Winnowing signals from massive data: SP for Big Data and its Relation to Systems Engineering

Alfred Hero

University of Michigan - Ann Arbor

March 21, 2013
1 Definitions

2 Requirements and challenges

3 SP for ranking

4 Conclusions
Outline

1. Definitions
2. Requirements and challenges
3. SP for ranking
4. Conclusions
Winnowing grain from chaff

Figure: Left: manual winnowing process. Right: Mechanical wind winnowing machine, illustration of the encyclopedia of education, St. Petersburg, Russian Empire, 1896
An engineer designs the function $f$ that computer evaluates on data and models:

$$\text{signal} = f(\text{data}, \text{models})$$
Winnowing signals from data and models

Feedback improves model and data acquisition
- Human-assisted processing, relevance feedback learning
- Plan-ahead sensing, sensor management, sequential DOE
Winnowing signals from data and models

Markov decision process (MDP) framework

• Human assistance can occur at any stage
• Full multistage optimization of MDP is intractible
• Useful framework for obtaining bounds and inspiration
Big Data: personalized health and medicine

Functional pathway model

Symptom model

Epidemic model

Jenner and Young, Nature Reviews 2005

Google Health Map

Metabolo/Geno/Proteo-mics

Clinical data

Close contact data

Ingenuity pathway analysis software

Aiello Research Group, UM SPH
**Big Data**

**Big data:** “A key tenet of big data is that the world and the data that describe it are *constantly changing* and organizations that can recognize the changes and react quickly and intelligently will have the upper hand…”

“…As the volume of data explodes, organizations will need analytic tools that are *reliable, robust and capable of being automated*. At the same time, the analytics, algorithms, and user interfaces they employ will need to *facilitate interactions* with the people who work with the tools.”

Big Data processing/analysis requirements

- Integration of very heterogenous data
  - Correlation mining in massive database
  - Data at vastly different scales and noise levels
  - Mixture of continuous and categorical variables

- Reliable and robust quantitative models
  - Uncertainty quantification
  - Adaptive to drift over time

- High throughput real-time processing
  - Smart adaptive sampling and compression
  - Distributed or parallel processing architectures

- Interactive user interfaces
  - Human-in-the-loop processing
  - Visualization and dimensionality reduction
Some signal processing challenges

• Heterogeneous data integration
  • Ranking signals for human-aided selection of relevant variables
  • Fusing graphs, tensors, and sequence data
  • Active visualization: dimensionality reduction

• Flexible low complexity modeling and computation
  • Scalable SP: distributed algorithms and implementation
  • Smart sampling: feedback controlled signal search and acquisition

• Reliable robust models for anomaly detection and classification
  • Parsimonious SP: Sparse correlation graphical models
  • Decomposable SP: factored models and algorithms
Signal processing toolbox

Primitives

- Linear equation solvers (Gauss, Givens, Householder)
- Spectral representations (FFT, SVD)
- Ensemble averaging (cross validation, bootstrap, boosting)
- Optimization (LLS, linear&quadratic programming, DP)
Signal processing toolbox

Primitives

- Linear equation solvers (Gauss, Givens, Householder)
- Spectral representations (FFT, SVD)
- Ensemble averaging (cross validation, bootstrap, boosting)
- Optimization (LLS, linear & quadratic programming, DP)

are used for

- **Linear & NL prediction**: Wiener, Kalman, particle filtering, Volterra filters
- **Signal reconstruction**: matrix factorization, matrix completion, robust PCA
- **Dimension reduction**: PCA, ICA, IPCA, CCA, LDA, NLE
- **Adaptive sampling**: compressive sensing, distilled sensing, sketching.
- **SP on graphs**: graph spectra, knn search, belief propagation
Search a database, e.g. Google, for best matches to an image query.

**Image size:**
3264 × 1840

No other sizes of this image found.

**Visually similar images** - Report images
Single query search: linear ordering

Matches to query $i$ sorted according to dissimilarity measure

$$f(i_1) < f(i_2) < \ldots < f(i_n)$$

This encourages people to interact with Google’s algorithm, leading to improvements
Single query search: linear ordering

Matches to query $i$ sorted according to dissimilarity measure

$$f(i_1) < f(i_2) < \ldots < f(i_n)$$

This encourages people to interact with Google’s algorithm, leading to improvements

SP dominated by algorithms that find a solution to an optimization

- Basis pursuit and dictionary learning find “a best match.”
- Parametric estimation produces a ML, MAP, or min MSE estimator.
- Compressive sensing, matrix completion give “the best signal reconstruction.”
Single query search: linear ordering

Matches to query $i$ sorted according to dissimilarity measure

$$f(i_1) < f(i_2) < \ldots < f(i_n)$$

This encourages people to interact with Google’s algorithm, leading to improvements

SP dominated by algorithms that find *a solution* to an optimization

- Basis pursuit and dictionary learning find “a best match.”
- Parametric estimation produces a ML, MAP, or min MSE estimator.
- Compressive sensing, matrix completion give “the best signal reconstruction.”

