Data Mining : Introduction

Chapter 1

Index

1. What is Data Mining

Data mining is the process of discovering interesting patterns (or knowledge) from large amounts of data.

The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

1. What is Data Mining

Communicates between users and data mining system. Visualizes results or perform exploration on data and schemas.

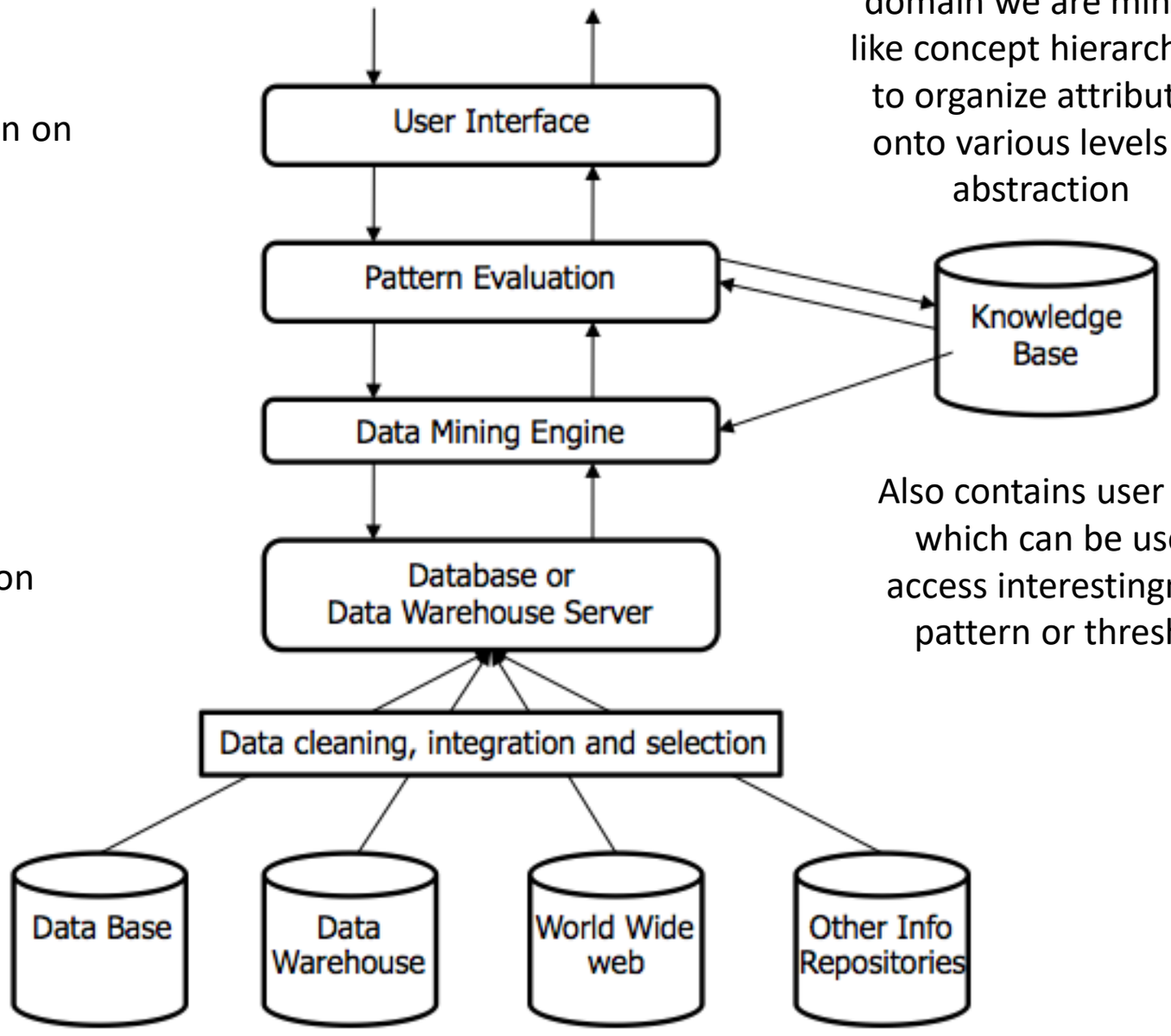
Tests for interestingness of a pattern

Performs functionalities like characterization, association, classification, prediction etc.

