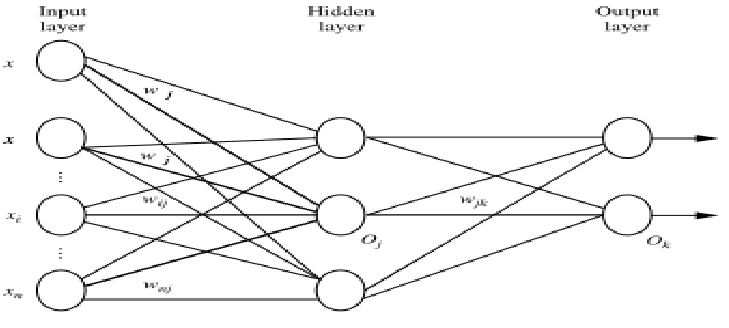
Classification by Back propagation

A Multilayer Feed-Forward Neural Network

The back propagation algorithm performs learning on a multilayer feed-forward neural network. It iteratively learns a set of weights for prediction of the class label of tuples.

A multi layer feed-forward neural network consists of an input layer,one or more hidden layers, and an output layer.Each layer is made up of units.The inputs to the network correspond to the attributes measured for each training tuple.

The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of "neuron like" units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used.



Multilayer feed-forward neural network.

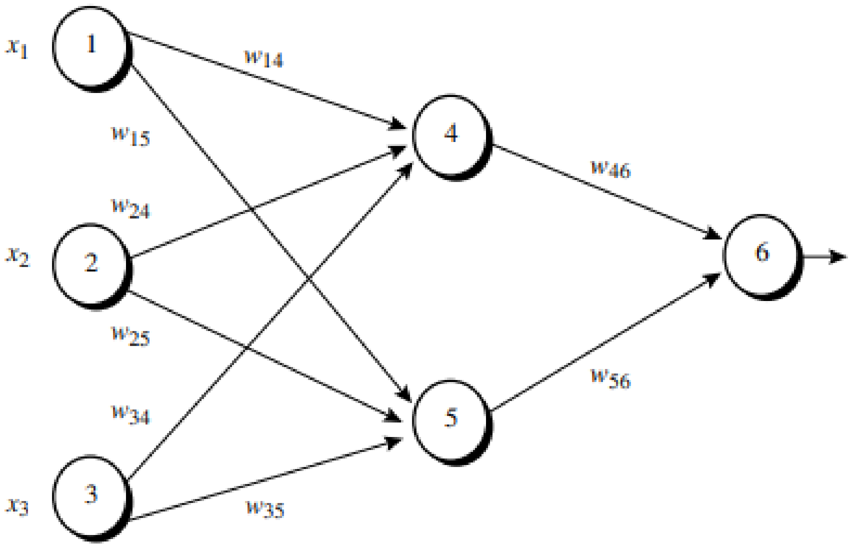
Algorithm:

Back propagation. Neural network learning for classification or numeric prediction, using the back propagation algorithm.

Input:

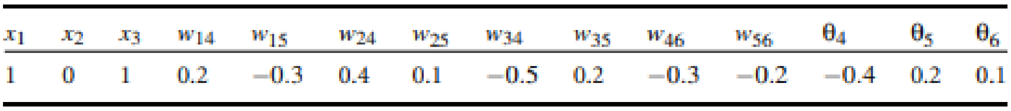
D, a data set consisting of the training tuples and their associated target values;
l, the learning rate; network, a multilayer feed-forward network.

Output: A trained neural network.



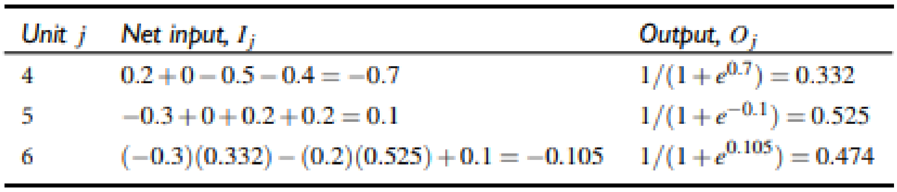
An example of a multilayer feed-forward neural network.

Initial input, weight, and bias values.



x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	−0.3	0.4	0.1	−0.5	0.2	−0.3	−0.2	−0.4	0.2	0.1

The net input and output calculations.



Unit j	*Net input, I_j*	*Output, O_j*
4	$0.2+0-0.5-0.4=-0.7$	$1/(1+e^{0.7})=0.332$
5	$-0.3+0+0.2+0.2=0.1$	$1/(1+e^{-0.1})=0.525$
6	$(-0.3)(0.332)-(0.2)(0.525)+0.1=-0.105$	$1/(1+e^{0.105})=0.474$

Video Content / Details of website for further learning (if any):
https://www.brainkart.com/article/Classification-by-Backpropagation_8324/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 327 - 330)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 24

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : III - Classification And Prediction Techniques

Date of Lecture: 06.10.2021

Topic of Lecture : Support Vector Machines

Introduction :

It is a new classification method for both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyperplane (i.e., decision boundary).

Prerequisite knowledge for Complete understanding and learning of Topic:

- Bayesian Classification
- Data Set
- Machine learning
- Back Propagation

Detailed Content of the Lecture:

Support Vector Machines (SVM) : A new classification method for both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyperplane (i.e., decision boundary). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. SVM finds this hyperplane using support vectors (essential training tuples) and margins (defined by the support vectors).

Features: training can be slow but accuracy is high owing to their ability to model complex nonlinear decision boundaries (margin maximization). Used both for classification and prediction.

Applications: handwritten digit recognition, object recognition, speaker identification, bench marking time-series prediction tests.

The Case When the Data Are Linearly Separable : An SVM approaches this problem by searching for the maximum marginal hyperplane. Consider the below Figure, which shows two possible separating hyper planes and their associated margins. Before we get into the definition of margins, let's take an intuitive.

Both hyper planes can correctly classify all of the given data tuples. Intuitively,however,we expect the hyperplane with the larger margin to be more accurate at classifying future data tuples than the hyperplane with the smaller margin.

This is why (during the learning or training phase), the SVM searches for the hyperplane with the largest margin, that is, the maximum marginal hyperplane (MMH).

The associated margin gives the largest separation between classes. Getting to an informal definition of margin, we can say that the shortest distance from a hyperplane to one side of its margin is equal to the shortest distance from the hyperplane to the other side of its margin, where the sides of the margin are parallel to the hyperplane.

When dealing with the MMH, this distance is, in fact, the shortest distance from the MMH to the closest training tuple of either class.

The Case When the Data Are Linearly Inseparable : We learned about linear SVMs for classifying linearly separable data, but what if the data are not linearly separable no straight line can be found that would separate the classes.

The linear SVMs we studied would not be able to find a feasible solution here.
The approach described for linear SVMs can be extended to create nonlinear SVMs for the classification of linearly inseparable data (also called non linearly separable data, or nonlinear data, for short).

Such SVMs are capable of finding nonlinear decision boundaries (i.e., nonlinear hyper surfaces) in input space.

Lazy Learners (or Learning from Your Neighbors) : The classification methods like decision tree induction, Bayesian classification, rule-based classification, classification by back propagation, support vector machines, and classification based on association rule mining are all examples of eager learners.
Eager learners, when given a set of training tuples, will construct a generalization (i.e., classification) model before receiving new (e.g., test) tuples to classify.

We can think of the learned model as being ready and eager to classify previously unseen tuples.

A contrasting lazy approach, in which the learner instead waits until the last minute before doing any model construction to classify a given test tuple. That is, when given a training tuple, a lazy learner simply stores it (or does only a little minor processing) and waits until it is given a test tuple.

Only when it sees the test tuple does it perform generalization to classify the tuple based on its similarity to the stored training tuples.

Video Content / Details of website for further learning (if any):
https://en.wikipedia.org/wiki/Support-vector_machine

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 337 - 340)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 25

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : III - Classification And Prediction Techniques

Date of Lecture: 08.10.2021

Topic of Lecture : K-Nearest Neighbor Algorithm

Introduction :

The k-nearest-neighbor method was first described in the early 1950s. The method is labor intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification Methods
- SVM
- Machine learning Approach
- Statistics Approach

Detailed Content of the Lecture:

K-Nearest Neighbor Algorithm

K-Nearest-neighbor classifiers are based on learning by analogy,that is,by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all of the training tuples are stored in an n-dimensional pattern space.

When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k-nearest neighbors‖ of the unknown tuple. Closeness is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, X1 = (x11, x12, : : : , x1n) and X2 = (x21, x22, : : , x2n), is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

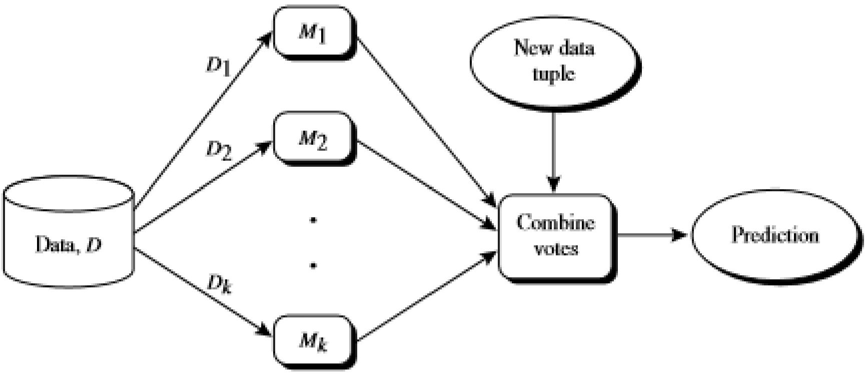
Case-Based Reasoning : Case-based reasoning (CBR) classifiers use a database of problem solutions to solve new problems. Unlike nearest-neighbor classifiers, which store training tuples as points in Euclidean space, CBR stores the tuples or ―cases‖ for problem solving as complex symbolic descriptions.

Business applications of CBR include problem resolution for customer service help desks, where cases describe product-related diagnostic problems. CBR has also been applied to areas such as engineering and law, where cases are either technical designs or legal rulings, respectively. Medical education is another area for CBR, where patient case histories and treatments are used to help diagnose and treat new patients.

Introducing Ensemble Methods

Bagging, boosting, and random forests are examples of ensemble methods. An ensemble combines a series of k learned models (or base classifiers), M1, M2,..., Mk, with the aim of creating an improved composite classification model, M∗.

A given data set, D, is used to create k training sets, D1, D2,..., Dk, where Di (1≤i≤k−1) is used to generate classifier Mi.
Given a new data tuple to classify, the base classifiers each vote by returning a class prediction.
The ensemble returns a class prediction based on the votes of the base classifiers.



General approaches for improving the classification accuracy of class-imbalanced data.
These approaches include (1) oversampling, (2) under sampling, (3) threshold moving, and (4)

The first three do not involve any changes to the construction of the classification model.
That is, oversampling and under sampling change the distribution of tuples in the training set; threshold moving affects how the model makes decisions when classifying new data.

Both oversampling and under sampling change the training data distribution so that there are(positive) class is well represented.Over sampling works by resampling the positive tuples so that the resulting training set contains an equal number of positive and negative tuples.

Under sampling works by decreasing the number of negative tuples. It randomly eliminates tuples from the majority (negative) class until there are an equal number of positive and negative tuples.

The threshold-moving approach to the class imbalance problem does not involve any sampling. It applies to classifiers that, given an input tuple, return a continuous output value That is, for an input tuple, X, such a classifier returns as output a mapping, f (X)→[0,1].

In the simplest approach,tuples for which f (X)≥t, for some threshold, t, are considered positive, while all other tuples are considered negative.

Other approaches may involve manipulating the outputs by weighting.

Video Content / Details of website for further learning (if any):
https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 348 - 350)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 26

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : III - Classification And Prediction Techniques

Date of Lecture: 20.10.2021

Topic of Lecture : Linear Regression, Nonlinear Regression, Other Regression-Based Methods

Introduction :

Linear and Nonlinear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function. Simple linear regression relates two variables (X and Y) with a straight line (y = mx + b), while nonlinear regression relates the two variables in a nonlinear (curved) relationship.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification Methods
- SVM
- Machine learning Approach
- K- Nearest Neighbor Method.

Detailed Content of the Lecture:

The goal of the model is to make the sum of the squares as small as possible. The sum of squares is a measure that tracks how far the Y observations vary from the nonlinear (curved) function that is used to predict Y.

It is computed by first finding the difference between the fitted nonlinear function and every Y point of data in the set. Then, each of those differences is squared. Lastly, all of the squared figures are added together. The smaller the sum of these squared figures, the better the function fits the data points in the set. Nonlinear regression uses logarithmic functions, trigonometric functions, exponential functions, power functions, Lorenz curves, Gaussian functions, and other fitting methods.

Nonlinear regression modeling is similar to linear regression modeling in that both seek to track a particular response from a set of variables graphically. Nonlinear models are more complicated than linear models to develop because the function is created through a series of approximations (iterations) that may stem from trial-and-error. Mathematicians use several established methods, such as the Gauss-Newton method and the Levenberg-Marquardt method.

Example of Nonlinear Regression :

One example of how nonlinear regression can be used is to predict population growth over time. A scatterplot of changing population data over time shows that there seems to be a relationship between time and population growth, but that it is a nonlinear relationship, requiring the use of a nonlinear regression model. A logistic population growth model can provide estimates of the population for periods that were not measured, and predictions of future population growth.

Nonlinear Regression Prediction Model

Due to the complex nature and variety of real-world data, it is very clumsy and inaccurate to use a simple linear relationship to describe the changing and trend of a time series.

The nonlinear regression modeling should be used to better describe those data as shown by the red curve (dark gray in print versions) in Fig. 3A. Between each two partition blocks, a nonlinear regression function should be calculated for prediction. For example, as shown in Fig. 3B, with the sampled data points from i to i+ l, some function should be calculated by a nonlinear regression model for inner section prediction.

How can we model data that does not show a linear dependence? For example, what if a given response variable and predictor variable have a relationship that may be modeled by a polynomial function?" Think back to the straight-line linear regression case above where dependent response variable, *y*, is modeled as a linear function of a single independent predictor variable, *x*.