Emerging area in Machine Learning and SP: “Learning to rank”


*Jamieson and Nowak, Active ranking using pairwise comparisons, arxiv, 2011.*

Dual query search

Sometimes a single query is not adequate
Alternative: dual query and multi-objective optimization

The similarity between query pair and database image $i$ is now a vector $[f_1(i), f_2(i)]$.

One idea: rank according to scalarization $f_\lambda = \lambda f_1 + (1 - \lambda) f_2$.

$$f_\lambda(i_1) < f_\lambda(i_2) < \ldots < f_\lambda(i_n)$$
Skyline search: non-dominated (Pareto) ranking

Drawback of scalarization: need fix $\lambda$; unknown user-dependent.

Non-dominated sorting: a point $i$ is non-dominated if there exists no other point $j$ such that $f_1(j) < f_2(i)$ and $f_2(j) < f_2(i)$.

The set of non-dominated points is an “antichain” called the Pareto front. Set of Pareto fronts is canonical antichain partition.

Papadias, Tao, Fu, Seeger, An Optimal and Progressive Algorithm for Skyline Queries, SIGMOD 2003
Searching for matches to dual queries over Pareto fronts

Another application: multicriteria anomaly detection

Motivation: Detect anomalous pedestrian trajectories.
Question: Which one of these groups of trajectories are anomalous?

Curve features: curve length, shape, walking speed.

Another application: multicriteria anomaly detection

Speed and shape similarity between trajectories \( T_i(t), T_j(t) \in \mathbb{R}^2 \):

\[
D_1(i, j) = \| \text{hist}(\| \Delta T_i \|) - \text{hist}(\| \Delta T_j \|) \|
\]

\[
D_2(i, j) = \| T_i - T_j \|
\]

1. Scalarization:
   \[ D_\lambda(i, j) = \lambda D_1(i, j) + (1-\lambda)D_2(i, j) \]

2. Pareto depth analysis:
   \((D_1(i, j), D_2(i, j)) \rightarrow \text{one dyad}\)

Performance of multicriteria anomaly detection

PDA Algorithm:
- Embed N choose 2 dyads onto plane
- Build Pareto fronts of non-dominated dyads.
- Compute anomaly scores = depth of front.

PDA outperforms scalarization
Questions on non-dominated (Pareto) sorting

• Is there an asymptotic theory (large $n$) for shape of the Pareto front $T$?

• What is average number of points on $T$?

• Can computational complexity of finding Pareto fronts be reduced?
Questions on non-dominated (Pareto) sorting

- Is there an asymptotic theory (large $p$) for shape of the Pareto front $T$?
  \[ \Rightarrow \text{Yes. The Pareto front is solution to a pde on } \mathbb{R}^d. \]

- What is average number of points on $T$?
  \[ \Rightarrow E[N_{\text{Pareto}}] = \gamma n^{(d-1)/d} + O(n^{(d-2)/d}) \text{ with} \]
  \[ \gamma = d^{-1} (d!)^{1/d} \int_T \frac{d^{-1}}{2} (u(z))(u_1(z) \cdots u_d(z))^{1/d} dz \]

- Can computational complexity of finding Pareto fronts be reduced?
  \[ \Rightarrow \text{Yes. In principle can reduce from } O(dn^2) \text{ to } O(1). \]
Asymptotic theorem

Let there be $d$ criteria giving non-negative similarity score $X_i = [f_1(i), \ldots, f_d(i)]$ with $i$th image in the database, $i = 1, \ldots, n$.

Assume that $\{X_i\}_{i=1}^n$ are i.i.d. from multivariate density $f(x_1, \ldots, x_d)$.

**Theorem**

*As $n \to \infty$ the Pareto fronts converge uniformly to the level sets of the value function $U(x_1, \ldots, x_d)$ where $U$ is the non-viscosity solution to the Hamilton-Jacobi partial differential equation:*

$$\frac{\partial U}{\partial x_1} \cdots \frac{\partial U}{\partial x_d} = \frac{1}{d^d} f$$

Calder, Esedoglu and H, “A Hamilton-Jacobi equation for the continuum limit of non-dominated sorting,”
Illustration of Asymptotic Theory

Conclusions

- Signal processing meets big data in processing of mixed data types
  - Spatio-temporal, continuous-valued, categorical, graphs, human inputs
  - Feedback and active processing

- Non-dominated sorting is an interesting and useful framework for multi-criteria anomalies, human-machine interaction, or multiple end-users
Human evolution
Acknowledgements

Thanks to my sponsors:

DARPA (N66001-07-C-2024)
AFOSR (FA9550-13-1-0043, FA8650-07-D-1220, FA8650-09-D-5037)
ARO (W911NF-09-1-0310, W911NF-11-1-0391, W911NF-12-1-0443)
NSF (CCF-1217880)
NIH (2P01CA087634-06A2)
