On the Role of Data Mining Techniques in Uncertainty Quantification

Chandrika Kamath
Lawrence Livermore National Laboratory
Livermore, CA, USA
kamath2@llnl.gov

USA/South America Symposium on Stochastic Modeling and Uncertainty Quantification
August 2011
Outline

1. Understanding data mining: A sample problem

2. Overview of the data mining process

3. Data mining techniques in uncertainty quantification
   - Preprocessing data
   - Identifying objects
   - Extracting features
   - Reducing the dimension
   - Building the models
   - Generating the data

4. Closing thoughts
Understanding scientific data mining using a simple example
Understanding data mining: A sample problem

Classification of orbits in Poincaré plots

Task: Given a training set of example orbits, with their associated classes, build a predictive model which, given an orbit, will assign it a class.

Quasiperiodic  Island chain  Separatrix  Stochastic
First, we need to identify a training set - this is challenging

- The labeling is tedious, subjective, and error-prone.
- There is variation within orbits of a class.
- The orbit may change class as more points are added.
- The orbits exhibit multi-scale and fractal behavior.
- The data are noisy.
Next, we need to find a suitable representation for the orbit

**Given:** the \((x, y)\) coordinates of the points in an orbit

**Want:** to differentiate among the orbits based on their “shape”

⇒ we need to find an appropriate representation of the orbit

**Objects:** the parts of the data which are of interest (e.g., an orbit)

**Features:** low level measurements representing the objects (e.g., \#islands)

Good features

- are representative
- can be used to differentiate between classes
- are invariant to scale, rotation, and translation
- are robust (insensitive to small changes in data)
We extracted several features for the orbits

- Convert the data to polar coordinates
- Consider the points in a window around each point; scale so $\delta r = 1.0$
- Extract simple features: concentration of points, error in fitting a second order polynomial, quadrat distribution of points, ...
- Extract derived features: number of times the peaks/valleys in concentration and error are aligned, ...

Appropriate features are essential to building an accurate predictive model.
The original raw data is now an object x feature matrix

Given the orbits represented as points in d-dimensional feature space, ...

<table>
<thead>
<tr>
<th></th>
<th>(f_1)</th>
<th>(f_2)</th>
<th>(\ldots)</th>
<th>(f_d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(O_1)</td>
<td>(f_{11})</td>
<td>(f_{12})</td>
<td>(\ldots)</td>
<td>(f_{1d})</td>
</tr>
<tr>
<td>(O_2)</td>
<td>(f_{21})</td>
<td>(f_{22})</td>
<td>(\ldots)</td>
<td>(f_{2d})</td>
</tr>
<tr>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td>(O_n)</td>
<td>(f_{n1})</td>
<td>(f_{n2})</td>
<td>(\ldots)</td>
<td>(f_{nd})</td>
</tr>
</tbody>
</table>

... we want to build a predictive model to separate the classes.

But first, it is helpful to take a closer look at the matrix

- Some features may be correlated
- Others may be irrelevant to the class label
- The samples required to build an accurate model grow exponentially with number of features
We can use visualization tools to explore the matrix, ...

Parallel plot: $f_7$, $f_{10}$, and $f_{15}$ are discriminatory  

Scatter plot of feature $f_{10}$ vs $f_7$

Simple visual tools are invaluable in

- identifying outliers and correlated variables
- evaluating the quality of features

But unsuitable if the data sets are very large or high dimensional
... or dimension reduction to identify key features

<table>
<thead>
<tr>
<th>distance filter</th>
<th>t-statistic filter</th>
<th>chi-squared filter</th>
<th>stump filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>f3</td>
<td>f8</td>
<td>f8</td>
<td>f18</td>
</tr>
<tr>
<td>f8</td>
<td>f3</td>
<td>f7</td>
<td>f3</td>
</tr>
<tr>
<td>f5</td>
<td>f5</td>
<td>f3</td>
<td>f6</td>
</tr>
<tr>
<td>f6</td>
<td>f6</td>
<td>f5</td>
<td>f5</td>
</tr>
<tr>
<td>f7</td>
<td>f7</td>
<td>f6</td>
<td>f8</td>
</tr>
<tr>
<td>f18</td>
<td>f16</td>
<td>f18</td>
<td>f7</td>
</tr>
<tr>
<td>f16</td>
<td>f4</td>
<td>f16</td>
<td>f4</td>
</tr>
</tbody>
</table>

