Block II: Data preprocessing and analysis

Readings Assignment

Introduction

In order to gain a general level of knowledge on this subject,

- a set of readings assignments (articles), together with
- a set of topics (field of questions)

will be assigned.

The students have to learn individually about the topics covered. The basic readings are assigned, additional search of information is however expected! The knowledge will be evaluated in a **form of a test.**

Readings assignment

(This is just a literature to begin with)

1. Fair Isaac White Paper. "*A discussion of Data Analysis, Prediction and Decision Techniques*", May 2003. (Selected chapters)
2. C. W. Kirkwood, Decision Tree Primer, 2002.
3. T. Bellemans, B. De Schutter, and B. De Moor, "Data acquisition, interfacing and pre-processing of highway traffic data," Proceedings of Telematics Automotive
4. Stephen K. Lower "Matter and measure" Simon Fraser University (http://www.sfu.ca/person/lower/TUTORIALS/matmeas/)
5. Linh N. Nguyen. Dr. William T. Scherer. "Imputation Techniques to Account for Missing Data in Support of Intelligent Transportation Systems Applications". Research Report No. UVACTS-13-0-78. May 2003.

… more at www.scitopia.org , www.ieeexplore.com etc.

Topics and sample questions

General

- Data types (categorical, numerical, ordinal, …)
- Measurement error (random, systematic, absolute, relative, multiplicative, additive, …), accuracy and precision (definition, explanation, source of errors, …)

Data Preprocessing

- Major steps in data preprocessing
  - What is data preprocessing? Why is data preprocessing important? How is it used in transportation? What is the effect of not using data preprocessing?
- Data imputation techniques - Data estimation techniques: Time-of day historical average, regression upon neighbouring detectors, ARIMA and STARMA models (equations, explanation, usage)
- Outliers/Corrupt or missing data
- Data reduction
  - What is data reduction? Why do we want to reduce data? Transportation examples? Approaches to data reduction?

Time Series Analysis

- What is a time series?
- Decomposition of time series?
- Filtering of time series: Moving average, exponential smoothing (equations, explanation, usage)?
- Autocorrelation function (equations, explanation, usage)?
- ARIMA model (equations, explanation, usage)?

Decision trees

- What is it and how does it work?
- Types of DTs
- How to create a DT? what are the imputation criteria?
- Advantages, disadvantages?

Cluster analysis

- Why cluster analysis?
- What is it and how does it work?
- Types of algorithms in cluster analysis (hierarchical/optimization)?
  - k-means, k-medoids, …
- Limitations of particular methods?
- How to determine an optimal number of clusters?
- What is a dendrogram
- Usage of cluster analysis in transportation