Lecture-4

Data Mining

Introduction

(Classification & Major Issues of Data Mining System)

Content

- Are all patterns interesting?
- Classification of data mining systems
- Data mining Primitives
- Major issues in data mining.
- Summary

Are All of the Patterns Interesting?

- A data mining system/query may generate thousands of patterns, not all of them are interesting.

➢ **Interestingness measures**:

- A pattern is interesting if it is
  - Easily understood by humans
  - Valid on new or test data with some degree of certainty
  - Potentially useful
  - Novel, or validates some hypothesis that a user seeks to confirm

Are All of the Patterns Interesting? [Cont..]

➢ **Objective vs. Subjective interestingness measures:**

- Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
- Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, action ability, etc.

Are All of the Patterns Interesting? [Cont..]

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns?
- Search for only interesting patterns: Optimization
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns—mining query optimization

Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views, different classifications
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

Data Mining: Classification Schemes [Cont..]

➢ Classification according to the kinds of databases mined:

- Database systems can be classified according to different criteria (such as data models, or the types of data or applications involved), each of which may require its own data mining technique.

➢ Classification according to the kinds of knowledge mined:

- kinds of knowledge they mine, that is, based on data mining functionalities, such as characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis.

Data Mining: Classification Schemes [Cont..]

➢ Classification according to the kinds of techniques utilized:

- The techniques can be described according to the degree of user interaction involved or the methods of data analysis employed.

➢ Classification according to the applications adapted:

- Data mining systems may be tailored specifically for finance, telecommunications, DNA, stock markets, e-mail, and so on. Different applications often require the integration of application-specific methods.

Data Mining Task Primitives

- A data mining task can be specified in the form of a data mining query, which is input to the data mining system.
- A data mining query is defined in terms of data mining task primitives.
- The data mining primitives specify the following:
  - ➢ The set of task-relevant data to be mined
    - This specifies the portions of the database or the set of data in which the user is interested.
  - ➢ The kind of knowledge to be mined
    - This specifies the data mining functionalities.

Data Mining Task Primitives [Cont..]

- The background knowledge
  - This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found.
- The interestingness measures and thresholds for pattern evaluation
  - They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns.
- The expected representation for visualizing the discovered patterns
  - This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

Major Issues in Data Mining

➢Mining methodology and user interaction

- **Mining different kinds of knowledge in databases**
  - Data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis.

- **Interactive mining of knowledge at multiple levels of abstraction**
  - Because it is difficult to know exactly what can be discovered within a database, the data mining process should be *interactive*.

Major Issues in Data Mining [Cont..]

➢Mining methodology and user interaction [Cont..]

- **Incorporation of background knowledge**
  - Information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction.

- **Data mining query languages and ad-hoc data mining**
  - Relational query languages (such as SQL) allow users to pose ad hoc queries for data retrieval.

Major Issues in Data Mining [Cont..]

➢Mining methodology and user interaction [Cont..]

- **Presentation and visualization of data mining results**
  - Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.
- **Handling noise and incomplete data**
  - The data stored in a database may reflect noise , exceptional cases, or incomplete data objects.

Major Issues in Data Mining [Cont..]

➢ Mining methodology and user interaction [Cont..]

- **Pattern evaluation: the interestingness problem**
  - A data mining system can uncover thousands of patterns.
  - Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack novelty.

Major Issues in Data Mining [Cont..]

➢Performance and scalability

- **Efficiency and scalability of data mining algorithms**
  - To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.
- **Parallel, distributed and incremental mining methods**
  - The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms.

Major Issues in Data Mining [Cont..]

➢Issues relating to the diversity of data types

- **Handling relational and complex types of data**
  - Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important.
- **Mining information from heterogeneous databases and global information systems**
  - Internet connect many sources of data, forming huge, distributed, and heterogeneous databases.
  - The discovery of knowledge from different sources of structured, semistructured, or unstructured data with diverse data semantics poses great challenges to data mining.

Major Issues in Data Mining [Cont..]

➢Issues related to applications and social impacts

- **Application of discovered knowledge**
  - Domain-specific data mining tools
  - Intelligent query answering
  - Process control and decision making
- Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
- Protection of data security, integrity, and privacy

Summary

- A pattern represents knowledge if it is easily understood by humans.
- Measures of pattern interestingness, either objective or subjective, can be used to guide the discovery process.
- Data mining systems can be classified according to the kinds of databases mined, the kinds of knowledge mined, the techniques used, or the applications adapted.
- We have studied five primitives for specifying a data mining task in the form of a data mining query.
- Efficient and effective data mining in large databases poses numerous requirements and great challenges to researchers and developers.
- The issues involved include data mining methodology, user interaction, performance and scalability, and the processing of a large variety of data types.

Thank You