Features selected as important relate to
- errors in fitting polynomial to points in a window (f3, f5, f6)
- the spread of points in a window (f7, f8)
- the trend in error vs. concentration (f16)
- gaps in the orbit larger than 10 degrees (f18, a binary variable)
We can now use the data to build a predictive model.

**Decision tree**

1884 examples:
- 778 quasiperiodic
- 440 island chains
- 416 stochastic
- 250 separatrix

Splitting criterion:
- **Gini**: split which most reduces node impurity
- **Information gain**: split which most reduces entropy

To assign a label to an orbit, extract the features and apply the model.
Overview of the scientific data mining process
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 underlying structure
- **Feature**: an extractable measurement or attribute

Data mining borrows ideas from many fields, including machine learning, statistics, pattern recognition, signal/image/video processing, linear algebra, mathematical optimization, high-performance computing, ...
The steps in mining science data are driven by the raw data.

An iterative and interactive process.
Observations on the data mining process

- Data mining is an iterative and interactive process.
- It requires close collaboration with the domain scientists.
- The pre-processing step is critical; it is problem-dependent, and often the most time consuming part.
- Software can seldom be used as a black-box.
- Not all tasks are used in all problems.
- Some algorithms can be used in more than one task.
Data mining techniques in uncertainty quantification
Preprocessing data to improve their quality
Experimental data often require cleanup before analysis

- Denoising: filters (mean, Gaussian), model-based methods, PDEs, ...
- Contrast enhancement: histogram equalization, Retinex, ...

Comparing simulations to experiments for Richtmyer-Meshkov instability

Extracting statistics from images of materials fragmentation
Identifying objects of interest in the data
We can focus on the boundaries of the objects, ...

- Gradient-based methods: simple filters, Canny edge detector, SUSAN, ...
- PDE-based methods: level sets

Comparing simulations to experiments in Richtmyer-Meshkov instability
... or, we can focus on the interior of the objects, ...

- Region growing methods, watershed approaches, ...
- PDE-based methods: active contours without edges

Finding the bubble surface in 3-D

Identifying the bubbles in 2-D

Identifying bubbles and spikes in Rayleigh-Taylor instability
... or, we can use domain specific methods

- **Identifying objects**
  - Fit elliptic Gaussians to the blobs
  - Identifying bent double galaxies in the FIRST survey

- **Extracting statistics on material fragments**
  - Find threshold using histogram

- **Finding coherent structures in fusion plasma simulations**
  - Find threshold using heuristics
Extracting representative features for the objects
Features are usually problem- and data-dependent

Extract angle-based features

Identifying bent double galaxies in the FIRST survey

Comparisons in Richtmyer-Meshkov instability

Extract features representing texture, shape, ...

Similarity-based object retrieval for simulations
Reducing the dimension of the data
There are several reasons we need to reduce the dimension

The dimension of a problem = number of features describing an object

- We want to identify the important variables to monitor.
- Lower dimensional data are easier to visualize and understand.
- Irrelevant features may lower the accuracy of classifiers.
- More samples are needed to cover a high dimensional space.
- “Nearest-neighbor” may lose meaning in high dimensional spaces.
- The computational cost to extract, store, and use the features may be a consideration.
- Data structures for fast searches do not scale to high dimensions.

Domain scientists can also identify features that are likely to be important.
Dimension reduction can be done by feature selection, ...

Feature selection techniques are used in classification and regression.

Filter methods (distance filter, stump filter, Relief, ...) evaluate how well the features discriminate among the classes.

Wrapper methods use the model to evaluate the subset selected:
- Forward search
- Backward search
or by using feature transform methods

Principal component analysis is a commonly used linear technique.

