Distributed Data Mining for Pervasive and Privacy-Sensitive Applications

Hillol Kargupta
Dept. of Computer Science and Electrical Engg,
University of Maryland Baltimore County
http://www.cs.umbc.edu/~hillol
hillol@cs.umbc.edu

Roadmap

- Distributed Data Mining (DDM)
- Pervasive and Privacy-Sensitive Applications of DDM
- Dealing with ensemble of data mining models
- Linear representations for advanced meta-level analysis of models
- Conclusions

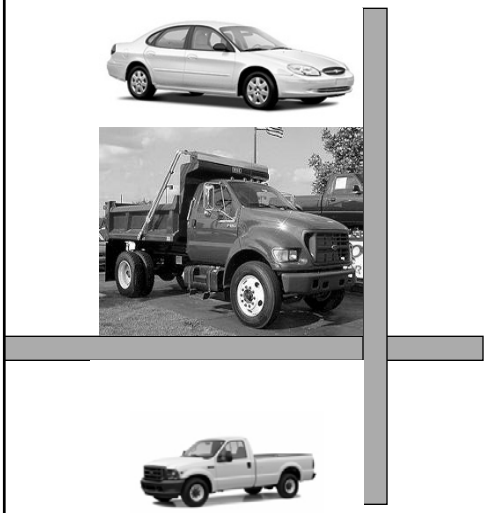
Distributed Data Mining (DDM)

- Distributed resources
  - data
  - Computation and communication
  - users
- Data mining by properly exploiting the distributed resources

Distributed Resources and DDM

- Distributed compute nodes connected by first communication network
  - Partition data if necessary and distribute computation
- Inherently distributed data that may not be collected to a single site or re-partitioned
  - Connected by limited bandwidth network
  - Privacy-sensitive data

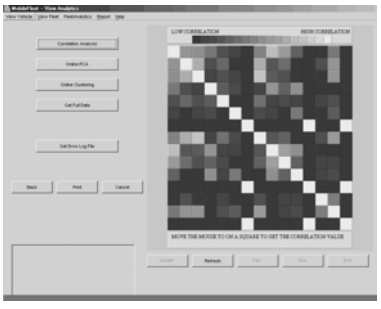
Pervasive Applications: UMBC Fleet Health Monitoring

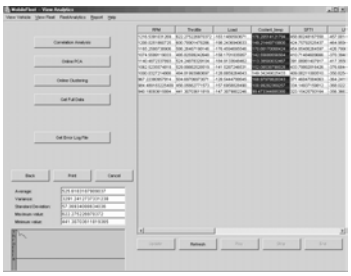


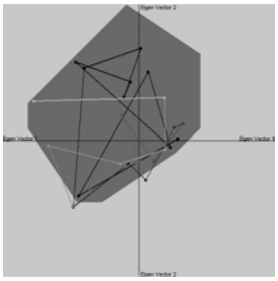
- **Vehicle Health Monitoring Systems**
- **Collect and analyze vehicle related information.**
- **On-board/*in situ* data analysis**
- **Send out interesting patterns**
- **Analyze data for the entire fleet**
- **UMBC fleet operations management**

