

Computation Enabling Information Sciences: A Data Miner's Perspective

Zoran Obradovic

Great impressions from yesterday:

- Data mining and HPC technologies enable new scientific discoveries through analysis of almost unimaginably large volumes of data

My concern:

- It is not clear to what extent are these two used together

Bad news:

- Less than 1% of papers submitted to SDM'09 conference were using HPC (3 out of 351 submissions from 26 countries)

Good news:

- Acceptance rate for HPC related papers at SDM'09 was a lot larger than the overall rate (66% vs <30%).

Data Mining: Needs and Aims

Needs for data mining research include:

- Inexpensive storage and new data collection technologies result in exponential growth of stored data
- Demand for a new kind of information management/analysis (shift from computers supporting main processes to playing central role in testing, and even formulation, of hypothesis)

Data mining aims include:

- Developing better methods for analysis of large data
- Applying developed methods towards better understanding of some challenging phenomena

Data Mining:

Achievements and Limitations

Many new methods are developed: pattern discovery, summarization, trend, anomaly detection, prediction etc

However, many challenging issues are not studied enough

Some recognized challenges still need solutions: e.g.

- Spatio-temporal dependency in high dimensional data
- Data collection bias

For a review of these challenges and our proposed solutions see:

Obradovic, Z. and Vucetic, S. (2004) “Challenges in Scientific Data Mining: Heterogeneous, Biased, and Large Sample,” at *The Next Generation Data Mining*

Challenging applications are needed to identify limitations for existing methods: e.g.

- Data fusion (*beyond multiple modalities and multiple resolutions*)

Das, D., Obradovic, Z., Vucetic, S. (2009) “Active Selection of Sensor Sites in Remote Sensing Applications,” *IEEE International Conference on Data Mining*

- Missing values (*beyond multiple imputation*)

Ayuyev, V., Jupin, J., Harris, P., Obradovic, Z. (2009) “Density Based Clustering for Estimation of Missing Values in Mixed Type Data” at *Data Warehousing and Knowledge Discovery*

Challenging Data Mining Applications: Examples from Zoran's Lab

Earth Sciences

- Estimation of geophysical parameters from sensors on satellites (*NSF*)

Social Sciences

- Data mining for juvenile recidivism investigation (*NIJ*)

Biosciences

- Bioinformatics of protein disorder (*NIH, NSF*)
- Data mining in brain image databases (*NIH, NSF*)
- Gene expression data analysis (*NIH, PA Dept. of Health*)
- Bioinformatics core facility (*PA Dept. of Health*)

Other disciplines

- Precision agriculture (spatial-temporal data reduction) (*NSF, DOE*)
- Power systems (distributed data; deregulated markets) (*NSF*)
- Public affairs (text mining) (*PA Dept. of Health*)

Data Mining in Biosciences: Bioinformatics of Protein Disorder

(NIH & NSF: Dunker A.K and Obradovic Z.)

Aim: Understanding protein disorder and its functions

Developed effective methods for:

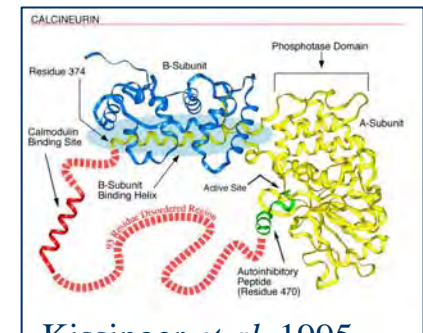
- Classification of sequences from a mixed distribution,
- Feature selection,
- Learning from a biased sample,
- High dimensional distribution learning,
- Anomaly detection,
- Text mining,
- Uncertainty reduction, etc.

Main Results: (>80 articles)

- Protein disorder is highly predictable (our predictors winners at 3 CASP competitions)
- Protein disorder is very common
- Fraction and type of disorder varies a lot by genomes
- Involved in a variety of functions
- Characterized w.r.t. major diseases

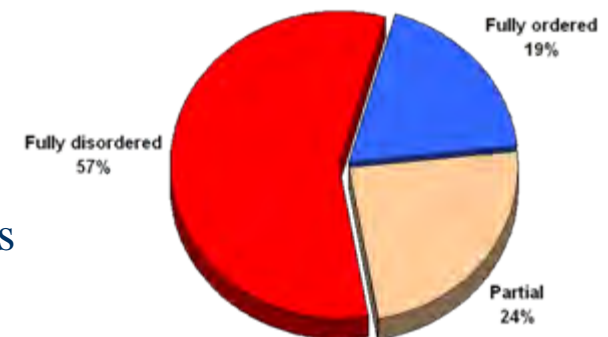
Current students: U. Midic, Z. Ping, R. Siyuan, Q. Xu

Disorder has function



Kissinger *et al.*, 1995

Alternative Splicing occurs mostly
in disordered regions



Romero *et al.*, *PNAS*. 2006

Data Mining in Atmospheric Sciences: Multiple-Source Spatial-Temporal Data Analysis

(NSF SIII : Obradovic Z., Vucetic S, Li Z)

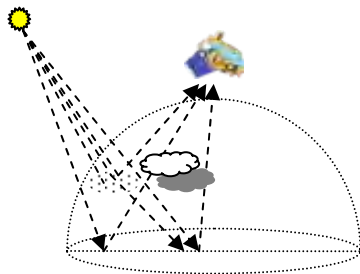
Aim: Accurate and efficient estimation of geophysical parameters from multiple satellites and ground based observations (*huge data streams*)

Why: Support climate and environmental protection studies, global trend analysis

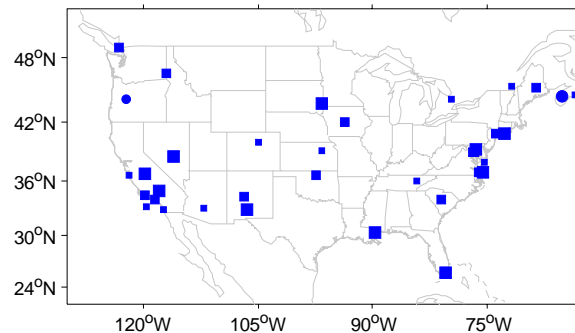
Main Results:

- ◆ Discovered major sources of correctable aerosol retrieval error of deterministic models
- ◆ Improved aerosol retrieval accuracy by integrating multisource data
- ◆ Found good spatio-temporal partitioning of Earth w.r.t. aerosol properties

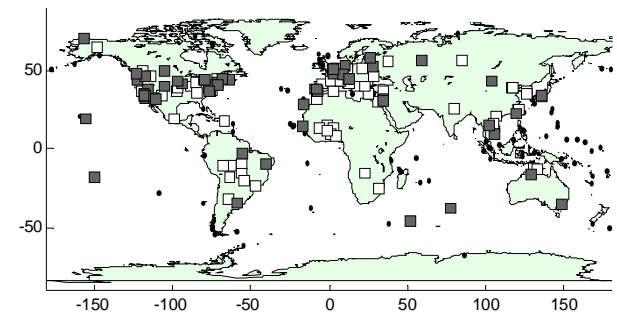
Challenges: Clouds
and bright surface



Our statistical estimation is more
accurate vs. deterministic



We found a good two cluster
partitioning for summer – fall



Current Students: D. Das, Q. Lou, V. Radosavljevic, K. Ristovski, H. Shi



Thank you

More information:

<http://www.ist.temple.edu>

Contact:

Zoran Obradovic, director
Information Science and Technology Center
Temple University
+1 215-204-6265
zoran@ist.temple.edu