Unit - 2
Introduction to Data Mining
(2 Hrs.)

Compiled By: Madan Nath
BSc. CSIT 7th Semester

Contents:

- Motivation for data mining,
- Introduction to data mining system,
- Data mining functionalities,
- KDD,
- Data object and attribute types,
- Statistical description of data,
- Issues and Applications

What is Data Mining?

- Data mining is the efficient discovery of valuable, non obvious information from a large collection of data.
- Knowledge discovery in databases is the nontrivial process of identifying valid novel potentially useful and ultimately understandable patterns in the data.
- It is the automatic discovery of new facts and relationships in data that are like valuable nuggets of business data.
- It is not a complex query where the user already has a suspicion about a relationship in the data and wants to pull all such information.
- The information discovered should give competitive advantage in business.
- Data mining is the induction of understandable models and patterns from a database.

What is Data Mining?

- It is the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions.
- It is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, visualization, and neural networks.
- Data mining is streamlining the transformation of masses of information into meaningful knowledge. It is a process that helps identify new opportunities by finding fundamental truths in apparently random data.

Why We Need Data Mining?

- Data mining is the field where huge amount of data is collected and being processed to extract some useful data i.e information. As you asked what motivates it, the need of era motivates it.
- Everyone wants the concise and precise information which is possible through it, as it is not a easy task but it becomes possible through series of process and science.
- **Some of the fields where data mining is used are:**
  - News Channels
  - In Industries : To know the reviews of people and likes of people.
  - In info-tech companies
  - Research Analysis
  - Healthcare etc.

Data Mining functionalities

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data Mining functions are used to define the trends or correlations contained in data mining activities.
- **Data mining activities can be divided into 2 categories:**
- **Descriptive Data Mining:**
  It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set.
  For examples: count, average etc.
- **Predictive Data Mining:**
  It helps developers to provide unlabeled definitions of attributes. Based on previous tests, the software estimates the characteristics that are absent.
  For example: Judging from the findings of a patient's medical examinations that is he suffering from any particular disease.

Data Mining functionalities

1. Class/Concept Descriptions:

- Classes or definitions can be correlated with results. In simplified, descriptive and yet accurate ways, it can be helpful to define individual groups and concepts.
- These class or concept definitions are referred to as class/concept descriptions.
  - **Data Characterization:**
    This refers to the summary of general characteristics or features of the class that is under the study. For example. To study the characteristics of a software product whose sales increased by 15% two years ago, anyone can collect these type of data related to such products by running SQL queries.
  - **Data Discrimination:**
    It compares common features of class which is under study. The output of this process can be represented in many forms. Eg., bar charts, curves and pie charts.

Data Mining functionalities

2. Mining Frequent Patterns, Associations, and Correlations:

Frequent patterns are nothing but things that are found to be most common in the data.

There are different kinds of frequency that can be observed in the dataset.

- **Frequent item set:**
  This applies to a number of items that can be seen together regularly for eg: milk and sugar.
- **Frequent Subsequence:**
  This refers to the pattern series that often occurs regularly such as purchasing a phone followed by a back cover.
- **Frequent Substructure:**
  It refers to the different kinds of data structures such as trees and graphs that may be combined with the itemset or subsequence.

Association Analysis:

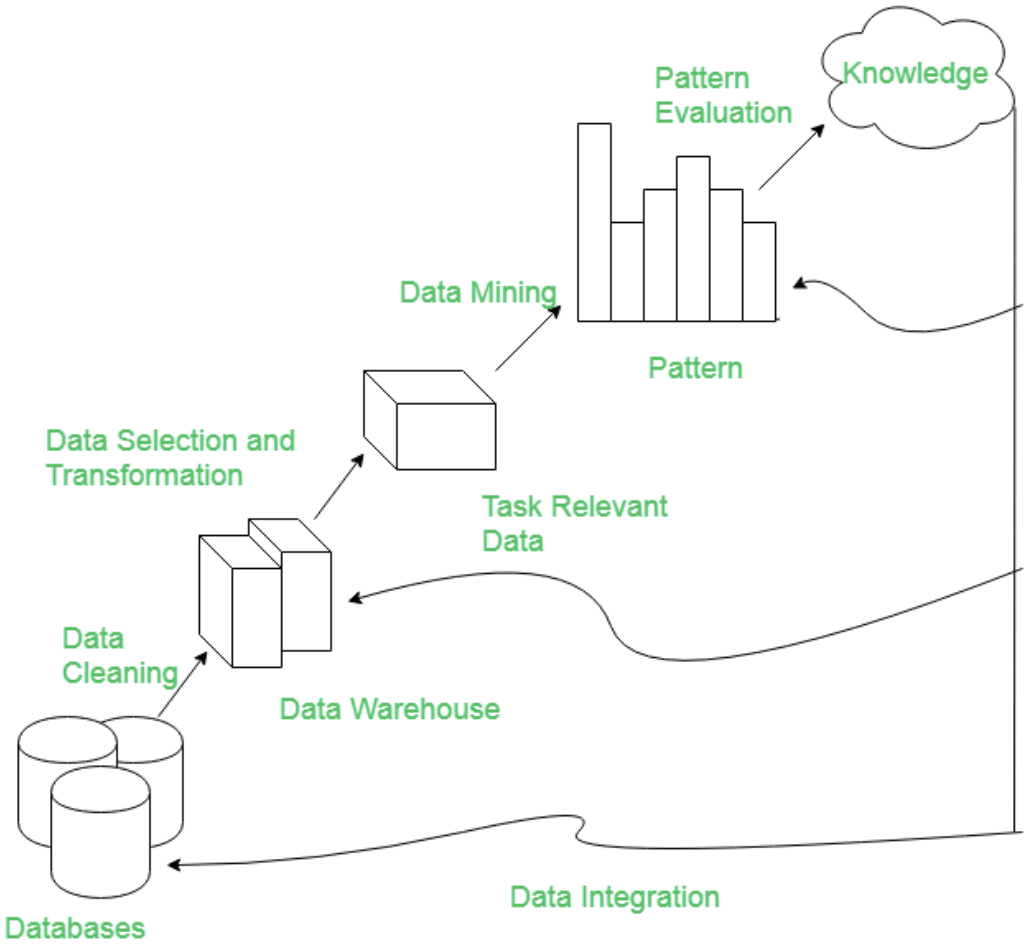
- The process involves uncovering the relationship between data and deciding the rules of the association.
- It is a way of discovering the relationship between various items. for example, it can be used to determine the sales of items that are frequently purchased together.

Correlation Analysis:

- Correlation is a mathematical technique that can show whether and how strongly the pairs of attributes are related to each other. For example, Highted people tend to have more weight.

Knowledge Discovery in Databases(KDD)

- Data Mining also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.



Knowledge Discovery in Databases(KDD)

1. **Data Cleaning:** Data cleaning is defined as removal of noisy and irrelevant data from collection.
   - Cleaning in case of Missing values.
   - Cleaning noisy data, where noise is a random or variance error.
   - Cleaning with Data discrepancy detection and Data transformation tools.
2. **Data Integration:** Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).
   - Data integration using Data Migration tools.
   - Data integration using Data Synchronization tools.
   - Data integration using ETL(Extract-Load-Transformation) process.

Knowledge Discovery in Databases(KDD)

3. Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.

- Data selection using **Neural network.**
- Data selection using **Decision Trees.**
- Data selection using **Naive bayes.**
- Data selection using **Clustering, Regression**, etc.

4. Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data Transformation is a two step process:

- **Data Mapping:** Assigning elements from source base to destination to capture transformations.
- **Code generation:** Creation of the actual transformation program.

Knowledge Discovery in Databases(KDD)

5. Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.

- Transforms task relevant data into patterns.
- Decides purpose of model
  using classification or characterization.

6. Pattern Evaluation: Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.

- Find interestingness score of each pattern.
- Uses summarization and Visualization to make data understandable by user.
  - program.

Knowledge Discovery in Databases(KDD)

7. Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

- Generate reports.
- Generate tables.
- Generate discriminant rules, classification rules, characterization rules, etc.

Data Objects and Attribute Types

- Data sets are made up of data objects. A data object represents an entity
  - in a sales database, the objects may be customers, store items, and sales;
  - in a medical database, the objects may be patients;
  - in a university database, the objects may be students, professors, and courses.
- Data objects are typically described by attributes.
- Data objects can also be referred to as samples, examples, instances, data points, or objects.
- If the data objects are stored in a database, they are data tuples. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.

Data Objects and Attribute Types

- **Attribute:**
  It can be seen as a data field that represents the characteristics or features of a data object. For a customer, object attributes can be customer Id, address, etc. We can say that a set of attributes used to describe a given object are known as attribute vector or feature vector.
- **Type of attributes :**
- This is the First step of Data Data-preprocessing. We differentiate between different types of attributes and then preprocess the data. So here is the description of attribute types.
  1. Qualitative (Nominal (N), Ordinal (O), Binary(B)).
  2. Quantitative (Numeric, Discrete, Continuous)

Qualitative Attributes:

1. Nominal Attributes – related to names: The values of a Nominal attribute are names of things, some kind of symbols. Values of Nominal attributes represents some category or state and that's why nominal attribute also referred as categorical attributes and there is no order (rank, position) among values of the nominal attribute.

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

Qualitative Attributes:

2. Binary Attributes: Binary data has only 2 values/states. For Example yes or no, affected or unaffected, true or false.