Continued…

- Onboard real-time vehicle-mining system over a wireless network

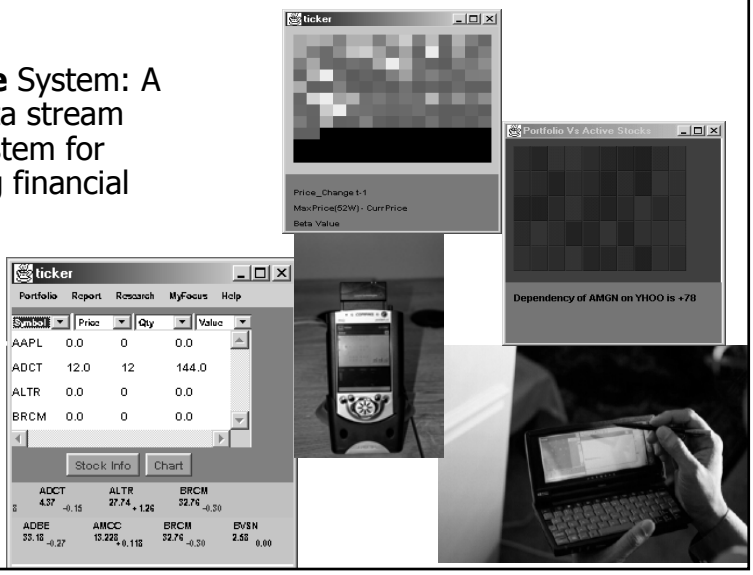




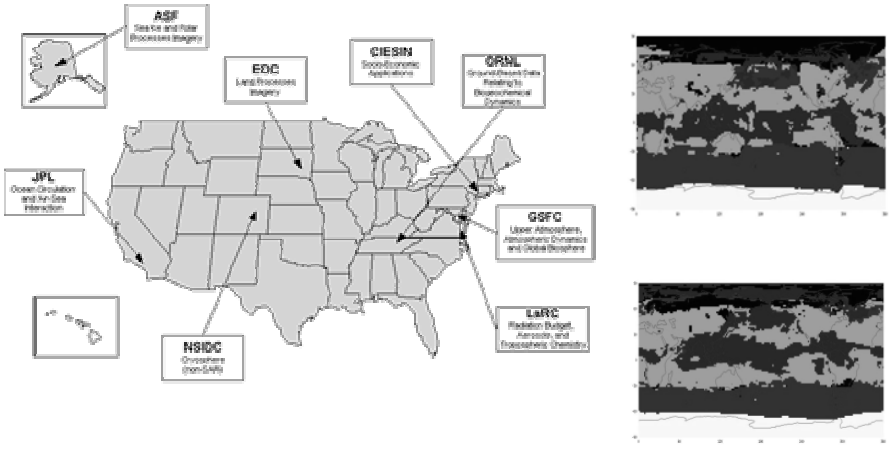


Pervasive Applications: MobiMine

- **MobiMine** System: A mobile data stream mining system for monitoring financial data



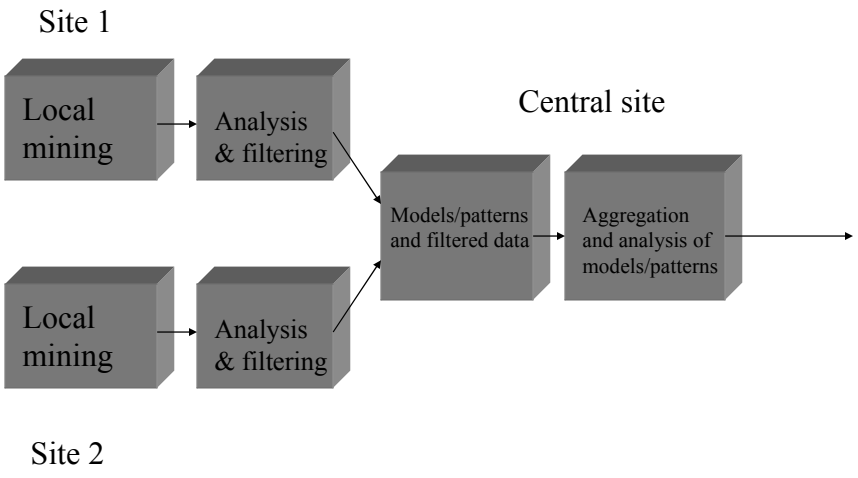
DDM from NASA EOS Distributed Data Repositories