What if we can get a more accurate model using a nonlinear model, such as a parabola or some other higher-order polynomial? Polynomial regression is often of interest when there is just one predictor variable.

It can be modeled by adding polynomial terms to the basic linear model. By applying transformations to the variables, we can convert the nonlinear model into a linear one that can then be solved by the method of least squares.

Frequent patterns and their corresponding association or correlation rules characterize interesting relationships between attribute conditions and class labels, and thus have been recently used for effective classification.

Association rules show strong associations between attribute-value pairs (or *items*) that occur frequently in a given data set. Association rules are commonly used to analyze the purchasing patterns of customers in a store. Such analysis is useful in many decision-making processes, such as product placement, catalog design, and cross-marketing.

Other Regression-Based Methods :

The discovery of association rules is based on *frequent item set mining*. Many methods for frequent item set mining and the generation of association rules were described s section, we look at associative classification, where association rules are generated and analyzed for use in classification. The general idea is that we can search for strong associations between frequent patterns (conjunctions of attribute-value pairs) and class labels. Because association rules explore highly confident associations among multiple attributes, this approach may overcome some constraints introduced by decision - tree induction, which considers only one attribute at a time.

Video Content / Details of website for further learning (if any):
https://statisticsbyjim.com/regression/difference-between-linear-nonlinear-regression-models/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 355- 359)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 27

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : III - Classification And Prediction Techniques

Date of Lecture: 22.10.2021

Topic of Lecture : Prediction

Introduction :

Prediction is the task of predicting continuous (or ordered) values for given input. For example, we may wish to predict the salary of college graduates with 10 years of work experience, or the potential sales of a new product given its price. By far, the most widely used approach for numeric prediction.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Regression
- Machine learning
- Pattern Evaluation

Detailed Content of the Lecture:

Regression analysis is a good choice when all of the predictor variables are continuous valued as well. Many problems can be solved by linear regression, and even more can be tackled by applying transformations to the variables so that a nonlinear problem can be converted to a linear one.

For reasons of space, we cannot give a fully detailed treatment of regression. Instead, this section provides an intuitive introduction to the straight-line regression analysis (which involves a single predictor variable) and multiple linear regression analysis (which involves two or more predictor variables)

It provides some pointers on dealing with nonlinear regression.

It regression-based methods, such as generalized linear models, Poisson regression, log-linear models, and regression trees.

Several software packages exist to solve regression problems. Examples include SAS (www.sas.com), SPSS (www.spss.com), and S-Plus (www.insightful.com).
Another useful resource is the book Numerical Recipes in C, by Press, Flannery, Teukolsky, and Vetterling, and its associated source code.

Linear Regression Straight-line regression analysis involves a response variable, y, and a single predictor variable, x. It is the simplest form of regression, and models y as a linear function of x.

That is, $y = b+wx$

where the variance of y is assumed to be constant, and b and w are regression coefficients specifying the Y-intercept and slope of the line, respectively. The regression coefficients, w and b, can also be thought of as weights, so that we can equivalently write,

$y = w0 +w1x.$

These coefficients can be solved for by the method of least squares, which estimates the best-fitting straight line as the one that minimizes the error between the actual data and the estimate of the line.

Let D be a training set consisting of values of predictor variable, x, for some population and their associated values for response variable, y.
The training set contains |D| data points of the form **(x1, y1), (x2, y2),..., (x|D| , y|D|).**
The regression coefficients can be estimated using this method with the following equations:

$w1 = |D| \sum i=1 (xi - x)(yi - y) |D| \sum i=1 (xi - x)$

Note that earlier, we had used the notation (Xi , yi) to refer It.

Accuracy − Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

Speed − This refers to the computational cost in generating and using the classifier or predictor.

Robustness − It refers to the ability of classifier or predictor to make correct predictions from given noisy data.

Scalability − Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.

Interpret ability − It refers to what extent the classifier or predictor understands.

Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

Note − Regression analysis is a statistical methodology that is most often used for numeric prediction.

Video Content / Details of website for further learning (if any):
https://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 354 - 358)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 28

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : IV - Clustering Techniques

Date of Lecture: 26.10.2021

Topic of Lecture : Cluster Analysis

Introduction :

Cluster: A collection of data objects.similar (or related) to one another within the same group. dissimilar (or unrelated) to the objects in other groups. Cluster analysis (or clustering, data segmentation, …) Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Regression
- Prediction
- Pattern Evaluation

Detailed Content of the Lecture:

Clustering is a Unsupervised learning Concepts.As a stand-alone tool to get insight into data distribution.As a pre processing step for other algorithms.

Applications of Clustering

Biology: Taxonomy of living things like kingdom, phylum, class, order, family, genus and species. Information retrieval: To document clustering.

Land use: Identification of areas of similar land use in an earth observation database.

Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.

Climate: Understanding earth climate, find patterns of atmospheric and ocean.

Economic Science: Used for Market Research.

What Is Good Clustering?

- A good clustering method will produce high quality clusters.
- High intra-class similarity: cohesive within clusters.
- Low inter-class similarity: distinctive between clusters.
- The quality of a clustering method depends on :
- The similarity measure used by the method.
- Its implementation, and
- Its ability to discover some or all of the hidden patterns.

Types of Data in Cluster Analysis
1.Interval-Scaled variables
2. Binary variables
3. Nominal, Ordinal, and Ratio variables
4. Variables of mixed types

1. Interval-Scaled variables :

Interval-scaled variables are continuous measurements of a roughly linear scale.

Example:

weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature.

2. Binary variables :

A binary variable is a variable that can take only 2 values.

Example : Generally gender variables can take 2 variables male and female.

3. Nominal, Ordinal, and Ratio variables:

Nominal (or) Categorical variables

A generalization of the binary variable in that it can take more than 2 states.

Example : red, yellow, blue, green.

Ordinal Variables:

An ordinal variable can be discrete or continuous.

Example : Rank.

Ratio variables :

It is a positive measurement on a nonlinear scale, approximately at an exponential scale.

Example : Ae^Bt or A^e-Bt.

4. Variables of mixed types :

A database may contain all the six types of variables **symmetric binary, asymmetric binary, nominal, ordinal, interval, and ratio. Those combinedly called as mixed-type variables.**

Video Content / Details of website for further learning (if any):

https://www.qualtrics.com/au/experience-management/research/cluster-analysis/

Important Books/Journals for further learning including the page nos.:

Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 383 - 393)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 29

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : IV - Clustering Techniques

Date of Lecture: 27.10.2021

Topic of Lecture : Partitioning Methods: k-Means and k- Mediods

Introduction :

Construct various partitions and then evaluate them by some Condition, e.g., minimizing the sum of square errors.
Typical methods: k-means, k- medoids , CLARANS

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Regression
- Prediction
- Clustering

Detailed Content of the Lecture :

Partitioning Methods :

Partitioning a database D of n objects into a set of k clusters, such that the sum of squared distances is minimized (where ci is the centroid or medoid of cluster Ci)

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

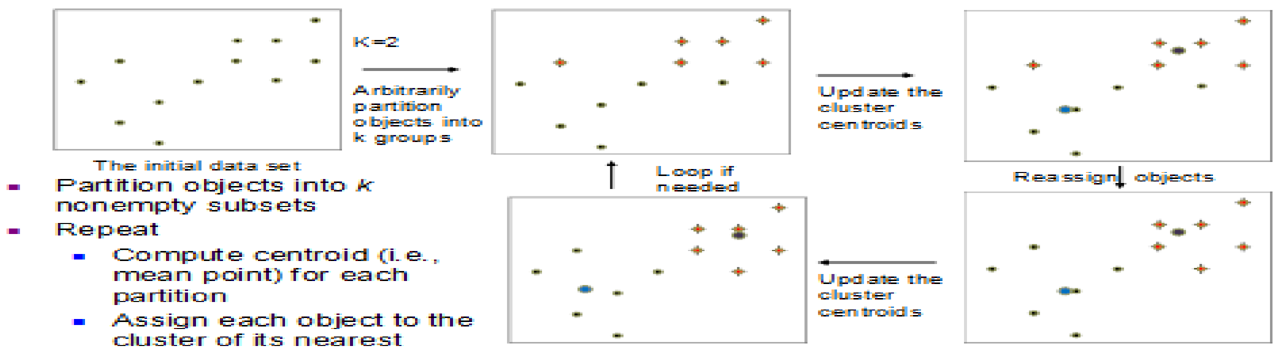
Given k, find a partition of k clusters that optimizes the chosen partitioning criterion.
Global optimal: exhaustively enumerate all partitions Heuristic methods: k-means and k-medoids algorithms
k-means : Each cluster is represented by the center of the cluster
k-medoids or PAM (Partition around medoids) : Each cluster is represented by one of the objects in the cluster

The K-Means Clustering Method :

Given k, the k-means algorithm is implemented in four steps:

Partition objects into k nonempty subsets Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., mean point, of the cluster) Assign each object to the cluster with the nearest seed point. Go back to Step 2, stop when the assignment does not change.
Strength: Efficient: O(taken), where n is # objects, k is # clusters, and t is # iterations. Normally, k, t << n.
Comparing: PAM: O(k(n-k)2), CLARA: O(ks2 + k(n-k))
Weakness: Applicable only to objects in a continuous n-dimensional space
Using the k-modes method for categorical data.
In comparison, k-medoids can be applied to a wide range of data.

An Example of *K-Means* Clustering


The initial data set
K=2
Arbitrarily partition objects into k groups
Update the cluster centroids
Reassign objects
Loop if needed
Update the cluster centroids

- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

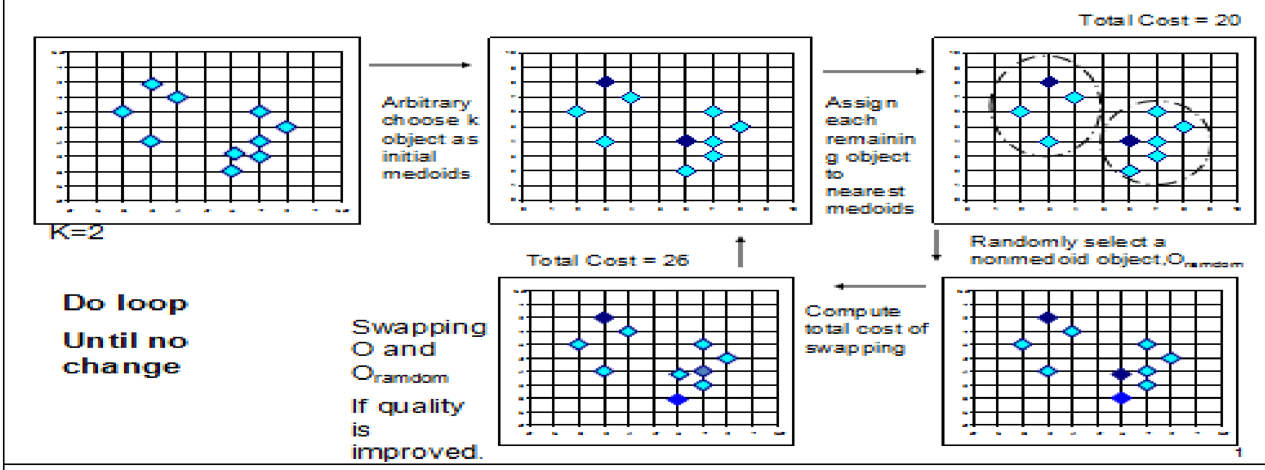
1

The K-Medoids Clustering Method :

K-Medoids Clustering: Find representative objects (medoids) in clusters.PAM (Partitioning Around Medoids) Starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering.PAM works effectively for small data sets, but does not scale well for large data sets (due to the computational complexity).

- Efficiency improvement on PAM
- CLARA (Kaufmann & Rousseeuw, 1990): PAM on samples.
- CLARANS (Ng & Han, 1994): Randomized re-sampling.

PAM: A Typical K-Medoids Algorithm


Total Cost = 20
K=2
Arbitrary choose k object as initial medoids
Assign each remaining object to nearest medoids
Randomly select a nonmedoid object,O$_{ramdom}$
Total Cost = 26
Compute total cost of swapping
Do loop
Until no change
Swapping O and O$_{ramdom}$
If quality is improved.
1

Video Content / Details of website for further learning (if any):
https://www.datamining365.com/2020/03/partitional-clustering-k-means.html

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 401 - 408)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 30

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : IV - Clustering Techniques

Date of Lecture: 02.11.2021

Topic of Lecture : Hierarchical Methods: Agglomerative and Divisive

Introduction :

Create a hierarchical decomposition of the set of data (or objects) using some Condition.

Typical methods: Agnes, Diana, BIRCH, CAMELEON.

Use distance matrix as clustering criteria.

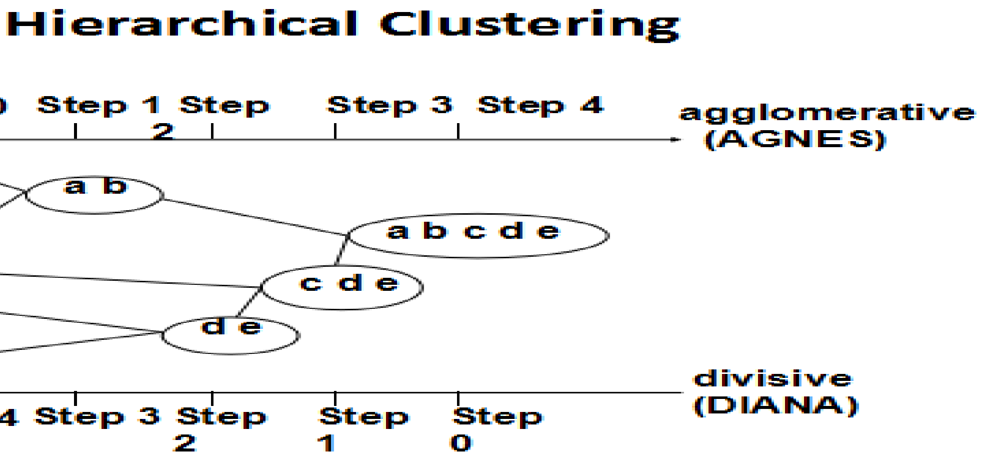
Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Regression
- Prediction
- Partitioning Method.

