Information Era - Knowledge Age

Objective

Acquire knowledge over large and distributed databases.

Problems

- Time and Space scaling
  - More data become available every day.
  - Existing learning algorithms are main-memory based.
  - Application and algorithm complexity is increased.
  - It may require several months to learn a million records.
- Information is distributed
  - Efficiency reasons prohibit transferring.
  - Proprietary reasons forbid sharing.
  - Distributed databases are inherently partitioned.
  - Learning from a small subset of data can be inaccurate.
- Existing searching engines have limited functionality.

Question

How can a system learn efficiently and accurately from large amounts of data that can be distributed?

Means of acquring knowledge: Data Mining/Meta Learning

- *Data Mining:* The application of machine learning algorithms that compute descriptive representations and summarizations or find repetitive patterns exhibited in large stores of data.
- *Classifier:* The result of a data mining application. Encodes a learned predictive theory of the data that partitions related data into semantically meaningfull abstructions.
- *Meta Learning:* A general strategy that provides the means of learning how to combine and integrate a number of separately learned classifiers or models.
- *Meta Classifier:* The result of meta learning. Represents the collective knowledge derived from individual knowledge sources.

The JAM Project: Java agents for Meta-Learning

It is a software system that supports intelligent information extraction in a network computing environment (i.e. the internet). This includes:

- A Java agent infrastructure that realizes the host environment for designing a meta-learning system.
- A collection of Java based agent modules implementing a variety of machine learning algorithms, as well as the meta-learning techniques architected by the host design system.
- A collection of Java based agent modules implementing classifiers or meta-classifiers as derived by learning agents, that can be applied to remote data over the network.
- Secure agent mechanisms to protect access and execution of the various agents used by the project.
- A convenient interface allowing end-users to launch Java learning agents in support of information extraction functions.

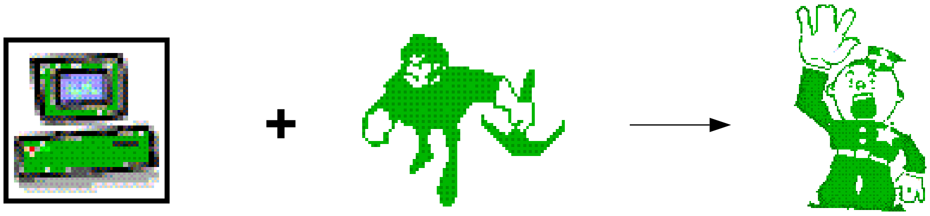
JAM: step by step

- Large datasets are partitioned into subsets.
- Java learning agents are dispached to each of these sites, and produce base classifier agents.
- Newly created classifier agents travel among participating cites. Local meta-learning agents use them to produce meta-level training data by generating base predictions on training data.
- Meta-learning agents produce meta-classifiers based on the meta-level training data.
- The newly created meta-classifier represents a function which combines classifiers from multiple parallel learning processes.
- Meta-classifiers serve as agent proxies that use learned knowledge to search and analyze other data.
- Results indicate that meta-classifiers can achieve at least the accuracy of the "Global Classifier" (if all data were present at one site).

Intrusion and fraud detection in financial information systems

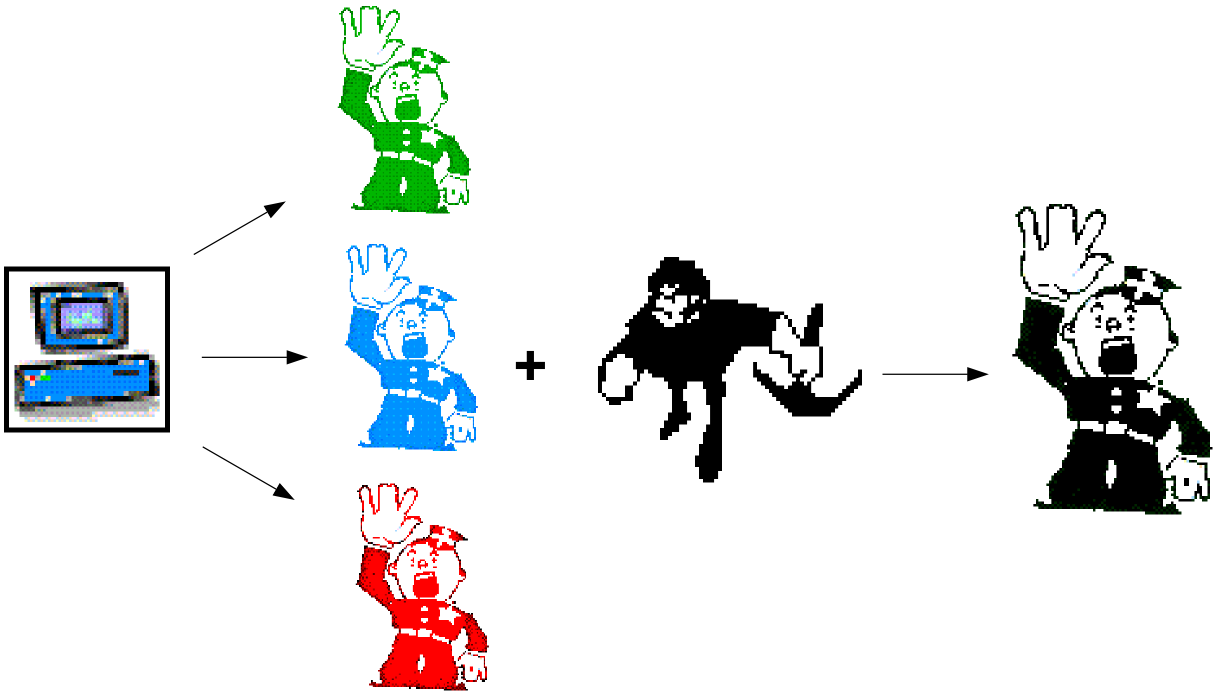
---

- Volume of electronic transactions is increasing, number of access points into the data network is growing rapidly.
- FSTC goal: attack the problem of fraudulent electronic transactions from a unified global viewpoint.
- Problem: need for privacy (competitive and legal issues prohibit financial institutions from sharing their raw data)
- Solution: share knowledge WITHOUT sharing data
  - Employ pattern-directed inference systems that use models of anomalous or errant transaction behaviors.
  - Employ the local fraud detection agents that learn how to detect fraud within a single corporate institution
  - Employ a secure and integrated meta-detection system that joins knowledge (but NOT data) acquired by individual local agents to gain a global view of the network transactions
  - Install the resultant meta-fraud-detectors as SENTRIES to warn of fraud. These meta-classifiers can pick up patterns of fraud that are not detectable at the local level.



The green learning agent digs into the green bank's data and creates the green fraud detector

Similarly, the red learning agent digs into the red bank's data and creates the red fraud detector. The fraud detectors are capable of inspecting and classifying new transactions as valid or fraudulent



The green and red fraud detectors migrate to the blue bank where they ally with the local blue fraud detector. The black meta-learning agent combines the blue bank's data and the predictions of the colored fraud detectors and creates the meta-fraud detector. The meta-fraud detector inspects all incoming transactions with global knowledge and hence increased accuracy