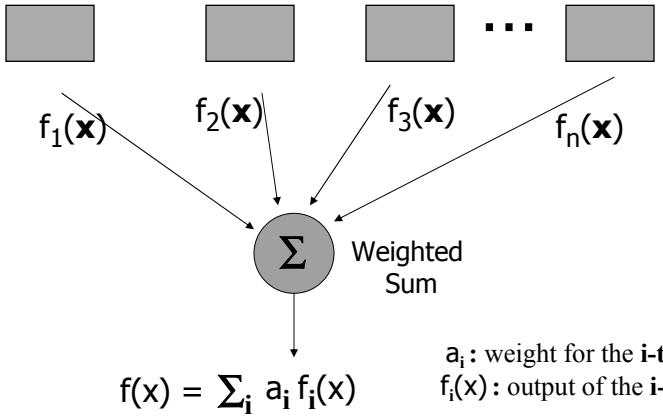
Mining from Distributed Privacy-Sensitive Data

- Analyze data without moving the data in its original form.
- Many DDM algorithms are privacy-friendly since they minimize data communication.

Distributed Data Mining



Ensemble of Classifiers and Clusters



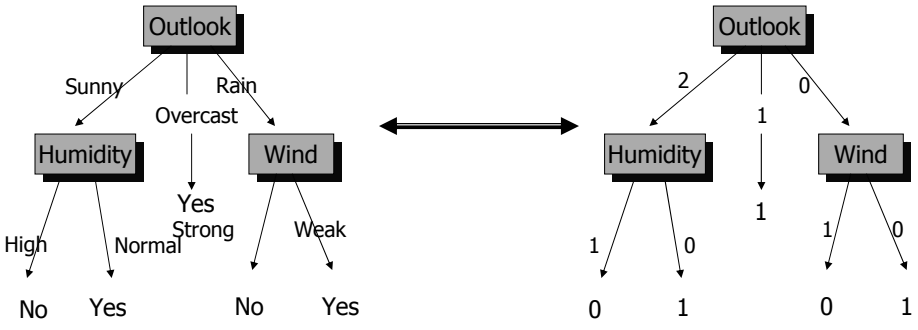
Discrete Structures for Data Mining Models

- Trees, in general Graphs are popular choices for data mining models:
  - Decision trees (Tree)
  - Neural networks (Graph)
  - Graphical models (Graph)
  - Clusters (Graph, hypergraph)
- Dealing with ensembles requires an algebraic framework.

Examples

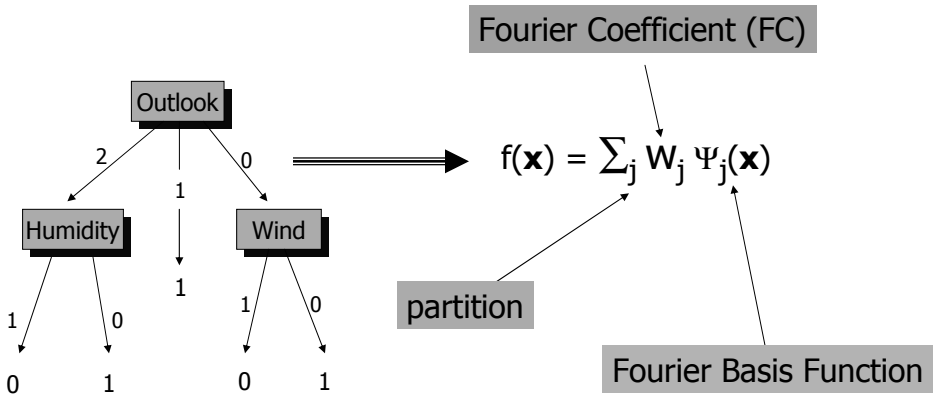
- Eigen analysis of graphs:
  - Graphs can be represented using matrices
  - Eigen analysis of the Laplacian of graphs (Chung, 1997).
- Wavelet, Fourier, or other representations of discrete structures??

Decision Trees as Functions



- Decision tree can be viewed as a numeric function.

Fourier Representation of a Decision Tree



Fourier Basis

$$f(x) = \sum_{j \in \Xi} w_j \Psi_j(x)$$

$\mathbf{j}, \mathbf{x} \in \{0, 1\}^l$

j-th Fourier basis function, $\Psi_j(x) = (-1)^{j.x}$

w_j is the corresponding Fourier coefficient;

$$w_j = \frac{1}{N} \sum_x f(x) \Psi_j(x)$$

Partitions

A partition **j** is an *l*-bit boolean string.

