Data Mining in Distributed and Ubiquitous Environments: Past, Present, and Future

Hillol Kargupta
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, MD 21250, USA
http://www.cs.umbc.edu/~hillol
hillol@cs.umbc.edu
&
AGNIK, LLC
Columbia, MD 21045
http://www.agnik.com
hillol@agnik.com

Roadmap

- Introduction
- What is Ubiquitous Data Mining?
- Applications
- Algorithms
- Benchmarking
- Products
Research & Development at UMBC DIADIC Laboratory and AGNIK, LLC

- Distributed and mobile data mining.
- Supported by NASA, US National Science Foundation CAREER award and other grants, US Air Force, TRW Research Foundation, Maryland Technology Development Council, and others.
- Agnik, LLC: A Spin-off from DIADIC Lab, specializing on mobile and distributed data mining and management.

Data Mining

- Data Mining: Scalable analysis of data by paying careful attention to issues in computing, storage, communication, and human-computer interaction.
Distributed Data Mining

- Distributed data mining (DDM): Mining data using distributed resources.
  - Pays careful attention to the distributed resources of data, computing, communication, and human factors in order to use them in a near optimal fashion.

What is Ubiquitous Data Mining?

- Distributed or resource aware
  - computing,
  - communication,
  - storage, and
  - human-computer interaction.
Early Days of the Community

- ACM SIGKDD Workshop on Distributed Data Mining, 1998.
- ACM SIGKDD Workshop on Distributed Data Mining, 2000.
- PKDD Workshop on Ubiquitous Data Mining for Mobile and Distributed Environments, 2001.

Data Mining in Distributed and Mobile Environments

- Mining databases from distributed sites
  - Earth Science, Astronomy, Counter-terrorism, Bioinformatics

- Monitoring multiple time critical data streams
  - Monitoring vehicle data streams in real-time
  - Onboard science

- Analyzing data in lightweight sensor networks
  - Limited network bandwidth
  - Limited power supply

- Preserving privacy
  - Security/Safety related applications
Evolution of Applications

Few Early Applications

- Work on multi-agent learning, ensemble learning


- Los Alamos National Laboratory, PADMA system for distributed text data mining, Kargupta, 1996.
MobiMiner: A Mobile Data Stream Mining System for Stock Market Data

- An interactive PDA-based data mining system for stock-market data
- Visualize the spectrum of the decision tree

Resource-Constrained Real-time Physiological Data Stream Monitoring

- Wearable sensors available in the market
  - SenseWear Armband from BodyMedia
  - Wearable West
  - LifeShirt Garment from Vivometrics
- SenseWear armband can measure heat flux, accelerometer, galvanic skin response, skin temperature, near body temperature
- Arm band can store up to about 5 days of data.

1. www.smartextiles.info

http://www.armband.it/
http://www.vivometrica.com
A Network of Monitoring Devices

Detecting emerging patterns in a group of health workers, soldiers, elderly individuals, animals.

MineFleet: A Vehicle Data Stream Management and Mining Software System

**On-board Module:**
- Continuous data streams from the vehicle data bus
- Onboard data stream mining
- Communicates with a remote control station
- Privacy management

**Central control station:**
- Data Management
- Data mining
- Communicates with the on-board modules over wireless networks
- Privacy management

Data Collection Module: Components

Optional GPS module

Closer Look

Power supply

OBD-II adapter connected to a Ford Van OBD-II port
Modes of Operations

- **Passive Mode:** Collect data and analyze later offline.

- **Active Mode:** Analyze data in real-time either on-board (cell-phone, PDA, embedded device) or remote desktop connected over wireless network.

Vehicle Data Stream Mining

- **Vehicle Health Monitoring and Maintenance:**
  - Several model and data driven fault tests
  - Detecting unusual behavior for a subsystem and accessing the data producing this behavior

- **Fuel Consumption Analysis:**
  - Is the vehicle burning fuel efficiently? Identify influencing factors and optimize
  - Detect influence of driver behavior on gas mileage and eliminate inefficient driving practices

- **Driver Behavior Monitoring:**
  - Route monitoring: Fixed and variable routes
  - Direct Cost Issues: e.g. Idling, braking habits
  - Safety Issues: e.g. speeding, trajectory monitoring (e.g. stopping, turns)

