Operations Research and Statistics Techniques: A Key to Quantitative Data Mining

Jorge Luis Romeu
IIT Research Institute, Rome NY
FCSM Conference, November 2001

Outline

- Introduction and Motivation
  - Why Data Mining/Knowledge Discovery now?
- Phases in a DM/KDD study
  - What are their intrinsic components?
  - How new (in what way) are they?
- Computer and Other Considerations
- Summary, Conclusions and Bibliography

Introduction

- Data, collection methods and computers
- The advent of Internet and its implications
- Data base explosion: start of data-profiles
- Traditional v. New data analysis paradigm
- Characterizing Data Mining (DM) and Knowledge Discovery in Databases (KDD)
- Quantitative (v. Qualitative) Data Mining

Motivation

- DM/KDD is a fast-growing activity
  - In dire need of good people (analysts)
- A Special Data Mining Characteristic:
  - research hypotheses and relationships between data variables are both obtained as a result
- Statistics and operations research areas
  - well-suited for data mining activities
- Paper objective: to provide a targeted review
  - Alert Stats/OR and Explain it to Others Players.

Phases in a DM/KDD study

- I) Determination of Objectives
- II) Preparation of the Data
- III) Mining the Data (***)
- IV) Analysis of Results
- V) Assimilation of the Knowledge Extracted from the Data

Determination of Objectives

- Having a clear problem statement includes:
  - review/validation of the basic information
  - re-statement of goals and objectives
  - technical context, to avoid ambiguity
  - gather and review background literature
    - about data, problem, component definitions
  - prepare comprehensive/detailed project plan
  - obtain a formal agreement from our “client”.

Preparation of the Data

- Most time-consuming phase (60%)
- Is divided into three subtasks:
- 1) Selection/collection of the Data
  - define, understand, identify, measure variables
  - data base (storage/retrieval) design issues
- 2) Data pre-processing task
  - ensuring the quality of the collected data
- 3) Data transformation subtask

Illustrative Example

- Internet Data Collection Project
  - Objective: forecast (specialized) Web usage
- Problem Components: Internet and user
- Indicators characterizing and relating them:
  - Hits, page requests, page views, downloads;
  - Dial-ups, unique-visitors, permanent connections
  - Internet subscribers: who, why, when, how, etc.
  - Web site (internal) movements (pages visited)
  - Traffic capacity, speed, rate, bandwidth

Mining the Data

- Traditional stats data analysis core
  - standard objectives: study, classify or predict
- Data mining techniques, into five classes, according to its objectives (Bradley et al):
  - Predictive modeling,
  - Clustering or segmentation
  - Dependency modeling,
  - Data summarization
  - Change and deviation detection

Regrouping the Approaches

- Overlap among Bradley's existing classes
  - not a partition of the existing methodology
- Define three (methodological) categories:
  - mathematically based procedures
  - statistically based procedures
  - and "mixed" algorithms
- A method can be used for multiple objectives
  - and we want to emphasize methods over uses.

Mathematically Based Algorithms

- *Mathematical programming*
  - linear, non-linear, integer programming
- *Network analyses/affinity analysis*
  - data flow is represented within a network
- *Memory-based reasoning*
  - nearest neighbors classified by their distances
  - form subsets of “similar” elements (neighbors)

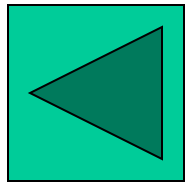
Some technical problems

- *Mathematical programming*
  - defining objective function and constraints
- *Network analyses/affinity analysis*
  - complex relations are difficult to represent
- *Memory-based reasoning*
  - definition/combination of metrics is difficult
  - pre-established subsets of “similar” neighbors
  - the selection of the “training set”

Mixed Algorithms

- *Neural networks*
  - mimics the way the brain is configured/works
- *Genetic algorithms*
  - mimics biological genetics in human evolution
- *Decision trees*
  - mirror image of neural nets (top-down)
- *Clustering methods*
  - for dividing population into similar groups

Technical Problems

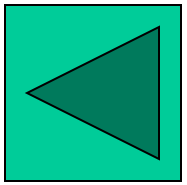


- *Neural networks*
  - select training data and subjective decisions
- *Genetic algorithms*
  - fitness function, defined by the model builder
- *Decision trees*
  - Definition of decision probabilities, training set
- *Clustering methods*
  - definition of the metric and of the number and components of each subgroup

Statistically Based Algorithms

- *Regression*
  - variables that better explain the relationship
- *Discrimination*
  - variables that better explain group differences
- *Time series*
  - black-box type of approach, no regressors
- *Factor analysis*
  - interest in detecting inter-variable associations

Technical Problems



- *Regression*
  - data base, variable selection/reduction methods
- *Discrimination*
  - partition into a pre-specified number of groups
- *Time series*
  - type (no. of parameters) of the ARIMA model
- *Factor analysis*
  - number and interpretation of resulting factors

Analysis of Results

- DM/KDD as an enhanced form of EDA (?)
- DM/KDD problem results are iterative:
  - are (*) inputs, used for a new iteration, or
  - are stage-final -but until new data is available!
- DM/KDD is a team effort (!!!)
  - one of its strongest assets and benefits
  - members, work together and jointly interpret
  - some may not be satisfied or would like to enhance or to experiment with other variants (*)

Assimilation of the Knowledge Extracted from the Data

- Main DM/KDD objective: problem solving
- By analyzing system data, we can obtain information to help resolve the problems
- Need to convert such information into adequate courses of action (solution)
- Define what changes would be necessary
- Assessment triggers next iteration phase

Computer/Other Considerations

- The size of the problem (database)
- The nature of the analysis (algorithms)
- Makes DM/KDD a computer-based process
  - requires combination of both theory and practice
- Needs software/high-performance computers
- More than mere statistical and mathematical
- Includes other important functions/algorithms
  - especially important are the “expert functions “

Summary of our Work

- Alerted statisticians and O.R. professionals
  - about DM/KDD and how they can partake in it
- Overviewed main stats and O.R. techniques
  - and the DM/KDD activities and paradigm
- Examples of uses, problems and limitations
  - in several methods and paradigm phases
- Discussed work statisticians and O.R. do
  - for other team members from different fields

Bibliography

Balasubramanian et al. Data Mining: Critical Review and Technology Assessment Report. IATAC/DTIC. 2000

Special Issue on Data Mining. INFORMS Journal on Computing. Vol. 11, Number 3. 1999.

Data Mining and Knowledge Discovery; On-line journal: http://www.digimine.com/usama/datamine/

Anderson, T. W. An Introduction to Multivariate Statistical Analysis. Wiley, NY. 1984.

Taha, H. A. Operations Research; an introduction. Prentice Hall. New Jersey. 1997