Is responsible for fetching relevant data based on user request

This is usually the source of data.
The data may require cleaning and integration.

This is the information of domain we are mining like concept hierarchies, to organize attributes onto various levels of abstraction



Also contains user beliefs, which can be used to access interestingness of pattern or thresholds

Architecture of data mining system

2. Data Mining Functionalities

Data Mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

Data Mining tasks can be classified into two categories

- **Descriptive**: Characterize general properties of data in the database
- **Predictive**: perform inference on data to make predictions

2.1 Data Mining Functionalities: Characterization and Discrimination

Data can be associated with **classes or concepts** that can be described in summarized, concise, and yet precise, terms.

Such descriptions of a concept or class are called **class/concept descriptions.**

These descriptions can be derived via

- Data Characterization
- Data Discrimination

2.1 Data Mining Functionalities: Characterization and Discrimination

Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.

ex: Description of all users who spent more than $10,000 a year at *AllElectronics?* A general profile of all customers, such as age, salary, location and credit ratings. Among all the customers meeting target condition (spent > $10,000), 10% are “Youth”, 60% are “Adults” and 30% are “Seniors”.

The output of data characterization can be presented in pie charts, bar charts, multidimensional data cubes, and multidimensional tables. They can also be presented in rule form.

2.1 Data Mining Functionalities Characterization and Discrimination

Data discrimination is a comparison of the target class data objects against the objects from one or multiple contrasting classes with respect to customers that share specified generalized feature(s).

ex: compare change is sales of software products for customers with given generalized feature: 40% of “Youth” have sales that increased by more 10% from last year; 10% of “Youth” have sales that decreased by at least 30% during the same period; the remaining 50% of “Youth” change in sales the fell in-between. “Youth” describes the generalized tuple, while increase in sales by > 10% is the target class. The other two amounts of change in sales are the contrasting classes.

The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help to distinguish between the target and contrasting classes.

2.2 Data Mining Functionalities: Mining Frequent Patterns

Frequent patterns are the patterns that occur frequently in the data. Patterns can include itemsets, sequences and subsequences.

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

ex: bread and milk

2.2 Data Mining Functionalities: Mining Frequent Patterns

Association Rules

if a customer buys a computer, there is a 50% chance that he will buy software as well

buys(X, “computer”)=>buys(X, “software”) [support =1%, confidence = 50%]

Single Dimension Association Rule

1% of all the transactions under analysis show
that computer and software are purchased together

age(X, “20..29”)^income(X, “40K..49K”)=>buys(X, “laptop”)

Multi-Dimension Association Rule

[support = 2%, confidence = 60%]

Association rules are discarded as uninteresting if they do not satisfy minimum support threshold and minimum confidence threshold

2.3 Data Mining Functionalities: Classification and Prediction

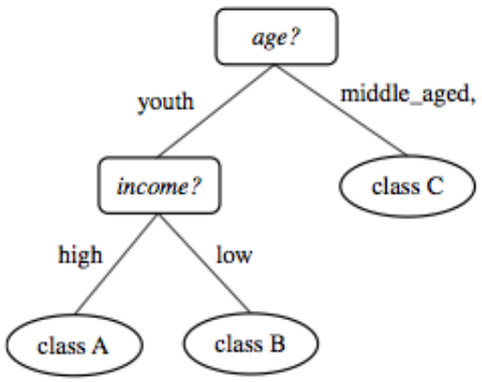
Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of training data and is used to predict the class label of objects for which the the class label is unknown.

