Data Mining Workshop for ISRO Scientists
December 21, 22 and 23, 2002
National Remote Sensing Organization, Hyderabad

Speaker Biography:

Prof. Iyengar is an ACM Fellow, IEEE Fellow, AAAS Fellow and Member of the New York Academy of Sciences, Williams Evans Fellow, Satish Dhawan Chaired Professor at Indian Institute of Science, SIAM Distinguished Visitor, IEEE Distinguished Scientist, IEEE Meritorious Award, IEEE Golden core member, author/coauthor/editor of 12 textbooks published by CRC press, John Wiley, & Prentice Hall, etc. He has also published over 350 research papers. Funded Research Projects from NSF, NASA DARPA, ONR, DOE, etc., and Member of the editorial board of several IEEE Journals.

http://www.csc.lsu.edu/~Iyengar/index.html
Essence:

Data mining is “the data-driven discovery and modeling of hidden patterns in large volumes of data.” Data mining differs from the retrospective technologies because it produces models that capture and represent the hidden patterns in the data. Via data mining, a user can discover patterns and build models automatically, without knowing exactly what the person looking for. The models are both descriptive and prospective. It is the science of storing, extracting, organizing, analyzing, interpreting, and utilizing information from large data sets.

Data Mining is the discovery of patterns, associations, anomalies, and statistically significant structures in data. It is a multi-disciplinary field, borrowing and enhancing ideas from diverse areas such as statistics, signal and image processing, image understanding, mathematical optimization, computer vision, and pattern recognition. It is the process of non-trivial extraction of implicit, previously unknown and potentially useful information from voluminous data. Advances in various fields like DNA Sequencing, Bioinformatics, Ecommerce, Fraud Detection, Knowledge Management, Remote Sensing Images, GIS, digital cartography, Stock Investment and Prediction Analysis, Real-Time Decisioning etc have fueled it.

Knowledge Discovery and Data Mining (KDD) techniques will play an increasingly important role in the analysis and discovery of sequence, structure and functional patterns or models from large sequence databases. Challenges like collecting Tera bytes of radar information and virtually reconstructing a 3-D model of the continental area, focusing on data from astronomical surveys to include spatial data (two- and three-dimensional, with multivariate fields), spatio-temporal data, data in the form of hierarchical structures, simulated vs. observed data, grid vs. “object” data, multi-spectral and multi-resolution data from multiple sensors etc, data registration and alignment techniques to fuse multi-sensor, multi-resolution, multi-spectral data can be met using KDD techniques. Creating distributed, heterogeneous knowledge networks is one of the greatest current science and engineering challenges. There is a pressing need for advanced approaches to combine, classify, and analyze geospatial data in decision support.

Spatial data mining, i.e., mining knowledge from large amounts of spatial data, is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing, to geographical information systems (GIS), computer cartography, environmental assessment and planning, etc. Spatial data mining is to mine high-level spatial information and knowledge from large spatial databases. Spatial data mining, or knowledge discovery in spatial database, refers to the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases. The huge amount of climate data acquired through satellites, terrestrial observations, and ecosystem models offers an unprecedented opportunity for predicting and preventing future ecological problems and such data include various atmospheric, land, and ocean variables. Due to the nature and scale of this data, data mining techniques can play a major role in the automatic extraction and analysis of interesting patterns, thus complementing existing statistical techniques.
With sensors becoming ubiquitous, computers simulating complex processes at an unprecedented pace, and ever-improving data storage capabilities, petabyte-scale datasets are becoming routine. This has led to the innovative application of data mining techniques to novel and challenging problems. All this finally boils down how to identify and exploit structure and patterns in the vast swirl of information we collect, for the sake of new insights and scientific discovery.

**PART A:**

**Discussion of Topics**

**December 21 9:00 AM – 5:00 PM (With some interruptions for lunch and tea)**

**Introduction**
- What is data mining?
- Drivers for data mining
- Data mining process
- Data mining standards
- Data mining functions
- Data mining techniques
- Data mining applications

**Data preprocessing**
- Data integration
- Data quality audit
- Data cleaning
- Data transformation
- Data sampling

**Classic Data mining techniques**
- Clustering
- Association rule mining
- Decision trees, rules and tables
- Artificial Neural Networks
- Case based reasoning

**Model validation**
- Criteria for evaluating the model
- Model scoring
- Bagging and boosting
- Comprehension
- Compactness

**December 22 9:00 AM – 5:00 PM (With some interruptions for lunch and tea)**

**Advanced data mining techniques – Fuzzy data mining**
- Fuzzy classifier
- Fuzzy predictive modeler
- Fuzzy rule induction
- Fuzzy k-nearest neighbor
Data mining in complex domains
- Mining time series data
- Web log mining
- Web content mining
- Mining spatial data
- Mining the multimedia data

Visual data mining
- Interactive data mining
- Independence diagram
- Stellar plot
- Pixel bar chart
- Radial co-ordinate plot
- Parallel co-ordinate plot

December 23 9:00 AM – 5:00 PM (With some interruptions for lunch and tea)

PART B:

Hands on experience with a few software projects using the following software:

Software/Tools/Resources:

PolyAnalyst V 4.5, is the most complete and versatile suite of unique knowledge discovery algorithms in the world. The system provides simple connectivity with data storage architectures, an easy-to-use visual workspace, and a broad range of machine learning algorithms.

CART® Decision Tree Software is a robust, easy to use decision tree tool that automatically sifts large, complex databases, searching for and isolating significant patterns and relationships. The discovered knowledge is then used to generate reliable, easy to grasp predictive models.

ENVI, the Environment for Visualizing Images, is one of the most advanced, yet easy-to-use remote sensing software available. ENVI will transform the way you work with remote sensing data. It includes hyperspectral image analysis, sensor specific radiometric and geometric corrections, extensive image and vector format support, interactive image enhancement and classification etc.

The Grid Analysis and Display System (GrADS) is an interactive desktop tool that is used for easy access, manipulation, and visualization of earth science data. The format of the data may be either binary, GRIB, NetCDF, or HDF-SDS (Scientific Data Sets). GrADS has been implemented worldwide on a variety of commonly used operating systems and is freely distributed over the Internet.

The Distributed Oceanographic Data System (DODS), developed by oceanographers at the University of Rhode Island and MIT, makes remote scientific data accessible through familiar
data analysis and visualization packages (and APIs). DODS uses client/server architecture with DODS servers providing access to collections of data.

**The Man computer Interactive Data Access System (McIDAS)**, developed by the University of Wisconsin's Space Science and Engineering Center--emphasizing comprehensive capabilities for image processing of data from all satellite-borne sensor systems.

**Motif System for Processing and Handling Images uNder X (Mspinx)**, developed at the Laboratoire d'Optique Atmospherique (LOA) is a flexible, simple-to-use image analysis and presentation package for climate and satellite remote sensing research.

**VisiMine- Image Mining Software** is a search engine for analyzing image databases and is designed for satellite imagery and aerial photos. Visimine provides a comprehensive workbench for image information mining, integrating state–of–the–art statistics, datamining and image processing to extract information and locate images from potentially huge databases. VisiMine can analyze large collections of remotely sensed images as a whole.

**Data Mining Software-EDM and DMSK** is a collection of routines for efficient mining of big data. It includes data preparation, feature reduction and selection, value reduction and smoothing, case reduction and sampling, prediction methods-classification and regression.