It can also be viewed as a subset of variables.

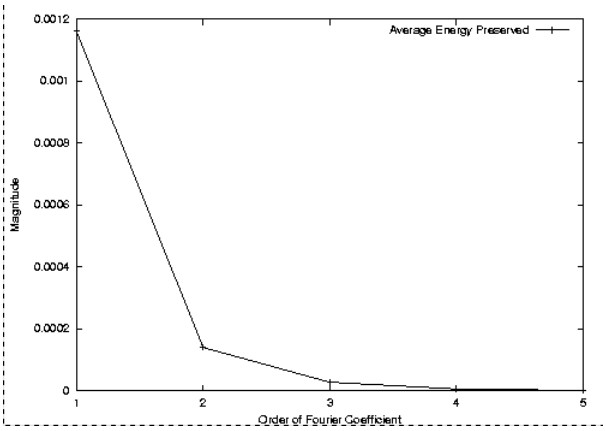
Example:
Partition 101 $\Rightarrow \{x_1, x_2\}$ contains the features associated with locations indicated by the 1-s in the partition.

Order of a partition = the number 1-s in a partition.

Fourier Spectrum of a Decision Tree

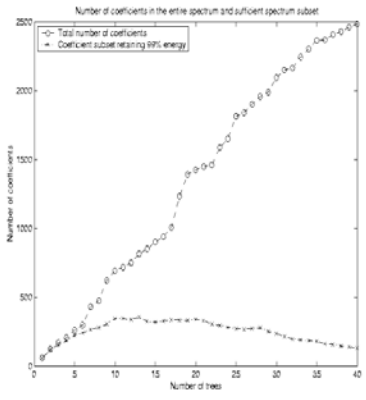
- Very sparse representation; polynomial number of non-zero coefficients. If k is the depth then all coefficients involving more than k features are zero.
- Higher order coefficients are exponentially smaller compared to the low order coefficients (Kushlewitz and Mansour, 1990; Park, Kargupta, 2001).
- Can be approximated by the low order coefficients with significant magnitude.
- Further details in [Linial, Mansour, Nisan, 89], [Park, Ayyagari Kargupta 01'], [Kargupta et al. 2001].

Exponential Decay of FCs (S&P 500 Index Data)

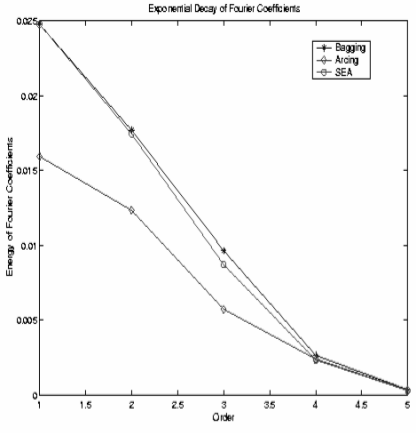


Compression

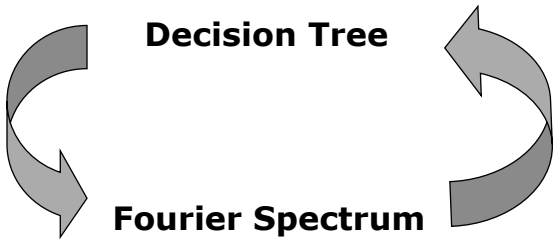
Sufficient spectrum (99% energy)



