Lecture 28

Data Mining and KDD Presentation (1 of 4)
KDD for Science Data Analysis

Wednesday, March 29, 2000

Arul Elumalai
Department of Computing and Information Sciences, KSU
arul@cis.ksu.edu

Article of the day
"KDD for Science Data Analysis: Issues and Examples"
- Fayyad, Haussler and Stolorz

Presentation preview

- Introduction
- Concept of data mining
- Fundamentals of data analysis
- Case studies
- Issues and Challenges
- Article critique

Introduction

Objective:
Application of KDD in creative data analysis for theory formation

Scope:
Analysis of scientific data

Scenario:
- Modern scientific instruments & data collection
- Data abundance

Issues:
- Gap between data collection and data analysis
- Large size and dimension of available data

Handling massive data

Data Reduction
Reducing data to an analyzable size and simplicity
Questions:
1. Is it representative of the complete phenomenon?
2. Is it only the redundant data that has been removed?
3. What strategy is to be deployed for data reduction?

Automated Analysis:
Mechanization of data analysis using intelligent agents.
Question:
1. Is it as efficient and foolproof as manual analysis?

Data in its many forms

- **Image Data**
  + Predefined display format
  - Mining image data is difficult
  - Mapping from pixel to feature is noisy
- **Time series and sequential data**
  - Rate of measurement may be random
  - Non stationary characteristics
- **Numerical Vs Categorical measurement**
  - The concept of "difference" is not defined (CM)

Data in its many forms (contd.)

- **Structured and sparse data**
  - Measured attributes may vary
  - Dimensionally complex (No available algorithms)
- **Reliability of data (sensor Vs model)**
  - Needs translation from sensor level

KDD and Data mining

Knowledge Discovery in Databases (KDD) refers to the overall process of discovering useful knowledge from data

Data mining refers to the application of algorithms for extracting patterns from data

Data mining consists of five major elements:

- **Extraction**
- **Storage**
- **Access/ retrieval**
- **Analysis (by application software.)**
- **Presentation.**

Case 1: Stellar classification

Problem statement:
Classification of stellar objects based on image data
Scale:
3 terra bytes of data; 2 billion objects; 3000 images; 40 attribs/ obj
Strategy deployed:
- Dimensional reduction (40 to 8)
- Tree learning algorithm

Achievement:
- Accommodated fainter images
- Achieved 94% prediction accuracy

Limitation:
- Inclusion of supervised learning

Case 2: Volcanoes in Venus

Problem statement:
Finding volcanoes on Venus using high resolution global maps
Scale:
30,000 images; 100 CD storage
Strategy deployed:
- Training via examples

Achievement:
- Detection of over 1 million volcanoes
- Flexible approach and allows reuse

Limitation:
- High false detection rate
- Sensitive to image illumination, scale and angle

Case 3: Extraction of genetic code

Problem statement:
Extraction of genetic code from stored values in databases
Scale:
400 million tokens (GENBANK); 200,000 sequences;
Inaccurate pattern finding algorithms;
Strategy deployed:
- Identification of a statistical model (HMM)
- Template structure for search is not provided and must be discovered

Achievement:
- Identification of new relations

Limitation:
- Slow database query and computational overheads
- Prerequisite bio knowledge and lab experimentation

Case 4: Earth Geophysics

Problem statement:
Measurement of tectonic motion based on images before and after quakes
Scale:
Lack of precision in resolution;
Strategy deployed:
- Repeated registration of local images to sub pixel precision
- Construction of systems that can work on massive data sets

Achievement:
- Not only measured known faults but also detected novel patterns

Limitation:
- Required “similar enough” images for comparison

Case 5: Atmospheric science

Problem statement:
Analysis of weather patterns and other spatio-temporal patterns
Scale:
Several gigabytes of data/ model; Complex queries; Large attributes
Strategy deployed:
- Use of parallel test beds
- Development of learning algorithms that identified novel patterns
- Content based indexing to increase query performance

Status quo:
- State of infancy

Issues and Challenges

- Feature extraction from raw data
- Attention to minority classes
- Demand of high degree of confidence and accuracy
- Basis for selection of data mining task
- Translation of derived models into useful knowledge
- Harnessing domain knowledge
- Scalable machine knowledge and algorithms

Content critique

Strengths:

- Good survey paper
- Elucidates practical application of KDD
- Gives an idea of the relevance of KDD to data analysis
- A valuable "documentation of experiences"

Weaknesses:

- More a BOK; not many findings
- Harps more on problems than solutions
- Haven't explicitly mentioned reasons for success

Presentation critique

+ Case based elucidation

+ Magnitude of the problem has been clearly mentioned

- Could have laid more emphasis on issues and challenges

- Does not give solutions to the problems encountered