Detailed Content of the Lecture:

Hierarchical Methods:

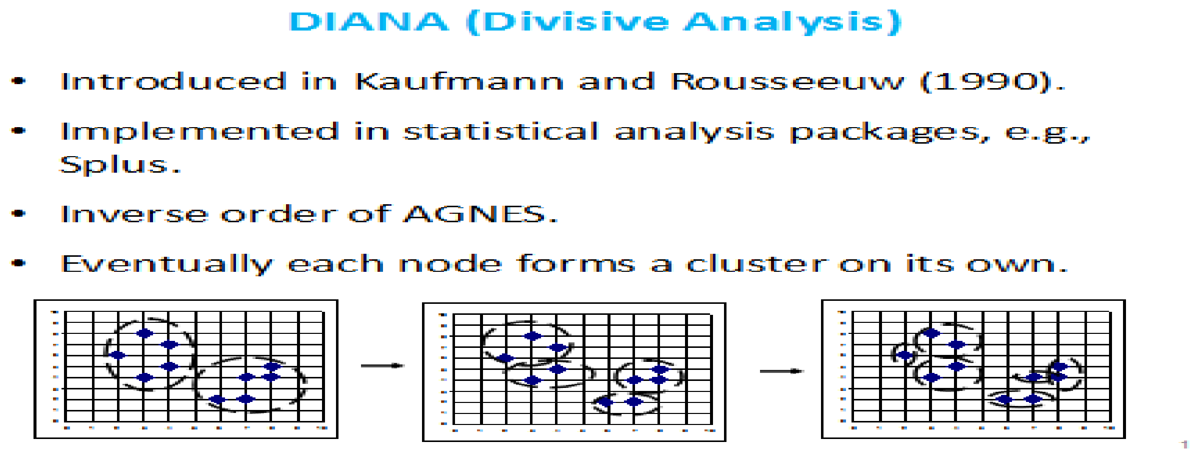
This method does not require the number of clusters k as an input, but needs a termination condition.



AGNES (Agglomerative Nesting) :

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical packages, e.g., Splus.
- Use the single-link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity.
- Go on in a non-descending fashion.
- Eventually all nodes belong to the same cluster.

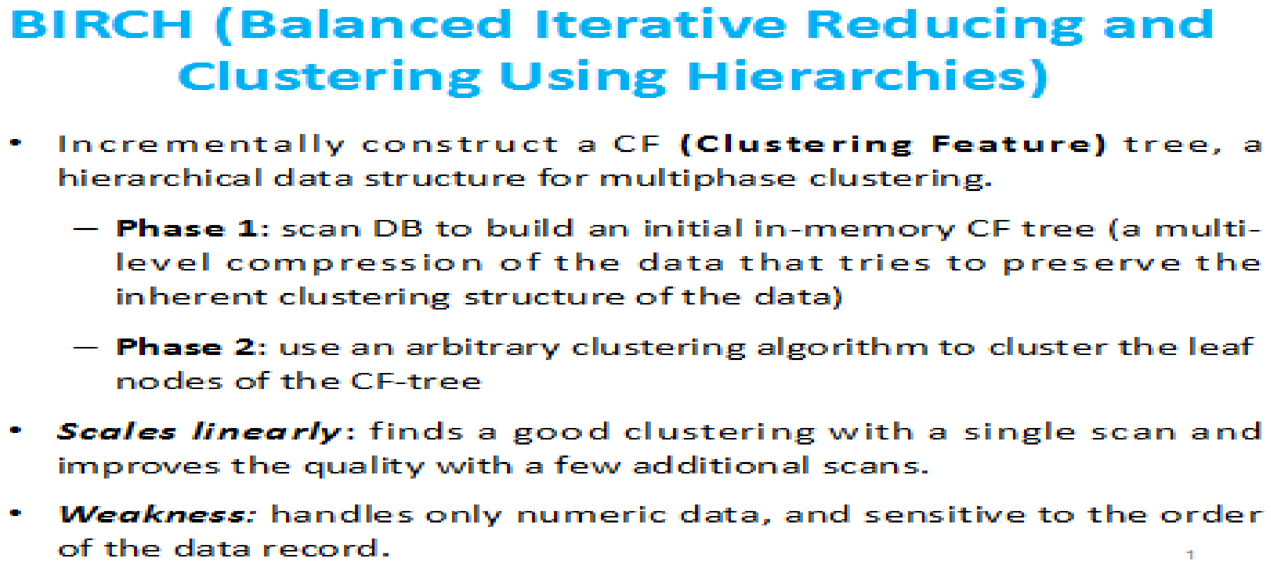
Dendrogram:
Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.
A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.



DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990).
- Implemented in statistical analysis packages, e.g., Splus.
- Inverse order of AGNES.
- Eventually each node forms a cluster on its own.

CHAMELEON: G. Karypis, E. H. Han, and V. Kumar, 1999 .Measures the similarity based on a dynamic model.Two clusters are merged only if the inter connectivity and closeness (proximity) between two clusters are high relative to the internal inter connectivity of the clusters and closeness of items within the clusters .Graph-based, and a two-phase algorithm Use a graph-partitioning algorithm: cluster objects into a large number of relatively small sub-clusters.Use an agglomerative hierarchical clustering algorithm: find the genuine clusters by repeatedly combining these sub-clusters.



BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)

- Incrementally construct a CF (**Clustering Feature**) tree, a hierarchical data structure for multiphase clustering.
  - **Phase 1:** scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)
  - **Phase 2:** use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree
- ***Scales linearly***: finds a good clustering with a single scan and improves the quality with a few additional scans.
- ***Weakness:*** handles only numeric data, and sensitive to the order of the data record.

Video Content / Details of website for further learning (if any):
https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 408 - 416)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 31

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : IV - Clustering Techniques

Date of Lecture: 09.11.2021

Topic of Lecture : Density–Based Method : DBSCAN

Introduction :

Based on connectivity and density functions.
Typical methods: DBSACN, OPTICS, DenClue
Clustering based on density (local cluster condition), such as density-connected points.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Prediction
- Clustering
- Hierarchical Methods

Detailed Content of the Lecture:

Major Features:

- Discover clusters of arbitrary shape
- Handle noise
- One scan
- Need density parameters as termination condition

Density-Based Clustering: Basic Concepts

Two parameters:

Eps: Maximum radius of the neighbourhood.
MinPts: Minimum number of points in an Eps-neighbourhood of that point.
NEps (p): {q belongs to D | dist(p,q) ≤ Eps}
Directly density-reachable: A point p is directly density-reachable from a point q w.r.t. Eps, Min Pts if p belongs to NEps(q) core point condition:
|NEps (q)| ≥ MinPts

Density-reachable:
A point p is density-reachable from a point q w.r.t. Eps, MinPts if there is a chain of points p1, …, pn, p1 = q, pn = p such that pi+1 is directly density-reachable from pi

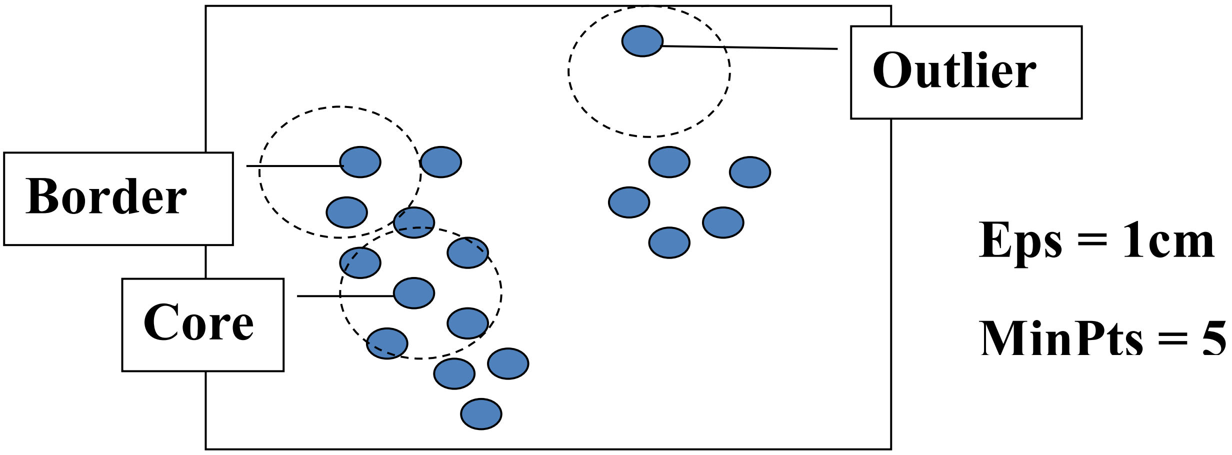
Density-connected :
A point p is density-connected to a point q w.r.t. Eps, MinPts if there is a point o such that both, p and q are density-reachable from o w.r.t. Eps and MinPts

DBSCAN: Density-Based Spatial Clustering of Applications with Noise :

Relies on a density-based notion of cluster: A cluster is defined as a maximal set of density-connected points.Discovers clusters of arbitrary shape in spatial databases with noise.

DBSCAN: The Algorithm

Arbitrary select a point p
Retrieve all points density-reachable from p w.r.t. Eps and MinPts
If p is a core point, a cluster is formed.
If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
Continue the process until all of the points have been processed.



Video Content / Details of website for further learning (if any):
https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 418 - 424)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 32

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : IV - Clustering Techniques

Date of Lecture: 10.11.2021

Topic of Lecture : Grid Based Method

Introduction :

- Based on a multiple-level granularity structure.
- Typical methods: STING, Wave Cluster, CLIQUE
- Using multi-resolution grid data structure
- A multi-resolution clustering approach using wavelet method.

Prerequisite knowledge for Complete understanding and learning of Topic:

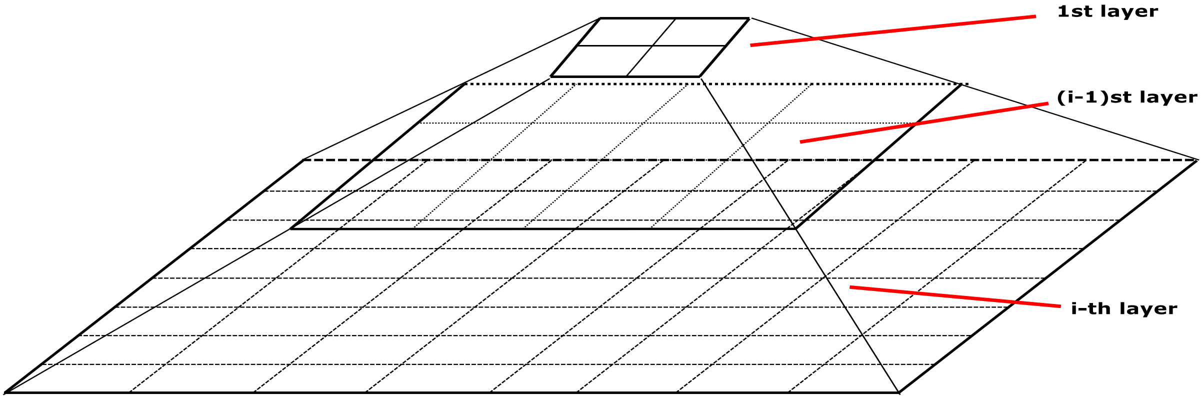
- Classification
- Prediction
- Clustering
- Density Based Method

Detailed Content of the Lecture:

Each cell at a high level is partitioned into a number of smaller cells in the next lower level.Statistical info of each cell is calculated and stored before hand and is used to answer queries.Parameters of higher level cells can be easily calculated from parameters of lower level cell.count, mean, s, min, max type of distribution—normal, uniform, etc.Use a top-down approach to answer spatial data queries.Start from a pre-selected layer—typically with a small number of cells. For each cell in the current level compute the confidence interval.

STING: A Statistical Information Grid Approach

The spatial area is divided into rectangular cells.
There are several levels of cells corresponding to different levels of resolution.



STING Algorithm and Its Analysis :

Remove the irrelevant cells from further consideration.When finish examining the current layer, proceed to the next lower level. Repeat this process until the bottom layer is reached.

Advantages:

Query-independent, easy to parallelize, incremental update.

O(K), where K is the number of grid cells at the lowest level.

Disadvantages:

All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected.

Wave Cluster: Clustering by Wavelet Analysis (1998) :

- A multi-resolution clustering approach which applies wavelet transform to the feature space; both grid-based and density-based.
- Wavelet transform: A signal processing technique that decomposes a signal into different frequency sub-band.
- Data are transformed to preserve relative distance between objects at different levels of resolution.
- Allows natural clusters to become more distinguishable.

The Wave Cluster Algorithm :

How to apply wavelet transform to find clusters Summarizes the data by imposing a multidimensional grid structure onto data space.These multidimensional spatial data objects are represented in a n-dimensional feature space. Apply wavelet transform on feature space to find the dense regions in the feature space. Apply wavelet transform multiple times which result in clusters at different scales from fine to coarse .

Major Features:

1. **Complexity O(N)**
2. **Detect arbitrary shaped clusters at different scales.**
3. **Not sensitive to noise, not sensitive to input order.**
4. **Only applicable to low dimensional data.**

CLIQUE (CLustering In QUEst) :

- Automatically identifying sub spaces of a high dimensional data space that allow better clustering than original space.
- CLIQUE can be considered as both density-based and grid-based
- It partitions each dimension into the same number of equal length interval.

Video Content / Details of website for further learning (if any):
https://www.datamining365.com/2020/04/grid-based-clustering.html

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 424 - 428)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 33

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : IV - Clustering Techniques

Date of Lecture: 12.11.2021

Topic of Lecture : Model Based Clustering Methods

Introduction :

A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other. Cluster analysis is to find hidden categories.

