Scientific Data Mining and Analysis

Chandrika Kamath
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
March 17, 2004

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.

http://www.llnl.gov/casc/sapphire
Data mining terminology

- **Data mining**: the semi-automatic discovery of patterns, associations, anomalies, and statistically significant structures in data

- **Pattern recognition**: the discovery and characterization of patterns

- **Pattern**: an ordering with an underlying structure

- **Feature**: extractable measurement or attribute

Pattern: radio galaxy with a bent-double morphology
Features: number of “blobs”
maximum intensity in a blob
spatial relationship between blobs (distances and angles)

FIRST: sundog.stsci.edu
Large-scale data mining - from a Terabyte to a Megabyte

An iterative and interactive process
First, we need to handle massive, multi-resolution data from different sensors

- Different data formats and output types
- Size of the data
  - sampling
  - multi-resolution techniques
- Data from different sensors at different resolutions at different times
  - data registration

Example: Images of the Crab Nebula (chandra.harvard.edu)
Next, we need to find the objects in image and mesh data

- Data can be noisy, with missing values
  - denoising can smooth the data

- Image processing techniques to identify the object
  - extensive variation across objects and images
  - algorithms have several parameters
  - must be robust across images
  - extensible to 3D meshes and unstructured grids

Example: the result of image segmentation
Once an object has been identified, we need to extract features to represent it:

- Scale, rotation, and translation invariant
- Robust: insensitive to small changes in data
- Appropriate to the problem
  - involve domain scientists
  - extract many and then reduce

Example: Position angle is not a robust feature
We may need to reduce the number of features through feature selection.

- May need to normalize features
- Number of features to keep
- Involve domain scientists
- Need to interpret the results
- Techniques may or may not depend on the pattern recognition algorithm used

\[ f_1, f_2, \ldots, f_n \rightarrow f_1', f_2', \ldots, f_p' \]

\[ p < n \]
There are several challenges in applying pattern recognition to scientific data

- **Classification**: given a training set, build a “model” that can assign a label to an unclassified object
  - training sets are often small
  - unbalanced
  - contain subjective errors

- **Clustering**: group the data items based on their similarity
  - number of groups
  - quality of clustering
  - effects of outliers
Other issues make scientific data mining challenging

- **Visualization and validation of results**
  - use to create a larger training set; subject to labeling “drift”
  - info-viz: one-off solution; problems with large datasets
- **Parallelism**: size of the data shrinks as it is processed
- **Real time responses**: streaming data
- **Spatio-temporal data**
- “Interesting” vs. “labeled” data
- **Lack of ground truth**
- **Mining a growing data set**