Non-linear methods include: Isomap, LLE, Laplacian eigenmaps, ...

Random projections (the Johnson-Lindenstrauss lemma) are an option.

- The connection to the original features is no longer clear.
- The non-linear techniques are based on nearest neighbors.
- Some methods are computationally expensive.
- The intrinsic dimensionality of the data is of concern.
Our experiences indicate feature selection works well

Identifying bent-double galaxies in the FIRST survey

Identifying weather conditions associated with ramp events in wind generation

Feature selection techniques tend to be more accurate, give interpretable results, and are computationally less expensive.
Building the models
We can build different kinds of models from data

**Scaling laws**

Power law model for bubble counts in Rayleigh-Taylor instability

**Predictive models (classification/regression)**

Examples: decision trees, support vector machines, neural networks, locally weighted learning, naïve Bayes, ..., and ensembles

**Descriptive models (clustering)**

Examples: k-means, agglomerative and divisive methods, graph-based approaches...

We have a choice of many algorithms. In our experience, the results are dependent more on the data and the pre-processing, than the algorithm.
There are several considerations in building a model

- **Classification:**
  - unbalanced training set
  - incorrect labels
  - need for interpretable models

- **Clustering:**
  - defining a similarity metric

- **Scaling laws:**
  - confirming that it is indeed a power law
The role of analysis techniques in generating the data
There is also interest in the process of generating the data

Motivation: low quality of data, easy availability of computational resources, new kinds of problems being addressed, ...

- There is a large amount of unlabeled data, but labeling is expensive - which samples should we label?
- We want to run simulations to understand how an experiment will perform - what should be the inputs? How do we ensure there is enough variation in the ensembles?
- We want to understand the design space to create a new material using simulations - where should we place our sample points?
- Inverse problems - what input should I use to get a desired output?
- My simulations are expensive; is there a “surrogate” I can use?

The intent is to close the gap between data acquisition and model building.
Several ideas from data mining are relevant

- Active learning, semi-supervised learning, relevance feedback, ...
- Code surrogates, meta-modeling, ...
- Dimension reduction
- Multi-task learning
- Sampling
- Design of (computational) experiments

This is an active area of research with many ideas being explored.
Closing Thoughts
Caveats and notes of caution

- Scientific data mining is a careful and considered application of techniques in close collaboration with the domain scientists.
- A blind use of techniques/software is not recommended.
- Data pre-processing is a critical, but time-consuming, part of the process.
- Using more than one technique to solve a problem is helpful.
- Analysis results must be interpreted carefully before making a decision.
- Many different fields contribute to data mining, each with its own perspectives and insights.

The first principle is that you must not fool yourself - and you are the easiest person to fool.

– Richard P. Feynmann
Challenges and current topics of research

- Streaming data: concept drift, real-time analysis, ...
- Size of the data, exa-scale computing, ...
- Complex data: multivariate, multi-sensor, multi-scale, multi-modal, ...
- New analysis problems: application of techniques to generation of data, problems in simulation data, ...
- Reasoning and decision making in the presence of uncertainty ...
Acknowledgments

- **Data miners and software developers:** Erick Cantú-Paz, Samson Cheung, Ya Ju Fan, Abel Gezahegne, Cyrus Harrison, Thinh Nguyen, Nu Ai Tang, ...

- **Domain scientists:** Robert Becker and the FIRST astronomers, Joshua Breslau, Omar Hurricane, Jeffrey Greenough, Jeffrey Jacobs, Scott Klasky, Zhihong Lin, Paul Miller, Donald Monticello, Neil Pomphrey, ...

- **Funding sources:** ASC program (DOE), ASCR Base Research Program (DOE), the LDRD program at Lawrence Livermore National Laboratory, SciDAC program (DOE), and WHTP/EERE (DOE).

For more details

https://computation.llnl.gov/casc/sapphire/
https://computation.llnl.gov/casc/StarSapphire/
Chandrika Kamath, kamath2@llnl.gov