Typical methods: EM, SOM, COBWEB

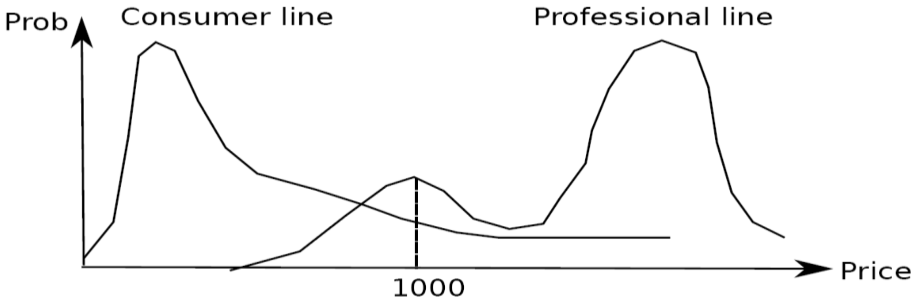
Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Clustering
- Prediction
- Hierarchical Methods.

Detailed Content of the Lecture:

Cluster analysis is to find hidden categories. A hidden category (i.e., probabilistic cluster) is a distribution over the data space, which can be mathematically represented using a probability density function (or distribution function).

Example :



Ex. 2 categories for digital cameras sold
consumer line vs. professional line
density functions f1, f2 for C1, C2
obtained by probabilistic clustering

The EM (Expectation Maximization) Algorithm

The k-means algorithm has two steps at each iteration:

Expectation Step (E-step): Given the current cluster centers, each object is assigned to the cluster whose center is closest to the object: An object is expected to belong to the closest cluster.

Maximization Step (M-step): Given the cluster assignment, for each cluster, the algorithm adjusts the center so that the sum of distance from the objects assigned to this cluster and the new center is minimized.

The (EM) algorithm: A framework to approach maximum likelihood or maximum a posteriori estimates of parameters in statistical models.
E-step assigns objects to clusters according to the current fuzzy clustering or parameters of probabilistic clusters. M-step finds the new clustering or parameters that maximize the sum of squared error (SSE) or the expected likelihood.

Computing Mixture Models with EM :

Given n objects O = {o1, …, on}, we want to mine a set of parameters Θ = {θ1, …, θk} s.t.,P(O|Θ) is maximized, where θj = (μj, σj) are the mean and standard deviation of the j-th univariate Gaussian distribution. We initially assign random values to parameters θj, then iteratively conduct the E- and M- steps until converge or sufficiently small change. At the E-step, for each object oi, calculate the probability that oi belongs to each distribution.

$$P(\Theta_j|o_i, \Theta) = \frac{P(o_i|\Theta_j)}{\sum_{l=1}^{k} P(o_i|\Theta_l)}$$

Advantages :

- Mixture models are more general than partitioning and fuzzy clustering .
- Clusters can be characterized by a small number of parameters.
- The results may satisfy the statistical assumptions of the generative models.

Disadvantages :

- Converge to local optimal (overcome: run multi-times w. random initialization)
- Computationally expensive if the number of distributions is large, or the data set contains very few observed data points.
- Need large data sets.
- Hard to estimate the number of clusters.

COBWEB

- A popular a simple method of incremental conceptual learning. Creates a hierarchical clustering in the form of a classification tree.Each node refers to a concept and contains a probabilistic description of that concept.
- It maps all the points in a high-dimensional source space into a 2 to 3-d target space, s.t., the distance and proximity relationship (i.e., topology) are preserved as much as possible.Similar to k-means: cluster centers tend to lie in a low-dimensional manifold in the feature space.Clustering is performed by having several units competing for the current object.
- SOMs are believed to resemble processing that can occur in the brain Useful for visualizing high-dimensional data in 2- or 3-D space.

Video Content / Details of website for further learning (if any):
https://www.lancaster.ac.uk/stor-i-student-sites/hamish-thorburn/2020/02/23/model-based-clustering/

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 429 - 433)**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

LECTURE HANDOUTS

L 34

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : IV - Clustering Techniques

Date of Lecture: 13.11.2021

Topic of Lecture : Clustering High Dimensional Data

Introduction :

Clustering high-dimensional data (How high is high-D in clustering?). Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Classification
- Clustering
- Prediction
- Model Based Clustering

Detailed Content of the Lecture:

Major challenges:

Many irrelevant dimensions may mask clusters.Distance measure becomes meaningless—due to equi-distance. Clusters may exist only in some sub spaces.

METHODS :

Subspace-clustering : Search for clusters existing in sub spaces of the given high dimensional data space.

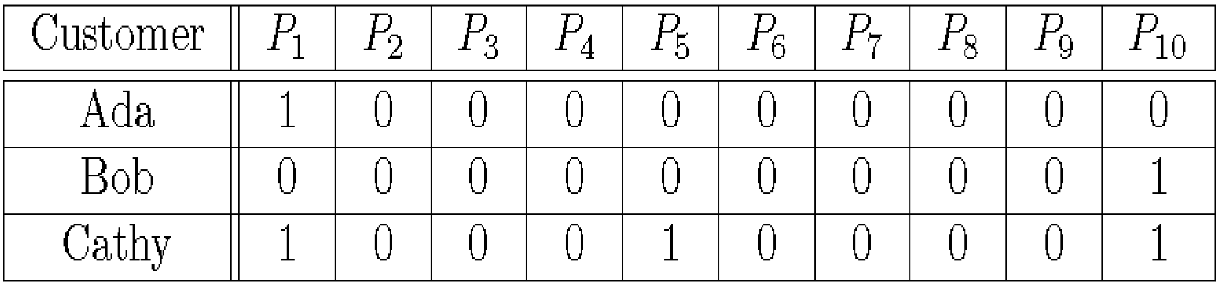
CLIQUE, ProClus, and bi-clustering approaches.

Dimensionality reduction approaches: Construct a much lower dimensional space and search for clusters there (may construct new dimensions by combining some dimensions in the original data) Dimensionality reduction methods and spectral clustering.

EXAMPLE :

Traditional distance measure could be dominated by noises in many dimensions.

Ex. Which pairs of customers are more similar?



Customer	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
Ada	1	0	0	0	0	0	0	0	0	0
Bob	0	0	0	0	0	0	0	0	0	1
Cathy	1	0	0	0	1	0	0	0	0	1

Subspace High-Dimensional Clustering Methods :

Subspace search methods:

- Search various sub spaces to find clusters
- Bottom-up approaches
- Top-down approaches

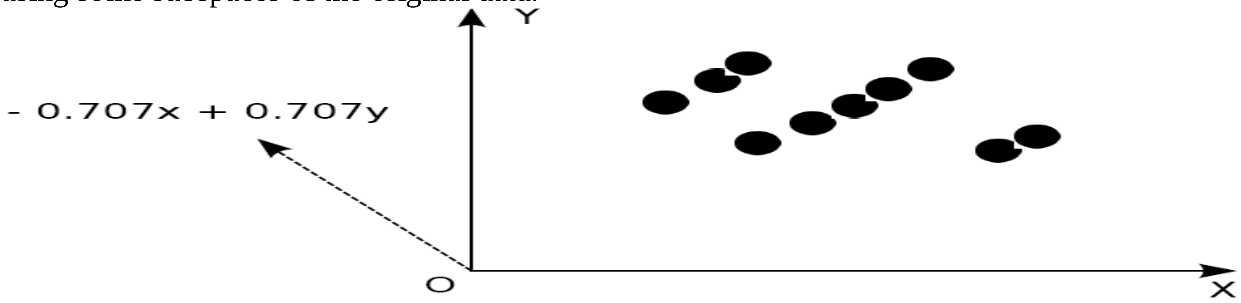
Correlation-based clustering methods

E.g., PCA based approaches

- Bi-clustering methods
- Optimization-based methods
- Enumeration methods

Dimensionality-Reduction Methods :

Dimensionality reduction: In some situations, it is more effective to construct a new space instead of using some subspaces of the original data.



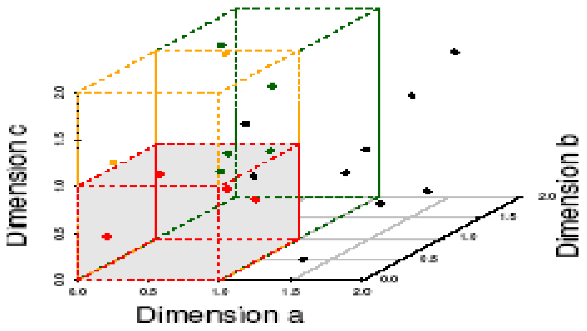
Dimensionality reduction methods :

Feature selection and extraction: But may not focus on clustering structure finding.

Spectral clustering: Combining feature extraction and clustering (i.e., use the spectrum of the similarity matrix of the data to perform dimensionality reduction for clustering in fewer dimensions)

Normalized Cuts (Shi and Malik, CVPR'97 or PAMI'2000)

The Ng-Jordan-Weiss algorithm (NIPS'01)



Video Content / Details of website for further learning (if any):

https://en.wikipedia.org/wiki/Clustering_high-dimensional_data

Important Books/Journals for further learning including the page nos.:

Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 434 - 443)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 35

MCA

II / III

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **IV - Clustering Techniques**

Date of Lecture: 19.11.2021

Topic of Lecture : Outlier Analysis

Introduction :

If the data objects does not belongs to any group it should be named as Outlier Analysis.
An outlier is a data point that differs significantly from other observations.
An outlier may be due to variability in the measurement or it may indicate experimental error.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Clustering
- Classification
- Prediction
- Learning Methods

Detailed Content of the Lecture:

Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution.outlier analysis tries to find unusual patterns in any data set. Most data mining methods discard outliers as noise or exceptions.

The analysis of outlier data is referred to as outlier mining.Outliers may be detected using statistical tests that assume a distribution or probability model for the data. Many data mining algorithms try to minimize the influence of outliers or eliminate them all together.Outlier detection and analysis is an interesting data mining task.

The easiest way to detect outliers is to create a graph. Plots such as Box plots, Scatter plots and Histograms can help to detect outliers. Alternatively, we can use mean and standard deviation to list out the outliers. Inter quartile Range and Quartiles can also be used to detect outliers.

Here is another illustration of an outlier. If you look at the Histogram below, you will see that one value lies far to the left of all other data. This data point is an outlier.

How Can Outlier Detection Improve Business Analysis?

Outlier data points can represent either a) items that are so far outside the norm that they need not be considered or b) the illustration of a very unique and singular category or variable that is worth exploring either to capitalize on a niche or find an area where an organization can offer a unique focus.

When considering the use of Outlier analysis, a business should first think about why they want to find the outliers and what they will do with that data. That focus will help the business to select the right method of analysis, graphing or plotting to reveal the results they need to see and understand.

When considering the use of Outlier analysis, it is important to recognize that, when the Outlier analysis is applied to certain datasets, the results will indicate that outliers should be discounted, while in other cases, the outlier results will indicate that the organization focus solely on those outliers.

For example, if an outlier indicates a risk or a mistake, that outlier should be identified and the risk or mistake should be addressed.

If an outlier indicates an exceptional result, such as a person that recovered from a particular disease in spite of the fact that most other patients did not survive, the organization will want to perform further analysis on the outlier result to identify the unique aspects that may be responsible for the patient's recovery.

When a business uses Outlier analysis, it is important to test the results and analyze the overall dataset and environment to be sure that the presence of outliers does not indicate that the data set may be more complex than anticipated and may require a different form of analysis.

The Smarten approach to augmented analytics and modern business intelligence focuses on the business user and provides tools for Advanced Data Discovery so users can perform early prototyping and test hypotheses without the skills of a data scientist.

Smarten Augmented Analytics tools include **assisted predictive modeling**, **smart data visualization**, **self-serve data preparation**, **Clickless Analytics** with natural language processing (NLP) for search analytics**, Auto Insights, Key Influencer Analytics, and SnapShot** monitoring and alerts.

These tools are designed for business users with average skills and require no specialized knowledge of statistical analysis or support from IT or data scientists.

Businesses can advance Citizen Data Scientist initiatives with in-person and online workshops and self-paced eLearning courses designed to introduce users and businesses to the concept, illustrate the benefits and provide introductory training on analytical concepts and the Citizen Data Scientist role.

The Smarten approach to data discovery is designed as an augmented analytics solution to serve business users.

Smarten is a representative vendor in multiple Gartner reports including the Gartner Modern BI and Analytics Platform report and the Gartner Magic Quadrant for Business Intelligence and Analytics Platforms Report.

Errors such as computational errors or incorrect entry of an object cause outliers. The differences of outliers to that of noise are:

Whenever some random error occurs in some measured variable or there is variance in the measured variable, then it is termed as noise.

Before detecting the outliers present in a dataset, it is advisable to remove the noise.

Video Content / Details of website for further learning (if any):
https://www.elegantjbi.com/blog/what-is-outlier-analysis-and-how-can-it-improve-analysis.htm

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 451 - 452)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 36

MCA

II / III

Course Name with Code : Data Mining And Data Warehousing / 19CAC16

Course Faculty : Mr. S.Nithyananth

Unit : IV - Clustering Techniques

Date of Lecture: 23.11.2021

Topic of Lecture : Outlier detection methods

Introduction :

If the data objects does not belongs to any group it should be named as Outlier Analysis.
An outlier is a data point that differs significantly from other observations.
An outlier may be due to variability in the measurement or it may indicate experimental error.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Clustering
- Classification
- Prediction
- Learning Methods

Detailed Content of the Lecture:

Methods of Outlier Detection :

There are four methods in outlier detection.

1. The statistical approach
2. The distance-based approach
3. The density-based local outlier approach
4. The deviation-based approach

1. The Statistical Approach :

This approach assumes a distribution for the given data set and then identifies outliers with respect to the model using a discordancy test. A statistical discordancy test examines **Two Hypotheses:**

A working hypothesis and
An alternative hypothesis.

A working hypothesis, H, is a statement that the entire data set of n objects comes from an initial distribution model, F, that is H:o_i ϵ F, where i= 1,2,....,n.
An alternative hypothesis, H, which states that oi comes from another distribution model (G), is adopted.
The result is very much dependent on which model is chosen because oi may be an outlier under one model and a perfectly valid value under another.