Energy preserved in the Lower Order Coefficients

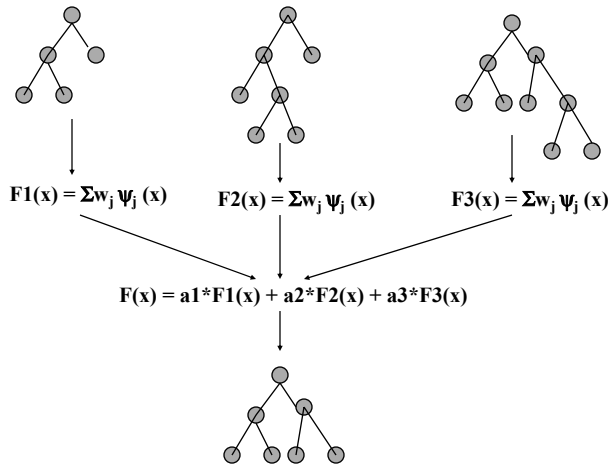


Fourier Spectrum and Decision Trees



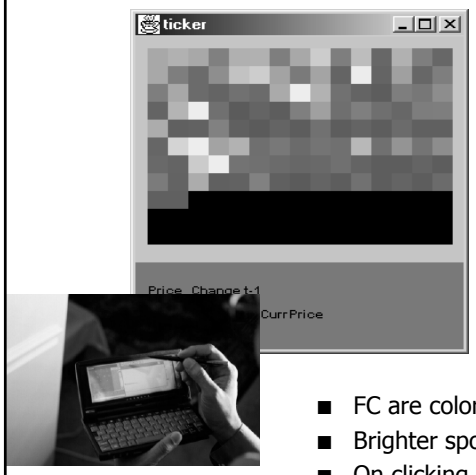
- Developed efficient algorithms to
  - Compute Fourier spectrum of decision tree
    (IEEE TKDE, SIAM Data Mining Conf., IEEE Data Mining Conf, ACM SIGKDD Explorations)
  - Compute tree from the Fourier spectrum
    (DMKD, SIGMOD 2002)

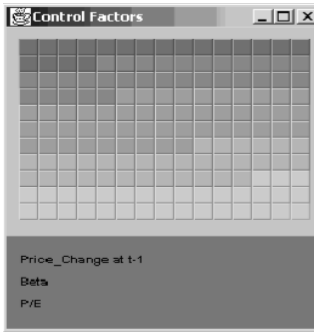
Aggregation of Multiple Decision Trees



- Weighted average of decision trees through Fourier analysis

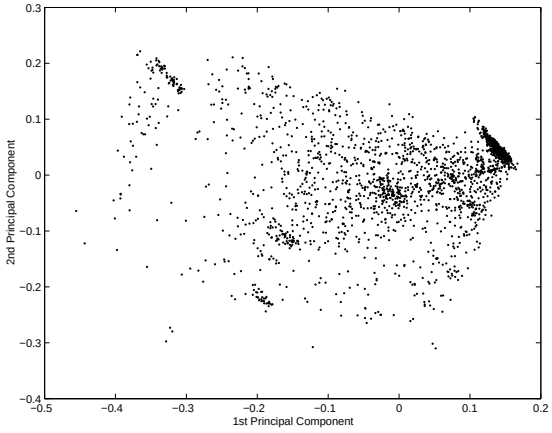
Visualization of Decision Trees





- FC are color-coded in accordance to the magnitude.
- Brighter spots are more significant coefficients.
- On clicking, partition corresponding to the coefficient is displayed.

PCA-Based Visualization of Decision Trees



Redundancy Reduction: Orthogonal Decision Trees

True output of the target function
1
-1
-1
1
1
1
-1
1

Matrix D

Tree1	Tree2	Tree3	Tree4
1	-1	1	-1
-1	-1	-1	-1
-1	1	-1	1
-1	1	1	1
1	1	1	-1
1	1	-1	-1
-1	-1	-1	1
1	-1	1	1

All domain members

PCA-Based Redundancy Reduction

- Trees may share underlying redundancy.
- Perform PCA; the eigenvectors tell us how to combine the trees for creating a basis set.
- Problems:
  1) Impractical, D is very very large for most applications.
  2) You only get the weights of the base classifiers.
- Approximating D over the training data (Merz and Pazzani, 1999).