- **Vehicle location related services**

- **Vehicular network security and privacy management**
**Fuel System**
- Oxygen Sensor Operating Condition Monitoring
- Long Term Fuel Related Combustion Efficiency Monitoring
- Air Intake Volume Inconsistency Monitoring
- Engine Intake Vacuum Inefficiency Monitoring
- Engine Thermal Event Detection
- Throttle Request Status Monitoring
- Idle Control Monitoring
- Intake Air Management Monitoring
- Quantitative Fuel Management Monitoring
- Vehicle System Temperature Management Monitoring
- Quantitative Fuel System Management monitoring

**Exhaust System**
- Combustion Temperature Inequality Monitoring
- Combustion Temperature Control Decay Monitoring

**Ignition System**
- Vehicle Ignition System Voltage Monitoring
- Spark Control Monitoring
- Vehicle Operating System Voltage Monitoring

---

Discoveries Across Multiple Databases

from dozens to hundreds of databases

**Courtesy Bob Grossman, Illinois**
Combining Microarray Data & Clinical Data


Aguilar-Ruiz et al. (2004). Data Mining Approaches to Diffuse Large B-Cell Lymphoma Gene Expression Data Interpretation.

Correlating Microarray Data and Clinical Data

- International Prognostic Indicator (IPI), a clinical indicator of prognosis, has been successfully used to define prognostic subgroups in DLBCL.

- The clusters in the microarray data provide additional prognostic information not available in the IPI.

- Virtually combining local columns in a clinical database with remote columns from a microarray database.
Hackers Target U.S. Power Grid
Government Quietly Warns Utilities To Beef Up Their Computer Security

By Justin Blum
Washington Post Staff Writer
Friday, March 11, 2005; Page E01

Hundreds of times a day, hackers try to slip past cyber-security into the computer network of Constellation Energy Group Inc., a Baltimore power company with customers around the country.

"We have no discernable way of knowing who is trying to hit our system," said John R. Collins, chief risk officer for Constellation, which operates Baltimore Gas and Electric. "We just know it’s being hit."

PURSUIT: Privacy-Sensitive Cross-Domain Intrusion Detection
- Cross-Domain Network Attack Detection system using Privacy-Preserving Distributed Data Mining
- Sponsor: US Department of Homeland Security
- Partners:
  - Agnik
  - Army High Performance Research Center, University of Minnesota
  - Tresys Inc.
- PURSUIT Consortium:
  - Purdue University
  - Ohio State University
  - Stevens University
  - SRI International
  - University of Illinois at Urbana-Champaign
Spatial Attack Distribution of IPs on the Same Day: (Left) IPs attacking the UFL network on 12/09/04 (712 scanners). (Middle) IPs attacking the UMN network on 12/09/04 (14,938 scanners). (Right) Intersection of the IPs attacking UFL and UMN (201 scanners). Courtesy: Vipin Kumar, UMN

PURSUIT: Objectives

- Discovering Attacker Signatures based on the Network of Zombie Hosts
- Discovering Attack Patterns on Coalition members
- Discovering New Distributed Stealth Attacks
P2P DDM Applications: An Exciting Area

- P2P Music mining
- User behavior data mining in a p2p network
- Mobile ad hoc vehicular networks

P2P Distributed Data Mining Algorithms

- P2P clustering
- P2P association rule learning
- P2P eigenstate monitoring
- P2P outlier detection
- Some exciting upcoming commercial applications
DDM Algorithms

- Distributed association rule learning
- Collective decision tree learning
- Collective PCA and PCA-based clustering
- Distributed hierarchical clustering
- Other distributed clustering algorithms
- Collective Bayesian network learning
- Collective multi-variate regression
- Distributed support vector machine learning
- Distributed construction ensemble models
- Ensemble-based aggregation

http://www.cs.umbc.edu/~hillol/DDMBIB

Benchmarking

- Scalability
  - computing,
  - communication,
  - storage, and
  - human-computer interaction?

- Benchmarking Privacy??
- Resource consumption
  - Power
Power Consumption Behavior of Data Mining Algorithms

- HP Jornada 690 (Hitachi SuperH SH-3).
- Two networks – CDPD and 802.11b.
- Agilent 54622A oscilloscope.

Building DDM Systems

- Public domain system: DDM Toolkit
- JADE-based distributed system in Java.
- Currently being beta-tested
- If you want a beta version please send me an e-mail.
Survey Articles & Text Books


- Upcoming text book on distributed data mining