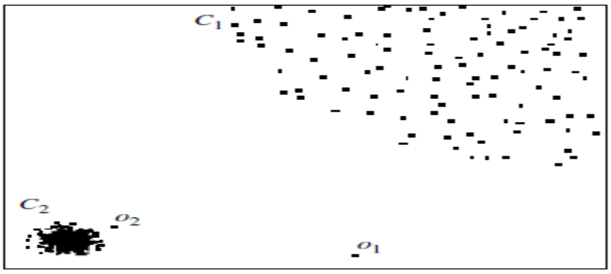
2. The Distance-Based Approach :

This approach generalizes the ideas behind discordancy testing for various standard distributions and its neighbors are defined based on their distance from the given object.
Several efficient algorithms for mining distance-based outliers have been developed.

1. Index-based algorithm
2. Nested-loop algorithm
3. Cell-based algorithm

3.The Density-Based Local Outlier Approach :

Distance based outlier detection faces difficulty in identifying outliers if data is not uniformly distributed.
Therefore this approach is used which depends on the overall distribution of the given set of data points.
Eg: The below Figure shows a simple 2-D data set containing 502 objects, with two obvious clusters.
Cluster C1 contains 400 objects while Cluster C2 contains 100 objects.
Two additional objects, o1 and o2 are clearly outliers.



4. The Deviation-Based Approach :

This approach identifies outliers by examining the main characteristics of objects in a group.
Objects that "deviate" from this description are considered outliers.
Hence, in this approach the term deviation is typically used to refer to outliers.
There are Two techniques for deviation-based outlier detection.

1. Sequentially compares objects in a set.
2. An OLAP Data Cube Approach.

Outlier data points can represent either a) items that are so far outside the norm that they need not be considered or b) the illustration of a very unique and singular category or variable that is worth exploring either to capitalize on a niche or find an area where an organization can offer a unique focus.

When considering the use of Outlier analysis, a business should first think about why they want to find the outliers and what they will do with that data. That focus will help the business to select the right method of analysis, graphing or plotting to reveal the results they need to see and understand.

When considering the use of Outlier analysis, it is important to recognize that, when the Outlier analysis is applied to certain datasets, the results will indicate that outliers should be discounted, while in other cases, the outlier results will indicate that the organization focus solely on those outliers.

Video Content / Details of website for further learning (if any):
https://towardsdatascience.com/5-outlier-detection-methods-that-every-data-enthusiast-must-know-f917bf439210

Important Books/Journals for further learning including the page nos.:
Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber Second Edition,Morgan Kaufmann Publication 2010 **(Pg.No : 452 - 458)**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 37

MCA

II / III

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **V - Data Warehouse** **Date of Lecture: 24.11.2021**

Topic of Lecture: Introduction & Need for Data Warehouse

Introduction :

- Data warehouse is an environment, not a product which is based on relational database management system that functions as the central repository for informational data.
- The central repository information is surrounded by number of key components designed to make the environment is functional, manageable and accessible.
- The data source for data warehouse is coming from operational applications.
- The data entered into the data warehouse transformed into an integrated structure and format.
- The transformation process involves conversion, summarization, filtering and condensation.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data
- Central repository
- Summarization
- Data source

Detailed content of the Lecture:

- They perform conversions, summarization, key changes, structural changes and condensation.
- The data transformation is required so that the information can by used by decision support tools.
- The transformation produces programs, control statements, JCL code, COBOL code, UNIX scripts, and SQL DDL code etc., to move the data into data warehouse from multiple operational systems.

Need for Data Warehouse:

- To remove unwanted data from operational db
- Converting to common data names and attributes
- Calculating summaries and derived data
- Establishing defaults for missing data
- Accommodating source data definition changes

Its purpose is to provide info to business users for decision making. There are five categories:

- Data query and reporting tools
- Application development tools

- Executive info system tools(EIS)
- OLAP tools
- Data mining tools

Query and reporting tools are used to generate query and report. There are two types of reporting tools. They are:

- Production reporting tool used to generate regular operationalreports
- Desktop report writer are inexpensive desktop tools designed for endusers.

Managed Query tools: used to generate SQL query. It uses Meta layer software in between users and databases which offers a point-and-click creation of SQL statement.

This tool is a preferred choice of users to perform segment identification, demographic analysis, territory management and preparation of customer mailing lists etc.

Application development tools: This is a graphical data access environment which integrates OLAP tools with data warehouse and can be used to access all dbsystems

OLAP Tools: are used to analyze the data in multi dimensional and complex views.

To enable multidimensional properties it uses MDDB and MRDB where MDDB refers multi dimensional data base and MRDB refers multi relational data bases.

Data mining tools: are used to discover knowledge from the data warehouse data also can be used for data visualization and data correction purposes.

Video Content / Details of website for further learning (if any):
https://www.herzing.edu/blog/what-data-warehousing-and-why-it-important

Important Books/Journals for further learning including the page nos.:
Alex Berson and Stephen J. Smith Data Warehousing, Data Mining & OLAP Tata McGraw Hill Edition 2007, **Pg.No 113-114**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

LECTURE HANDOUTS

L 38

MCA

II / III

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **V - Data Warehouse** **Date of Lecture: 26.11.2021**

Topic of Lecture: Data warehousing Components

Introduction :

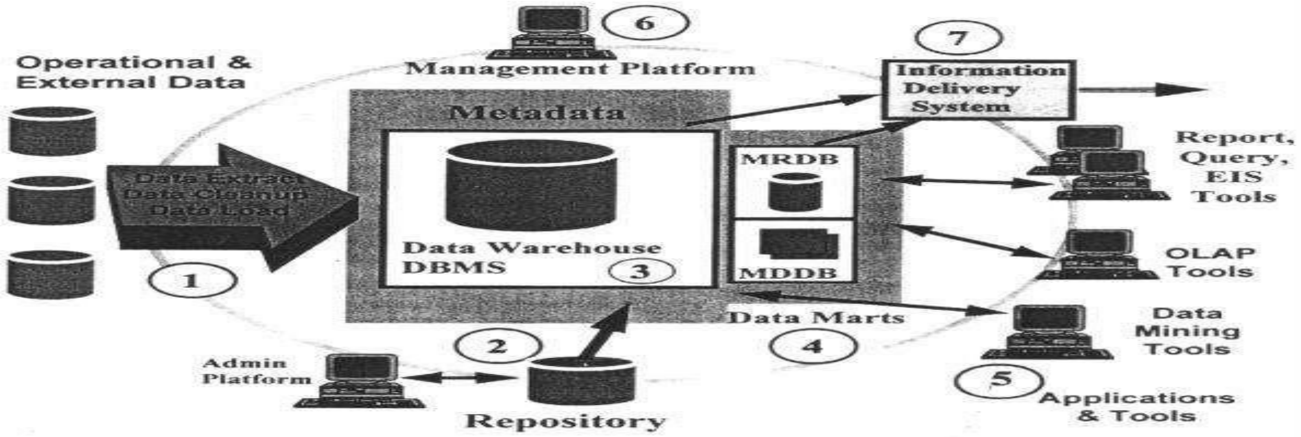
- Data warehouse is an environment, not a product which is based on relational database management system that functions as the central repository for informational data.
- They perform conversions, summarization, key changes, structural changes and condensation.
- The data transformation is required so that the information can by used by decision support tools.
- The transformation produces programs, control statements, JCL code, COBOL code, UNIX scripts

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data warehouse
- Relational Database
- Decision support
- Data source

Detailed content of the Lecture:

1. Data sourcing, cleanup, transformation, and migration tools
2. Metadata repository
3. Warehouse/database technology
4. Datamarts
5. Data query, reporting, analysis, and mining tools
6. Data warehouse administration and management
7. Information delivery system



Data warehouse environment.

Seven Major components:-

1. Data warehousedatabase

This is the central part of the data warehousing environment. This is the item number 2 in the above arch. diagram. This is implemented based on RDBMS technology.

2. Sourcing, Acquisition, Clean up, and TransformationTools

This is item number 1 in the above arch diagram. They perform conversions, summarization, key changes, structural changes and condensation. The data transformation is required so that the information can by used by decision support tools. The transformation produces programs, control statements, JCL code, COBOL code, UNIX scripts, and SQL DDL code etc., to move the data into data warehouse from multiple operationalsystems.

The functionalities of these tools are listed below:

- To remove unwanted data from operationaldb
- Converting to common data names andattributes
- Calculating summaries and deriveddata
- Establishing defaults for missingdata
- Accommodating source data definitionchanges

Issues to be considered while data sourcing, cleanup, extract and transformation:

Data heterogeneity: It refers to DBMS different nature such as it may be in different data modules, it may have different access languages, it may have data navigation methods, operations, concurrency, integrity and recovery processes etc.,

Data heterogeneity: It refers to the different way the data is defined and used in different modules. E.g Prism Solutions, Evolutionary Technology Inc., Vality, Praxis and Carleton

3. Metadata

It is data about data. It is used for maintaining, managing and using the data warehouse. It is classified intotwo:

Technical Metadata:

It contains information about data warehouse data used by warehouse designer, administrator to carry out development and management tasks. It includes,

- Information about datastores
- Transformation descriptions. That is mapping methods from operational db to warehouse db
- Warehouse Object and data structure definitions for targetdata
- The rules used to perform clean up, and dataenhancement
- Data mappingoperations
- Access authorization, backup history, archive history, info delivery history, data acquisition history, data accessetc.,

Business Meta data:

It contains info that gives info stored in data warehouse to users. It includes,

- Subject areas, and info object type including queries, reports, images, video, audio clips etc.
- Internet homepages
- Info related to info deliverysystem
- Data warehouse operational info such as ownerships, audit trailsetc.,

Meta data helps the users to understand content and find the data. Meta data are stored in a separate data stores which is known as informational directory or Meta data repository which helps to integrate, maintain and view the contents of the datawarehouse.

The following lists the **characteristics of info directory/ Meta data:**

- It is the gateway to the data warehouseenvironment
- It supports easy distribution and replication of content for high performance and availability
- It should be searchable by business oriented keywords
- It should act as a launch platform for end user to access data and analysistools
- It should support the sharing ofinformation
- It should support scheduling options forrequest
- It should support and provide interface to otherapplications
- It should support end user monitoring of the status of the data warehouse environment

4. Accesstools

Its purpose is to provide info to business users for decision making. There are five categories:

- Data query and reportingtools
- Application developmenttools
- Executive info system tools(EIS)
- OLAPtools
- Data miningtools

Query and reporting tools are used to generate query and report. There are two types of reporting tools. They are:

- Production reporting tool used to generate regular operationalreports
- Desktop report writer are inexpensive desktop tools designed for endusers.

Managed Query tools: used to generate SQL query. It uses Meta layer software in between users and databases which offers a point-and-click creation of SQL statement. This tool is a preferred choice of users to perform segment identification, demographic analysis, territory management and preparation of customer mailing lists etc.

Application development tools: This is a graphical data access environment which integrates OLAP tools with data warehouse and can be used to access all dbsystems

OLAP Tools: are used to analyze the data in multi dimensional and complex views. To enable multidimensional properties it uses MDDB and MRDB where MDDB refers multi dimensional data base and MRDB refers multi relational data bases.

Data mining tools: are used to discover knowledge from the data warehouse data also can be used for data visualization and data correction purposes.

5. Datamarts

Departmental subsets that focus on selected subjects. They are independent used by dedicated user group. They are used for rapid delivery of enhanced decision support functionality to end users. Data mart is used in the following situation:

- Extremely urgent userrequirement
- The absence of a budget for a full scale data warehousestrategy
- The decentralization of businessneeds
- The attraction of easy to use tools and mind sizedproject

Data mart presents two problems:

1. Scalability: A small data mart can grow quickly in multi dimensions. So that while designing it, the organization has to pay more attention on system scalability, consistency and manageabilityissues
2. Dataintegration

6. Data warehouse admin andmanagement

The management of data warehouse includes,

- Security and priority management
- Monitoring updates from multiple sources
- Data quality checks
- Managing and updating metadata
- Auditing and reporting data warehouse usage and status
- Purging data
- Replicating, sub setting and distributing data
- Backup and recovery
- Data warehouse storage management which includes capacity planning, hierarchical storage management and purging of aged data etc.,

7. Information delivery system

- It is used to enable the process of subscribing for data warehouse info. Delivery to one or more destinations according to specified scheduling algorithm

Video Content / Details of website for further learning (if any):
https://tdan.com/components-of-a-data-warehouse/4213

Important Books/Journals for further learning including the page nos.:
Alex Berson and Stephen J. Smith Data Warehousing, Data Mining & OLAP Tata McGraw Hill Edition 2007, **Pg.No 115-127**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 39

MCA

II / III

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **V - Data Warehouse** **Date of Lecture: 01.12.2021**

Topic of Lecture: Database versus Data Warehouse

Introduction :

- A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data warehouse
- Relational Database
- Decision support
- Data source

Detailed Content of the Lecture:

Operational Data:

- Focusing on transactional function such as bank card withdrawals and deposits
- Updateable
- Reflects current data

Informational Data:

- Focusing on providing answers to problems posed by decision makers
- Summarized
- Non updateable

Metadata

It is data about data. It is used for maintaining, managing and using the data warehouse. It is classified into two:

<u>Technical Metadata:</u>

It contains information about data warehouse data used by warehouse designer, administrator to carry out development and management tasks. It includes,

- Information about data stores
- Transformation descriptions. That is mapping methods from operational db to warehouse db
- Warehouse Object and data structure definitions for target data

- The rules used to perform clean up, and data enhancement
- Access authorization, backup history, archive history, info delivery history.

Business Meta data:

It contains info that gives info stored in data warehouse to users. It includes,

- Subject areas, and info object type including queries, reports, images, video, audio clips etc.
- Internet homepages
- Info related to info delivery system
- Data warehouse operational info such as ownerships, audit trailsetc.,

Data warehouse and database MS specialization

- Very large size of databases and need to process complex adhoc queries in a short time
- The most important requirements for the data warehouse database MS are performance, throughput and scalability.

Implementation considerations

- Collect and analyze business requirements
- Create a data model
- Define data sources
- Choose a data base technology
- Choose database access and reporting tools
- Choose database connectivity s/w

Access tools: Data warehouse implementation relies on selecting suitable data access tools. The following lists the various type of data that can be accessed:

- Simple tabular form data
- Complex textual search data
- Ranking data
- Multi variable data
- Time series data
- Graphing, charting and pivoting data

Benefits of data warehousing: The benefits can be classified into two:
Tangible benefits (quantified / measurable): Improvement in product inventory, Decrement in production cost,Improvement in selection of target markets, Enhancement in asset and liability management .