Inner Product of Decision Trees and Fourier Transformation

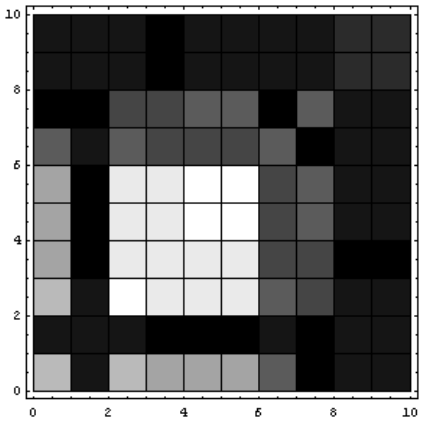
- Inner product between trees f1(x) and f2(x) :

$$< f1(x), f2(x) >= \sum_{x} f1(x) f2(x)$$

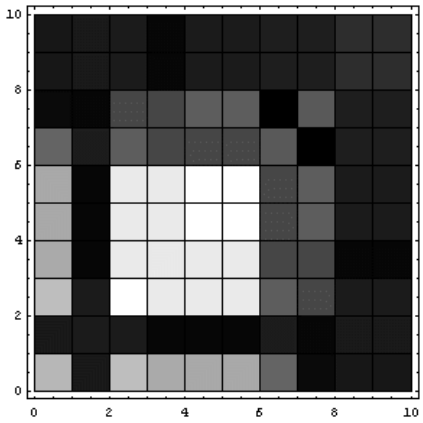
- If $\overline{W1}$ and $\overline{W2}$ are the corresponding Fourier spectra then:

$$< f1(x), f2(x) >=< \overline{w1}, \overline{w2} >$$

Inner Product Matrices



(a) Between Trees



(b) Between the Fourier Spectra

The Fourier Spectra Matrix

- Consider W, where $W_{i,j}$ is the Fourier coefficient of the i-th basis from the spectrum of the tree T_j.
- W^TW and D^TD are identical.
- W is a smaller matrix compared to D.
- So we can efficiently compute the eigenvectors using W^TW.

Conclusions

- Distributed data mining appears interesting for pervasive and privacy-sensitive applications.
- We need meta-level techniques to analyze aggregate the data mining models:
  - Stability of models/ensembles
  - Detecting changes in the model distribution
  - Many other issues….

Advertisement

- IEEE Transactions on System, Men, Cybernatics, Part B, Special Issue on Distributed and Mobile Data Mining
- Deadline: January 1, 2003.
  http://www.cs.umbc.edu/~hillol/DKD/smcb_dmdm.html

Hillol Kargupta

- **Hillol Kargupta** is an Assistant Professor in the Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County. He received his Ph.D. in Computer Science from University of Illinois at Urbana-Champaign in 1996. He is also a co-founder of Agnik LLC, a ubiquitous data intelligence company. His research interests include mobile and distributed data mining, computation in gene expression, and genetic algorithms.

  Dr. Kargupta won a National Science Foundation (NSF) CARRER award in 2001 for his research on ubiquitous and distributed data mining. His research is also funded by several other grants from NSF and NASA. He also received support from the TRW Research Foundation, American Cancer Society, US Department of Energy, and Caterpillar. He won the 1997 Los Alamos Award for Outstanding Technical Achievement. His dissertation earned him the 1996 Society for Industrial and Applied Mathematics (SIAM) annual best student paper prize. He has published more than fifty peer-reviewed articles in journals, conferences, and books. He is the distributed data mining consultant for DaimlerChrysler. He is the primary editor of a book entitled "Advances in Distributed and Parallel Knowledge Discovery", AAAI/MIT Press. His other recent activities include hosting the ACM SIGKDD-2000 workshop on Distributed and Parallel Knowledge Discovery (DPKD), KDD-98 workshop on distributed data mining, a special issue on DPKD in *Knowledge and Information Systems Journal*. He is the co-chair of the IJCAI-2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases. He is in the program/organizing committee for the 2001 & 2002 SIAM Data Mining Conference and the 2001 ACM SIGKDD Conference among several others. He is also the co-chair of a workshop on ubiquitous data mining in PKDD-2001. More information about him can be found at http://www.cs.umbc.edu/~hillol.