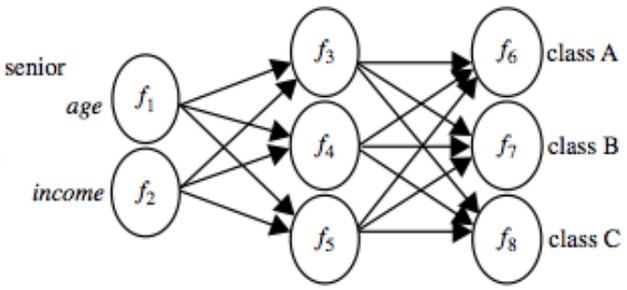
Representation of Derived model

IF-THEN Rules

age(X, "youth") AND income(X, "high") ⟶ *class(X, "A")*
age(X, "youth") AND income(X, "low") ⟶ *class(X, "B")*
age(X, "middle_aged") ⟶ *class(X, "C")*
age(X, "senior") ⟶ *class(X, "C")*

Decision Tree





Neural Network

2.3 Data Mining Functionalities: Classification and Prediction

Prediction values continuous valued functions, i.e. it is used to predict missing or unavailable numeric data values rather than class labels.

Prediction can be used for both numeric prediction and class label prediction.

Regression analysis is a statistical method used numeric prediction.

Classification and regression may need to be preceded by relevance analysis, which attempts to identify attributes that are significantly relevant to the classification and regression process. Such attributes will be selected for the classification and regression process. Other attributes, which are irrelevant, can then be excluded from consideration

2.4 Data Mining Functionalities: Cluster Analysis

Clustering analyzes data objects without consulting class labels.

Clustering can be used to generate class labels for a group of data which did not exist at the beginning.

The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity.

2.5 Data Mining Functionalities: Outlier Analysis

Outliers are data objects that do not comply with the general behavior or model of data.

Many data mining techniques discard outliers or exceptions as noise.

However, in some events these kind of events are more interesting. This analysis of outlier data is referred to as outlier analysis

ex: fraud detection.

2.6 Data Mining Functionalities Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.

This may include characterization, discrimination, association and correlation analysis, classification, prediction or clustering of time related data.

Distinct features of such data include time series data analysis, sequence or periodicity pattern matching and similarity based data analysis.

3. Are all Patterns Interesting?

We need to answer three questions to say if patterns are interesting

1. What makes a pattern interesting?
2. Can a data mining system generate all of the interesting patterns?
3. Can the system generate only the interesting ones?

3. Are all Patterns Interesting?

What makes a pattern is interesting?

Not known before

validates a hypothesis that user sought to confirm

Novel, Potentially useful or desired, understandable and valid

Valid on new set of data with a degree of certainty

Easily understood by humans

3. Are all Patterns Interesting?

Objective measures of interestingness (measurable)

Support: The percentage of transactions from transaction database that the given rule satisfies

$$support(X=>Y) = P(XUY)$$

Confidence: The degree of certainty of given transaction

$$Confidence(X=>Y)=P(Y|X)$$

3. Are all Patterns Interesting?

Many patterns that are interesting by objective standards may represent common sense and, therefore, are actually uninteresting.

So Objective measures are coupled with subjective measures that reflects users needs and interests.

Subjective interestingness measures are based on user beliefs in the data.

These measures find patterns interesting if the patterns are **unexpected** (contradicting user's belief), **actionable** (offer strategic information on which the user can act) or **expected**(confirm a hypothesis)

3. Are all Patterns Interesting?

- **Can a data mining system generate all of the interesting patterns?**

- A data mining algorithm is **complete** if it mines all interesting patterns.

- It is often unrealistic and inefficient for data mining systems to generate all possible patterns. Instead, user-provided constraints and interestingness measures should be used to focus the search.

- For some mining tasks, such as association, this is often sufficient to ensure the completeness of the algorithm.

3. Are all Patterns Interesting?

Can a data mining system generate only interesting patterns?

A data mining algorithm is **consistent** if it mines only interesting patterns. It is an optimization problem.

It is highly desirable for data mining systems to generate only interesting patterns. This would be efficient for users and data mining systems because neither would have to search through the patterns generated to identify the truly interesting ones.

Sufficient progress has been made in this direction, but it still a challenging issue in data mining.

4. Major Issues in Data Mining

1. Mining different kinds of data
2. Handling multiple levels of abstraction
3. Incorporation of background knowledge
4. Visualization of mining results
5. Handling of incomplete or noisy data
6. Scalability of algorithms