Intangible benefits (not easy to quantified): Improvement in productivity by keeping all data in single location and eliminating re - keying of data, Reduced redundant processing, Enhanced customer relation.

Video Content / Details of website for further learning (if any):
https://www.healthcatalyst.com/insights/database-vs-data-warehouse-a-comparative-review/

Important Books/Journals for further learning including the page nos.:
Alex Berson and Stephen J. Smith Data Warehousing, Data Mining & OLAP Tata McGraw Hill Edition 2007, **Pg.No 139-140**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 40

MCA

II / III

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **V - Data Warehouse** **Date of Lecture: 03.12.2021**

Topic of Lecture: Multidimensional Data Model

Introduction :

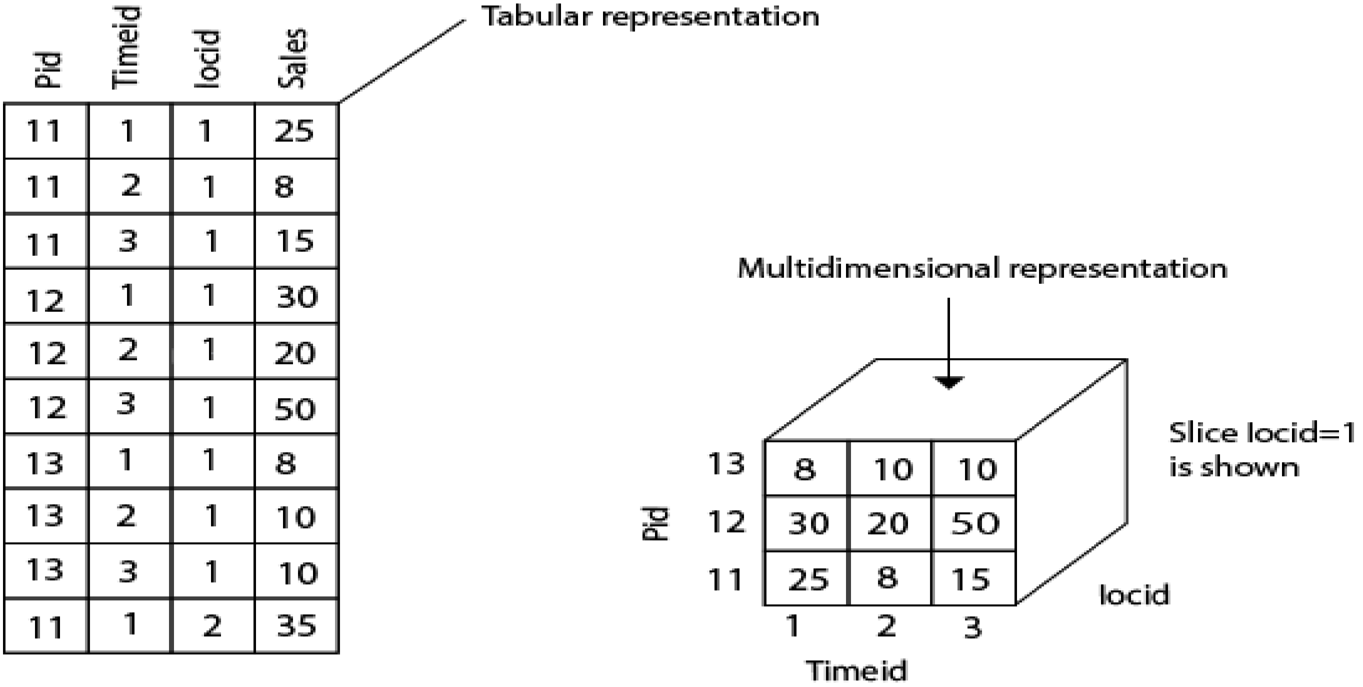
- In dimensional modeling, the transaction record is divided into either **"facts,"** which are frequently numerical transaction data, or **"dimensions,"** which are the reference information that gives context to the facts.
- For example, a sale transaction can be damage into facts such as the number of products ordered and the price paid for the products, and into dimensions such as order date, user name, product number, order ship-to, and bill-to locations, and salesman responsible for receiving the order.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Facts
- Dimensions
- Data-cube
- Data source

Detailed content of the Lecture:

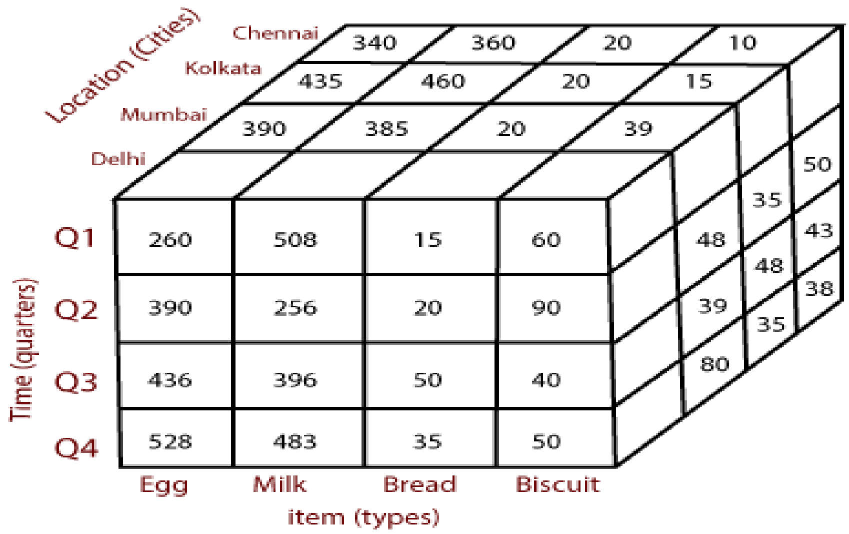
- A multidimensional model views data in the form of a data-cube. A data cube enables data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.
- The dimensions are the perspectives or entities concerning which an organization keeps records. For example, a shop may create a sales data warehouse to keep records of the store's sales for the dimension time, item, and location.
- These dimensions allow the save to keep track of things, for example, monthly sales of items and the locations at which the items were sold.
- Each dimension has a table related to it, called a dimensional table, which describes the dimension further. For example, a dimensional table for an item may contain the attributes item_name, brand, and type.
- A multidimensional data model is organized around a central theme, for example, sales. This theme is represented by a fact table. Facts are numerical measures. The fact table contains the names of the facts or measures of the related dimensional tables.



Pid	Timeid	locid	Sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35

Pid \ Timeid	1	2	3
13	8	10	10
12	30	20	50
11	25	8	15

Consider the data of a shop for items sold per quarter in the city of Delhi. The data is shown in the table. In this 2D representation, the sales for Delhi are shown for the time dimension (organized in quarters) and the item dimension (classified according to the types of an item sold). The fact or measure displayed in rupee_sold (in thousands).



	Egg	Milk	Bread	Biscuit
Q1	260	508	15	60
Q2	390	256	20	90
Q3	436	396	50	40
Q4	528	483	35	50

Video Content / Details of website for further learning (if any):
https://www.geeksforgeeks.org/multidimensional-data-model/

Important Books/Journals for further learning including the page nos.:
Alex Berson and Stephen J. Smith Data Warehousing, Data Mining & OLAP Tata McGraw Hill Edition 2007, **Pg.No 248-250**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 41

MCA

II / III

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **V - Data Warehouse** **Date of Lecture: 07.12.2021**

Topic of Lecture : Schemas for Multidimensional Databases - Star & Snowflake Schemas

Introduction :

- The basic concepts of dimensional modeling are: facts, dimensions and measures.
- A fact is a collection of related data items, consisting of measures and context data.
- It typically represents business items or business transactions.
- A dimension is a collection of data that describe one business dimension.
- Dimensions determine the contextual background for the facts; they are the parameters over which we want to perform OLAP.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Dimension modeling
- OLAP
- Facts
- Dimension

Detailed content of the Lecture:

- A measure is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions.

Considering Relational context, there are three basic schemas that are used in dimensional modeling:

1. Star schema
2. Snowflake schema
3. Fact constellation schema

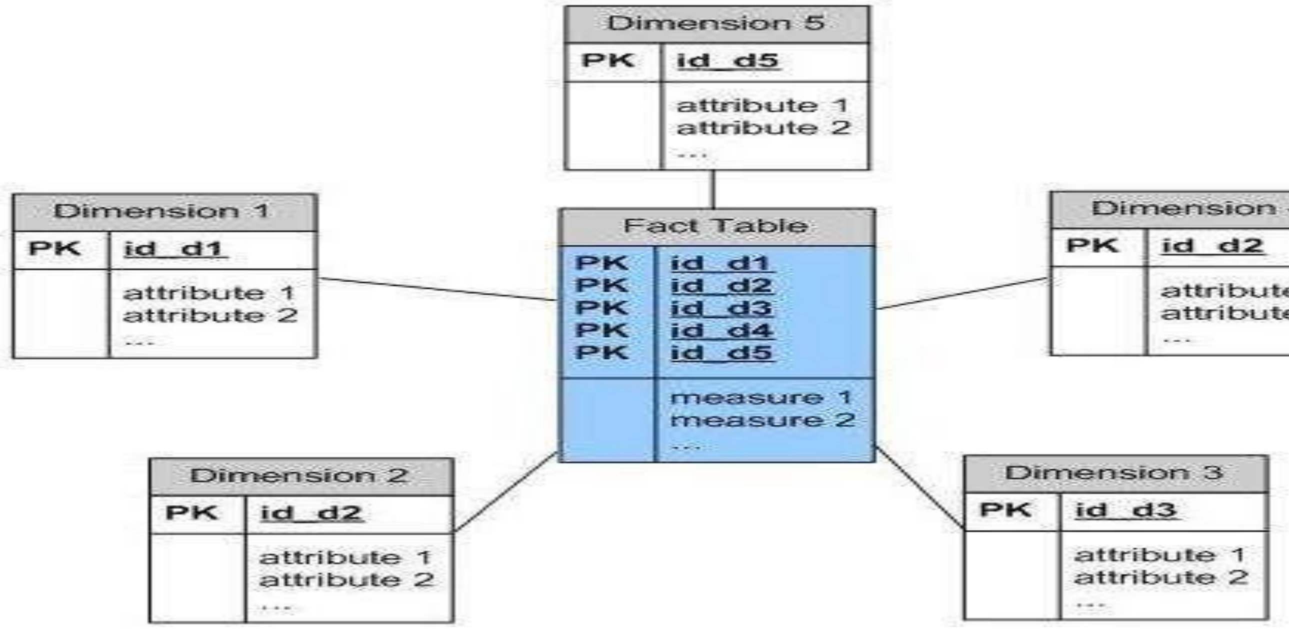
4.1. Star schema

The multidimensional view of data that is expressed using relational data base semantics is provided by the data base schema design called star schema. The basic of stat schema is that information can be classified into two groups:

- Facts
- Dimension

Star schema has one large central table (fact table) and a set of smaller tables (dimensions) arranged in a radial pattern around the centraltable.

Facts are core data element being analyzed while dimensions are attributes about the facts.



The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

- Each dimension in a star schema is represented with only one-dimensiontable.
- This dimension table contains the set ofattributes.
- There is a fact table at the center. It contains the keys to each of fourdimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

main characteristics of star schema:

- Simple structure -> easy to understand schema
- Great query effectives -> small number of tables tojoin
- Relatively long time of loading data into dimension tables -> de-normalization, redundancy data caused that size of the table could belarge.
- The most commonly used in the data warehouse implementations -> widely supported by a large number of business intelligence tools

The star schema suffers the following performance problems.

1.Indexing

Multipart key presents some problems in the star schema

model. (day->week-> month-> quarter-> year)

• It requires multiple metadata definition(one for each component) to design a singletable.

• Since the fact table must carry all key components as part of its primary key, addition or deletion of levels in the hierarchy will require physical modification of the affected table, which is time-consuming processed that limitsflexibility.

• Carrying all the segments of the compound dimensional key in the fact table increasesthe size of the index, thus impacting both performance and scalability.

2.Level Indicator.

The dimension table design includes a level of hierarchy indicator for every record.
Every query that is retrieving detail records from a table that stores details and aggregates must use this indicator as an additional constraint to obtain a correct result.

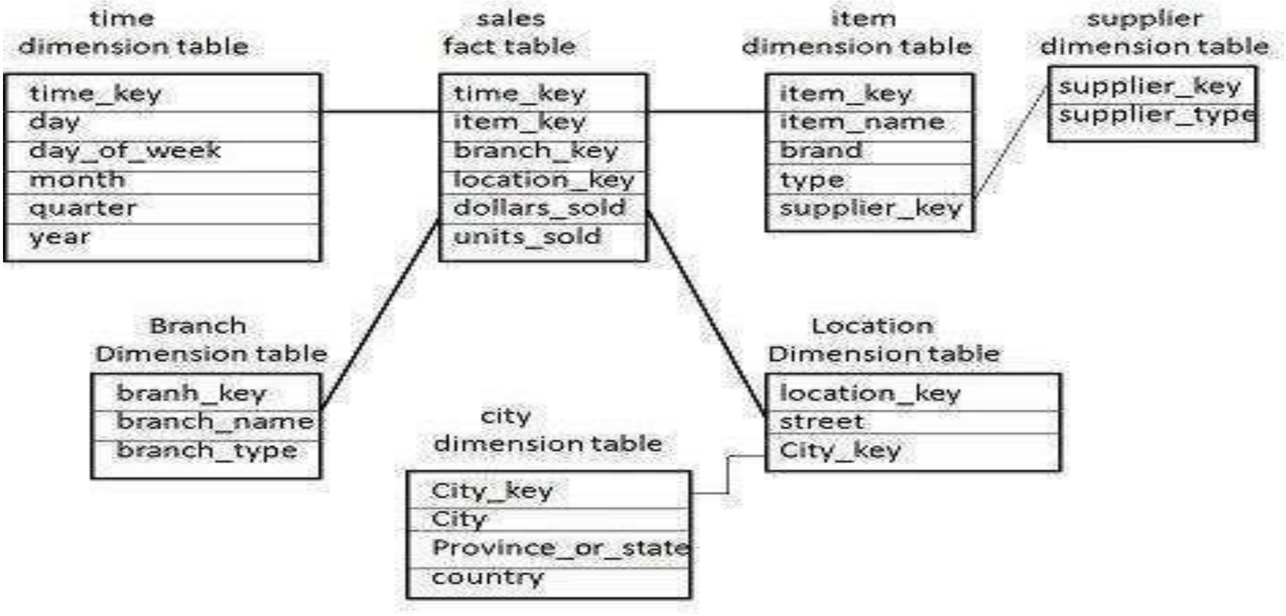
The user is not and aware of the level indicator, or its values are in correct, the otherwise valid query may result in a totally invalid answer.