- **Symmetric:** Both values are equally important (Gender).
- **Asymmetric:** Both values are not equally important (Result).

Attribute	Values
Gender	Male , Female

Qualitative Attributes:

3. Ordinal Attributes : The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is.

Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

Quantitative Attributes:

1. Numeric: A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, interval, and ratio.

- **An interval-scaled attribute** has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point, or we can call zero points. Data can be added and subtracted at an interval scale but can not be multiplied or divided. Consider an example of temperature in degrees Centigrade. If a day's temperature of one day is twice of the other day we cannot say that one day is twice as hot as another day.
- **A ratio-scaled attribute** is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range, and Five number summary can be given.

Quantitative Attributes:

2. Discrete : Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countably infinite set of values.

Example:

Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

Quantitative Attributes:

3. Continuous: Continuous data have an infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.

Example :

Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 .....etc

Statistical description of data:

- Statistics provides tools for understanding data.
  - In the wrong hands these tools can be dangerous!
- There is a typical data analysis cycle:
  1. Apply some formula to data to compute a "statistic".
  2. Find where value falls in a probability distribution computed on the basis of some "null hypothesis".
  3. If it falls in an unlikely spot (on distribution tail), conclude null hypothesis is false for your data set.
- Statistics and probability theory are closely related. Statistics can never prove things, only disprove them by ruling out hypotheses.
- Distinguish between model-independent statistics (e.g. mean, median, mode) and model dependent statistics (e.g. least-squares fitting).

Statistical description of data:

- The mean, median, and mode of distributions are called measures of central tendency.
- The most common description of data involves its moments, sums of integer powers of the values.
- The most familiar moment is the **mean:** $\bar{x} = \langle x \rangle = \frac{1}{N}\sum_{i=1}^{N} x_i$

NOTE: Why N -1? If the mean is known a priori, i.e. if it's not measured from the data, then use N, else N -1. If this matters to you, then N is probably too small!

- The width of the central value is estimated by its second moment, called the **variance:**

$$\mathrm{Var} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

or its square root, the **standard deviation** $\sigma = \sqrt{\mathrm{Var}}$

Statistical description of data:

- A clever way to minimize round-off error when computing the variance is to use the corrected two-pass algorithm. First compute <x>, then do:

$$\text{Var} = \frac{1}{N-1}\left\{\sum_{i=1}^{N}(x_i-\bar{x})^2-\frac{1}{N}\left[\sum_{i=1}^{N}(x_i-\bar{x})\right]^2\right\}$$

The second sum would be zero if <x> were exact, but otherwise it does a good job of correcting RE in Var.

- Higher moments, like skewness (3 rd moment) and kurtosis (4 th moment) are also sometimes used.

Statistical description of data:

- A distribution function (DF) p(x) gives the probability of finding value between x & x + dx.
  - The expected mean data value: $$\langle x \rangle = \frac{\int_{-\infty}^{\infty} x\, p(x)\, dx}{\int_{-\infty}^{\infty} p(x)\, dx}$$
  - For a discrete DF: $$\langle x \rangle = \frac{\sum_i x_i p_i}{\sum_i p_i}$$

- Similar to weighted means, e.g. center of mass.

Statistical description of data:

Median:

- The median of a DF is the value x med for which larger & smaller values of x are equally probable:

$$\int_{-\infty}^{x_{\text{med}}} p(x)\,dx = \frac{1}{2} = \int_{x_{\text{med}}}^{\infty} p(x)\,dx$$

- For discrete values, sort in ascending order, then:

$$x_{\text{med}} = \begin{cases} x_{(N+1)/2}, & N \text{ odd} \\ \frac{1}{2}\left(x_{N/2} + x_{(N/2)+1}\right), & N \text{ even} \end{cases}$$

Statistical description of data:

Mode:

- The mode of a probability DF p(x) is the value of x where the DF takes on a maximum value.
- Most useful when there is a single, sharp max, in which case it estimates the central value.
- Sometimes a distribution will be bimodal, with two relative maxima.
- In this case the mean and median are not very useful since they give only a "compromise" value between the two peaks.

Statistical description of data:

Comparing Distributions:

Often want to know if two distributions have different means or variances:

1. Student's t-test for significantly different means.

a) Find no. of standard errors $\sim \sigma/N^{1/2}$ between two means.

b) Compute statistic using nasty formula.

c) Small numerical value indicates significant difference.

2. F-test for significantly different variances.

a) Compute $F = Var_1 / Var_2$ and plug into nasty formula.

b) Small value indicates significant difference.

- Given two sets of data, can generalize to a single question: Are the sets drawn from the same DF? Recall can only disprove, not prove.