Alternative to using the level indicator is the snowflake schema. Aggregate fact tables are created separately from detail tables. Snowflake schema contains separate fact tables for each level of aggregation.

Other problems with the star schema design - Pairwise Join Problem

5 tables require joining first two tables, the result of this join with third table and so on. The intermediate result of every join operation is used to join with the next table.
Selecting the best order of pairwise joins rarely can be solve in a reasonable amount of time.
Five-table query has 5!=120 combinations

2 .Snowflake schema:
is the result of decomposing one or more of the dimensions. The many-to-one relationships among sets of attributes of a dimension can separate new dimension tables, forming a hierarchy. The decomposed snowflake structure visualizes the hierarchical structure of dimensions very well.



STAR join and STARIndex.

A STAR join is high-speed, single pass, parallelizable muti-tables join method. It performs many joins by single operation with the technology called Indexing. For query processing the indexes are used in columns and rows of the selected tables.

Red Brick's RDBMS indexes, called STAR indexes, used for STAR join performance. The STAR indexes are created on one or more foreign key columns of a fact table. STAR index contains information that relates the dimensions of a fact table to the rows that contains those dimensions. STAR indexes are very space-efficient. The presence of a STAR index allows Red Brick's RDBMS to quickly identify which target rows of the fact table are of interest for a particular set of dimension. Also, because STAR indexes are created over foreign keys, no assumptions are made about the type of queries which can use the STAR indexes.

Video Content / Details of website for further learning (if any):
https://www.guru99.com/star-snowflake-data-warehousing.html

Important Books/Journals for further learning including the page nos.:
Alex Berson and Stephen J. Smith Data Warehousing, Data Mining & OLAP Tata McGraw Hill Edition 2007 **Pg.No 169-185**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 42

MCA

II / III

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **V - Data Warehouse** **Date of Lecture: 08.12.2021**

Topic of Lecture :Fact Constellation Schemas

Introduction :

- The basic concepts of dimensional modeling are: facts, dimensions and measures.
- A fact is a collection of related data items, consisting of measures and context data.
- It typically represents business items or business transactions.
- A dimension is a collection of data that describe one business dimension.
- Dimensions determine the contextual background for the facts; they are the parameters over which we want to perform OLAP.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Dimension modeling
- OLAP
- Facts
- Dimension

Detailed content of the Lecture:

Fact constellation schema:

For each star schema it is possible to construct fact constellation schema(for example by splitting the original star schema into more star schemes each of them describes facts on another level of dimension hierarchies). The fact constellation architecture contains multiple fact tables that share many dimension tables.

The main shortcoming of the fact constellation schema is a more complicated design because many variants for particular kinds of aggregation must be considered and selected. Moreover, dimension tables are still large.

Bit MappedIndexing

- ▪ SYBASE IQ is an example of a product that uses a bit mapped index structure of the data stored in the SYBASEDBMS.
- ▪ Sybase released SYBASE IQ database targeted an "ideal" data mart solution for handle multi user adhoc(unstructured)queries.

Over view:

- ▪ SYBASE IQ is a separate SQLdatabase.
- ▪ Once loaded, SYBASE IQ converts all data into a series of bit maps, which are then highly compressed and stored ondisk.
- ▪ SYBASE positions SYBASE IQ as a read only database for data marts, with a practical size limitations currently placed at 100Gbytes.

Data cardinality: Bitmap indexes are used to optimize queries against low- cardinality data that is, data in which the total number of possible values is relativelylow. (Cardinal meaning – important)

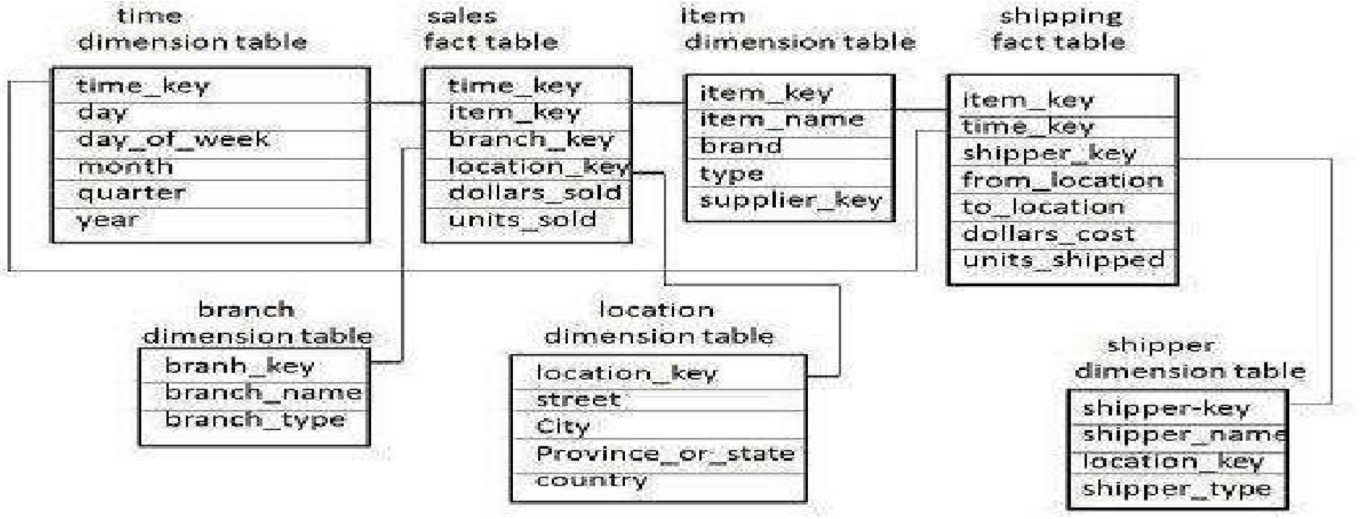
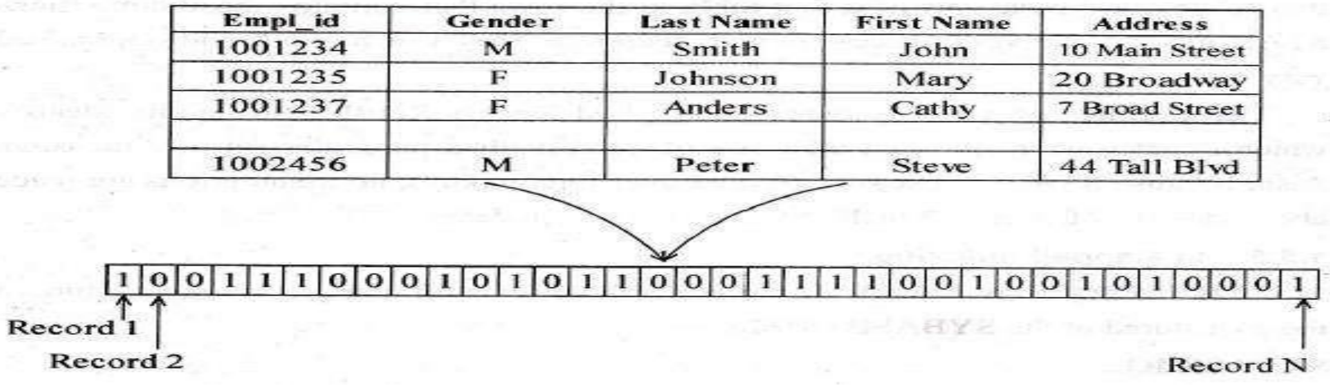


Fig: - Bitmap index

For example, address data cardinality pin code is 50 (50 possible values), and gender data cardinality is only 2 (male and female)..

If the bit for a given index is "on", the value exists in the record. Here, a 10,000 — row employee table that contains the "gender" column is bitmap-indexed for this value.

Bitmap indexes can become bulky and even unsuitable for high cardinality data where the range of possible values is high. For example, values like "income" or "revenue" may have an almost infinite number of values.



Empl_id	Gender	Last Name	First Name	Address
1001234	M	Smith	John	10 Main Street
1001235	F	Johnson	Mary	20 Broadway
1001237	F	Anders	Cathy	7 Broad Street
1002456	M	Peter	Steve	44 Tall Blvd

1 0 0 1 1 1 0 0 0 1 0 1 0 1 1 0 0 0 1 1 1 1 0 0 1 0 0 1 0 1 0 0 0 1

Video Content / Details of website for further learning (if any):
https://www.datawarehouse4u.info/Data-warehouse-schema-architecture-fact-constellation-schema.html

Important Books/Journals for further learning including the page nos.:
Alex Berson and Stephen J. Smith Data Warehousing, Data Mining & OLAP Tata McGraw Hill Edition 2007, **Pg.No 169-185**

Course Faculty

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L 43

MCA

II / III

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **V - Data Warehouse** **Date of Lecture: 13.12.2021**

Topic of Lecture : Online Analytical Processing (OLAP) Operations

Introduction :

- **OLAP** implement the multidimensional analysis of business information and support the capability for complex estimations, trend analysis, and sophisticated data modeling.
- It is rapidly enhancing the essential foundation for Intelligent Solutions containing Business Performance Management, Planning, Budgeting, Forecasting, Financial Documenting, Analysis, Simulation-Models, Knowledge Discovery, and Data Warehouses Reporting.
- OLAP enables end-clients to perform ad hoc analysis of record in multiple dimensions, providing the insight and understanding they require for better decision making.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Knowledge Discovery
- OLAP
- Analysis
- Dimension

Detailed content of the Lecture:

OLAP stands for **On-Line Analytical Processing**. OLAP is a classification of software technology which authorizes analysts, managers, and executives to gain insight into information through fast, consistent, interactive access in a wide variety of possible views of data that has been transformed from raw information to reflect the real dimensionality of the enterprise as understood by the clients.

Who uses OLAP and Why?

OLAP applications are used by a variety of the functions of an organization.

Finance and accounting:

- Budgeting
- Activity-based costing
- Financial performance analysis
- And financial modeling

Sales and Marketing

- Sales analysis and forecasting
- Market research analysis
- Promotion analysis
- Customer analysis
- Market and customer segmentation

Production

- Production planning
- Defect analysis

OLAP cubes have two main purposes. The first is to provide business users with a data model more intuitive to them than a tabular model. This model is called a Dimensional Model.

The second purpose is to enable fast query response that is usually difficult to achieve using tabular models.

OLAP Guidelines-Need

1) Multidimensional Conceptual View: This is the central features of an OLAP system. By needing a multidimensional view, it is possible to carry out methods like slice and dice.

2) Transparency: Make the technology, underlying information repository, computing operations, and the dissimilar nature of source data totally transparent to users. Such transparency helps to improve the efficiency and productivity of the users.

3) Accessibility: It provides access only to the data that is actually required to perform the particular analysis, present a single, coherent, and consistent view to the clients. The OLAP system must map its own logical schema to the heterogeneous physical data stores and perform any necessary transformations. The OLAP operations should be sitting between data sources (e.g., data warehouses) and an OLAP front-end.

4) Consistent Reporting Performance: To make sure that the users do not feel any significant degradation in documenting performance as the number of dimensions or the size of the database increases. That is, the performance of OLAP should not suffer as the number of dimensions is increased. Users must observe consistent run time, response time, or machine utilization every time a given query is run.

5) Client/Server Architecture: Make the server component of OLAP tools sufficiently intelligent that the various clients to be attached with a minimum of effort and integration programming. The server should be capable of mapping and consolidating data between dissimilar databases.

6) Generic Dimensionality: An OLAP method should treat each dimension as equivalent in both is structure and operational capabilities. Additional operational capabilities may be allowed to selected dimensions, but such additional tasks should be grantable to any dimension.

7) Dynamic Sparse Matrix Handling: To adapt the physical schema to the specific analytical model being created and loaded that optimizes sparse matrix handling. When encountering the sparse matrix, the system must be easy to dynamically assume the distribution of the information and adjust the storage and access to obtain and maintain a consistent level of performance.

8) Multiuser Support: OLAP tools must provide concurrent data access, data integrity, and access security.

9) Unrestricted cross-dimensional Operations: It provides the ability for the methods to identify dimensional order and necessarily functions roll-up and drill-down methods within a dimension or across the dimension.

10) Intuitive Data Manipulation: Data Manipulation fundamental the consolidation direction like as reorientation (pivoting), drill-down and roll-up, and another manipulation to be accomplished naturally and precisely via point-and-click and drag and drop methods on the cells of the scientific model. It avoids the use of a menu or multiple trips to a user interface.