Statistical description of data:

- May have continuous or binned data.
- May want to compare one data set with known DF, or two unknown data sets with each other. Popular technique for binned data is the **chi ^{2} test**. For continuous data, use the KS test.
- Suppose have N_i events in i^{th} bin but expect n_i :

$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i}$$

  - Large value of **chi^2** indicates unlikely match.
  - Compute probability Q(chi^2|*V*) from incomplete gamma function, where *V* is # degrees of freedom.
- For two binned data sets with events R_i and S_i :

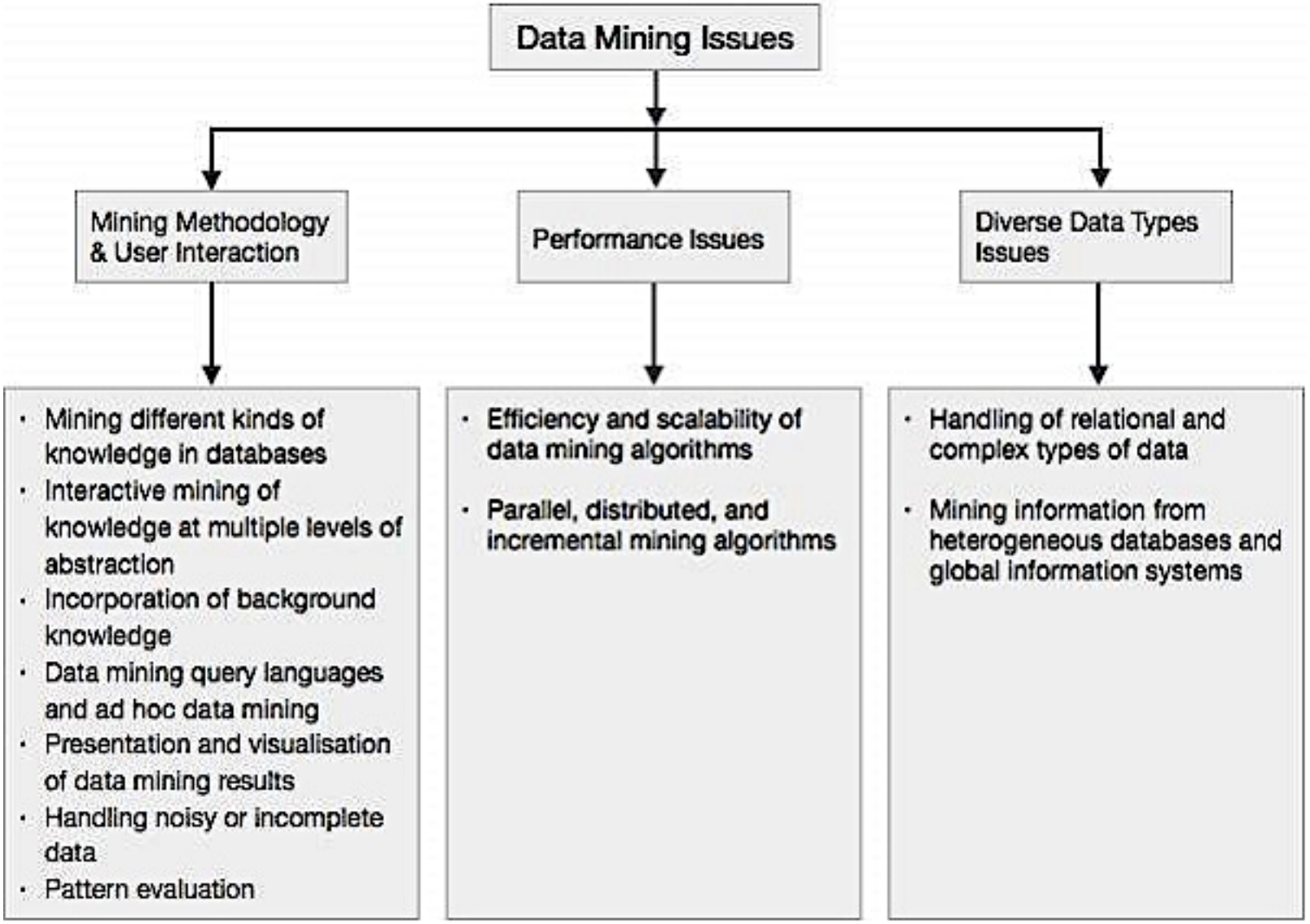
$$\chi^2 = \sum_i \frac{(R_i - S_i)^2}{R_i + S_i}$$

Issues and Applications:

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

- The following diagram describes the major issues.

Issues and Applications:



1. Mining Methodology and User Interaction Issues

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.

1. Mining Methodology and User Interaction Issues

- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

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. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

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. Therefore mining the knowledge from them adds challenges to data mining.

Thank you !!!