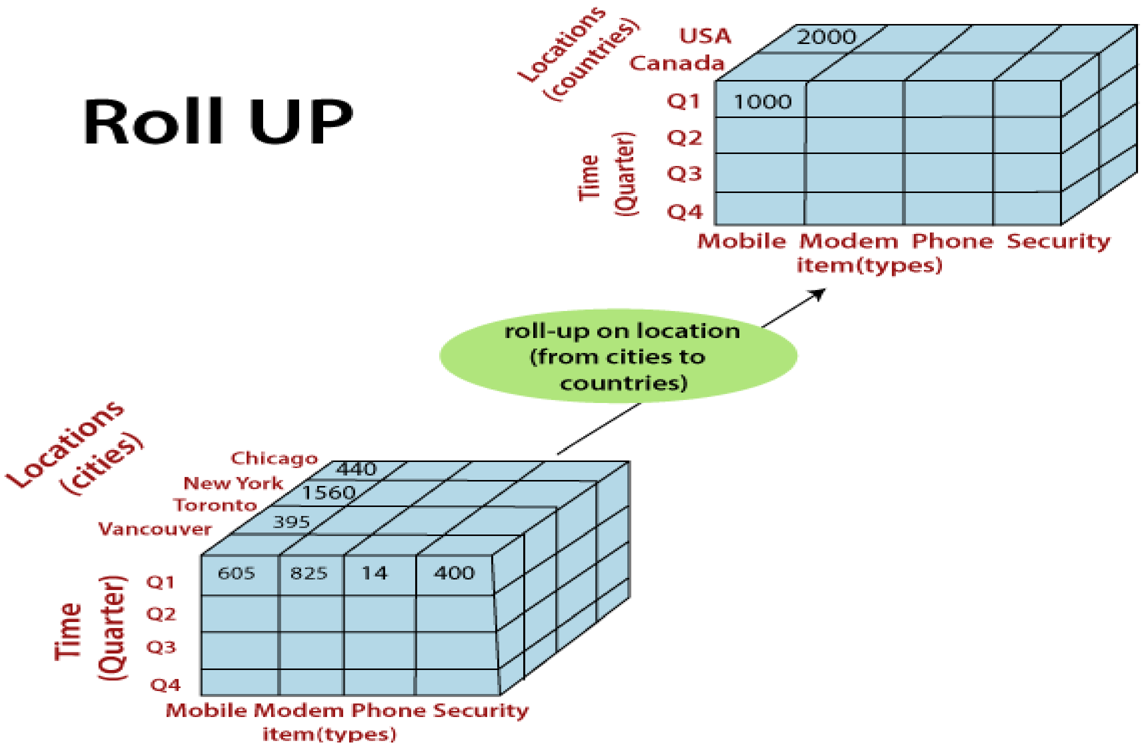
11) Flexible Reporting: It implements efficiency to the business clients to organize columns, rows, and cells in a manner that facilitates simple manipulation, analysis, and synthesis of data.

12) Unlimited Dimensions and Aggregation Levels: The number of data dimensions should be unlimited. Each of these common dimensions must allow a practically unlimited number of customer-defined aggregation levels within any given consolidation path.

OLAP Operations :

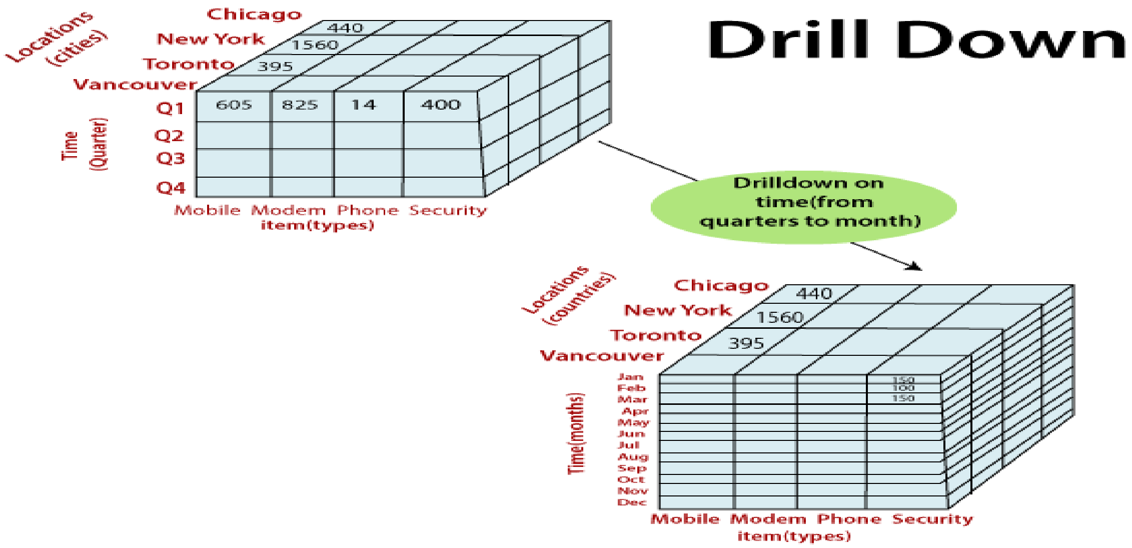
Roll-Up

- The roll-up operation **(also known as drill-up or aggregation operation)** performs aggregation on a data cube, by climbing down concept hierarchies, i.e., dimension reduction. Roll-up is like **zooming-out** on the data cubes.
- Figure shows the result of roll-up operations performed on the dimension location. The hierarchy for the location is defined as the Order Street, city, province, or state, country.
- The roll-up operation aggregates the data by ascending the location hierarchy from the level of the city to the level of the country.
- When a roll-up is performed by dimensions reduction, one or more dimensions are removed from the cube.
- For example, consider a sales data cube having two dimensions, location and time. Roll-up may be performed by removing, the time dimensions, appearing in an aggregation of the total sales by location, relatively than by location and by time.



Drill-Down

- The drill-down operation **(also called roll-down)** is the reverse operation of **roll-up**. Drill-down is like **zooming-in** on the data cube. It navigates from less detailed record to more detailed data. Drill-down can be performed by either **stepping down** a concept hierarchy for a dimension or adding additional dimensions.
- Figure shows a drill-down operation performed on the dimension time by stepping down a concept hierarchy which is defined as day, month, quarter, and year. Drill-down appears by descending the time hierarchy from the level of the quarter to a more detailed level of the month.
- Because a drill-down adds more details to the given data, it can also be performed by adding a new dimension to a cube. For example, a drill-down on the central cubes of the figure can occur by introducing an additional dimension, such as a customer group.



Video Content / Details of website for further learning (if any):

https://www.tutorialspoint.com/dwh/dwh_olap.htm

Important Books/Journals for further learning including the page nos.:
Alex Berson and Stephen J. Smith Data Warehousing, Data Mining & OLAP Tata McGraw Hill Edition 2007 **Pg.No 247-248**

Course Faculty

Verified by HOD

LECTURE HANDOUTS

L 44

MCA

II / III

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **V - Data Warehouse** **Date of Lecture: 14.12.2021**

Topic of Lecture : OLAP versus OLTP

Introduction :

- Logically, OLAP servers present business users with multidimensional data from data warehouses or data marts, without concerns regarding how or where the data are stored.
- However, the physical architecture and implementation of OLAP servers must consider data storage issues.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data warehouse
- OLAP.
- Fact Table
- Dimension Table

Detailed content of the Lecture:

Types of OLAP

There are three main types of OLAP servers are as following:
ROLAP stands for Relational OLAP, an application based on relational DBMSs.
MOLAP stands for Multidimensional OLAP, an application based on multidimensional DBMSs.
HOLAP stands for Hybrid OLAP, an application using both relational and multidimensional techniques.

Relational OLAP (ROLAP) servers:

- These are the intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces.
- ROLAP servers include optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services.
- ROLAP technology tends to have greater scalability than MOLAP technology. The DSS server of Microstrategy,
- for example, adopts the ROLAP approach.

Multidimensional OLAP (MOLAP) servers:

- These servers support multidimensional views of data through array-based multidimensional storage engines.
- They map multidimensional views directly to data cube array structures. The advantage of using a data cube is that it allows fast indexing to precomputed summarized data. Notice that with multidimensional data stores, the storage utilization may be low if the data set is sparse.

- In such cases, sparse matrix compression techniques should be explored Many MOLAP servers adopt a two-level storage representation to handle dense and sparse data sets: denser subcubes are identified and stored as array structures, whereas sparse subcubes employ compression technology for efficient storage utilization.

Hybrid OLAP (HOLAP) servers:

- The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP. For example, a HOLAP server may allow large volumes of detail data to be stored in a relational database, while aggregations are kept in a separate MOLAP store.
- The Microsoft SQL Server 2000 supports a hybrid OLAP server. Specialized SQL servers: To meet the growing demand of OLAP processing in relational databases, some database system vendors implement specialized.
- SQL servers that provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

Online Transaction Processing (OLTP)

- Online transaction processing provides transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization.
- OLAP applies complex queries to large amounts of historical data, aggregated from OLTP databases and other sources, for data mining, analytics, and **business intelligence** projects. In OLAP, the emphasis is on response time to these complex queries.
- OLAP databases and **data warehouses** give analysts and decision-makers the ability to use custom reporting tools to turn data into information. Query failure in OLAP does not interrupt or delay transaction processing for customers, but it can delay or impact the accuracy of business intelligence insights.

Video Content / Details of website for further learning (if any):
https://www.geeksforgeeks.org/difference-between-olap-and-oltp-in-dbms/

Important Books/Journals for further learning including the page nos.:
Alex Berson and Stephen J. Smith Data Warehousing, Data Mining & OLAP Tata McGraw Hill Edition 2007, **Pg.No** : **253-255**

Course Faculty

Verified by HOD

LECTURE HANDOUTS **L 45**

MCA **II / III**

Course Name with Code : **Data Mining And Data Warehousing / 19CAC16**

Course Faculty : **Mr. S.Nithyananth**

Unit : **V - Data Warehouse** **Date of Lecture: 15.12.2021**

Topic of Lecture: Data Warehouse Architecture

Introduction :

- Various kinds of data warehouse design tools are available.
- Data warehouse development tools provide functions to define and edit metadata repository contents (such as schemas, scripts, or rules), answer queries, output reports
- ship metadata to and from relational database system catalogues.
- Planning and analysis tools study the impact of schema changes and of refresh performance when changing refresh rates or time windows

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data warehouse
- Schema
- Metadata
- Meta Rules

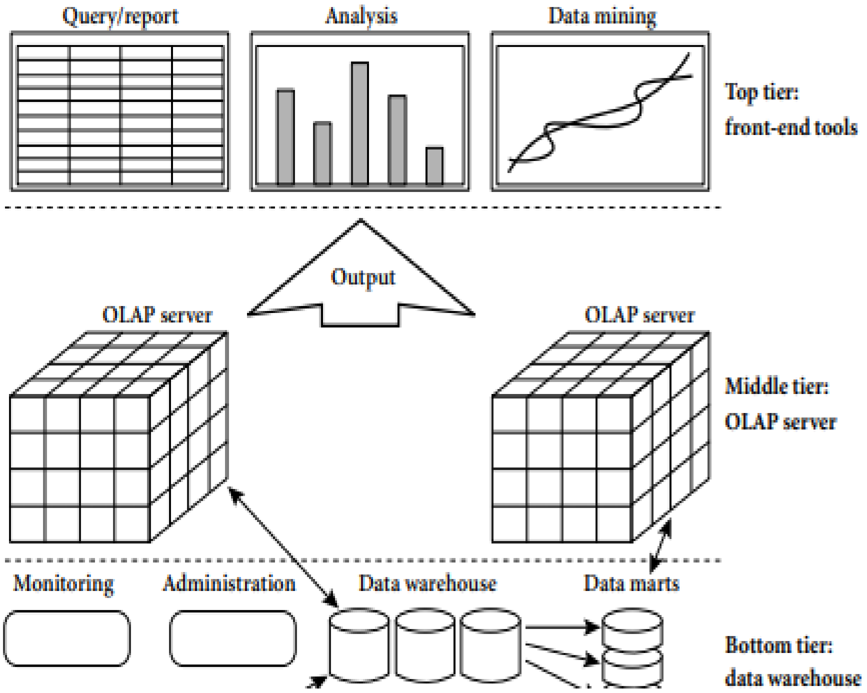
Detailed content of the Lecture:

Four different views regarding the design of a data warehouse must be considered: the top-down view, the data source view, the data warehouse view, and the business query view.

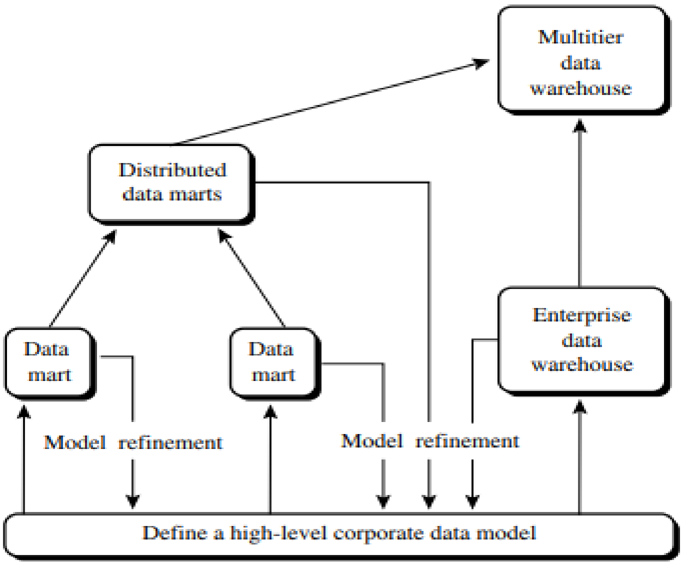
- The top-down view allows the selection of the relevant information necessary for the data warehouse. This information matches the current and future business needs.
- The data source view exposes the information being captured, stored, and managed by operational systems. This information may be documented at various levels of detail and accuracy, from individual data source tables to integrated data source tables.
- Data sources are often modeled by traditional data modeling techniques, such as the entity-relationship model or CASE (computer-aided software engineering) tools. The data warehouse view includes fact tables and dimension tables. It represents the information that is stored inside the data warehouse, including precalculated totals and counts, as well as information regarding the source, date, and time of origin, added to provide historical context.
- Finally, the business query view is the perspective of data in the data warehouse from the viewpoint of the end user

1. The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning, and transformation.

2.The middle tier is an OLAP server that is typically implemented using either (1) a relational OLAP (ROLAP) model, that is, an extended relational DBMS that: An Overview maps operations on multidimensional data to standard relational operations; or (2) a multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.



3. The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on). From the architecture point of view, there are three data warehouse models: the enterprise warehouse, the data mart, and the virtual warehouse.



Video Content / Details of website for further learning (if any):
https://www.astera.com/type/blog/data-warehouse-architecture/

Important Books/Journals for further learning including the page nos.:
Alex Berson and Stephen J. Smith Data Warehousing, Data Mining & OLAP Tata McGraw Hill Edition 2007, **Pg.No 114-117**

Course Faculty

Verified by HOD