KD Ubiq Summer School 2008
Behavioural Modelling of a Grid System

Michele Sebag
CNRS — INRIA — Université Paris-Sud
http://tao.lri.fr

March 8th, 2008

Overview of the Tutorial

Autonomic Computing

- ML & DM for Systems:
  Introduction, motivations, applications
- Zoom on an application: Performance management

Autonomic Grid

- EGEE: Enabling Grids for e-Science in Europe
- Data acquisition, Logging and Bookkeeping files
- (change of) Representation, Dimensionality reduction

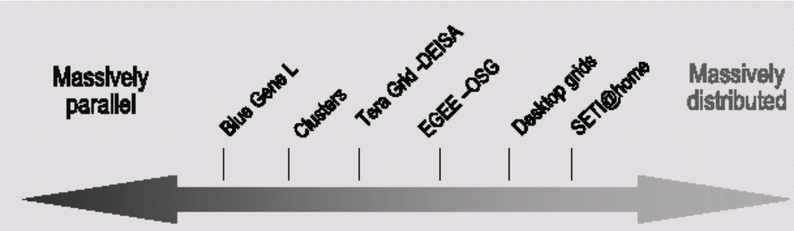
Modelling Jobs

- Exploratory Analysis and Clustering
- Standard approaches, stability, affinity propagation

Part 2

- Grid Systems
  Presentation of EGEE, Enabling Grids for e-Science in Europe
- Acquiring the data
  The grid observatory
- Preparation of the data
  - Functional dependencies
  - Dimensionality reduction
  - Propositionalization

Computing Systems: The landscape



parallel

- homogeneous soft and hard
- resources
  - dedicated
  - static
  - controlled
- reduced software stack
- no built-in fault tolerance

distributed

- heterogeneous soft and hard
- resources
  - shared
  - dynamic
  - aggregated
- middleware
- faults: the norm

Storage and Computation have to be distributed

EGEE: Enabling Grids for E-Science in Europe

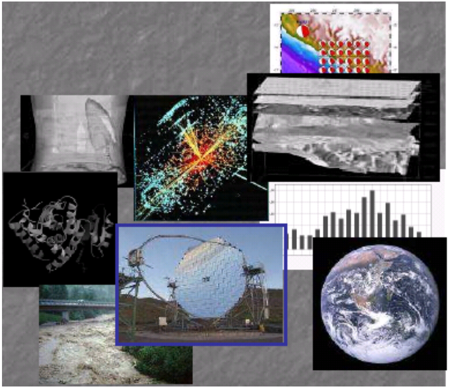
EGEE, 2

- Infrastructure project started in 2001 $\rightarrow$ FP6 and FP7
- Large scale, production quality grid
- Core node: Lab. Accelerateur Linéaire, Université Paris-Sud
- 240 partners, 41,000 CPUs, all over the world
- 5 Peta bytes storage
- 24 $\times$ 7, 20 K concurrent jobs
- Web: www.eu-egee.org

Storage as important as CPU

Applications

- High energy physics
- Life sciences
- Astrophysics
- Computational chemistry
- Earth sciences
- Financial simulation
- Fusion
- Multimedia
- Geophysics



Autonomic Grid

Requisite: The Grid Observatory

- Cluster in the EGEE-III proposal 2008-2010
- Data collection and publication: filtering, clustering

Workload management

- Models of the grid dynamics
- Models of requirements and middleware reaction: time series and beyond
- Utility based-scheduling, local and global: MAB problem
- Policy evaluations: very large scale optimization

Fault detection and diagnosis

- Categorization of failure modes from the Logging and Bookkeeping: feature construction, clustering,
- Abrupt changepoint detection

Autonomic Grid: The Grid Observatory

Data acquisition

- Data have not been stored with DM in mind — never
- Data [partially] automatically generated for EGEE services — here
  - redundant
  - little expert help

It's no longer: the expert feeds the machine with data. Rather, machines feed machines... — *J. Gama*

Data preprocessing

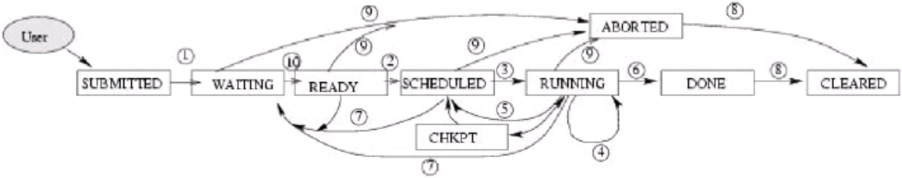
- 80% of the human cost
- Governs the quality of the output

The grid system and the data

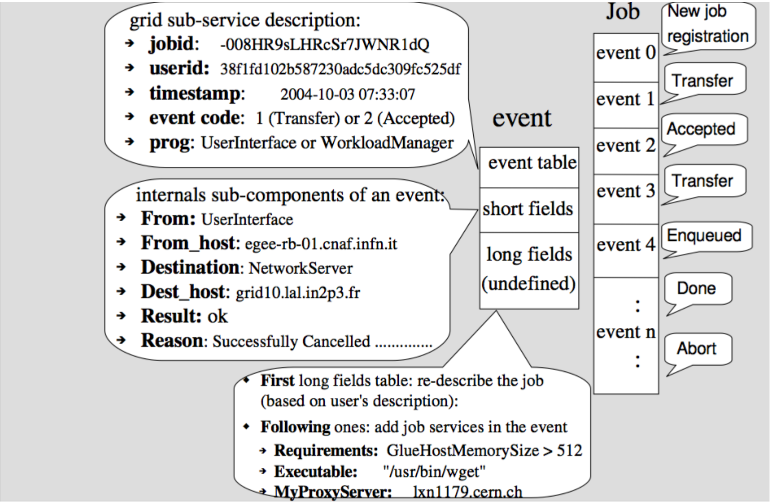
The Workload Management System

- User Interface — User submits job description and requirements, and gets the results
- Resource Broker — Decides Computing Element
- Job Submission Service — Submits to CE and Checks
- Logging and Bookkeeping Service — Archive the data

Job Lifecycle



The data



grid sub-service description:
jobid: -008HR9sLHRcSr7JWNR1dQ
userid: 38f1fd102b587230adc5dc309fc525df
timestamp: 2004-10-03 07:33:07
event code: 1 (Transfer) or 2 (Accepted)
prog: UserInterface or WorkloadManager
internals sub-components of an event:
From: UserInterface
From_host: egee-rb-01.cnaf.infn.it
Destination: NetworkServer
Dest_host: grid10.lal.in2p3.fr
Result: ok
Reason: Successfully Cancelled ....
event
event table
short fields
long fields (undefined)
First long fields table: re-describe the job (based on user's description):
Following ones: add job services in the event
Requirements: GlueHostMemorySize > 512
Executable: "/usr/bin/wget"
MyProxyServer: lxn1179.cern.ch
Job
event 0
event 1
event 2
event 3
event 4
:
event n
:
New job registration
Transfer
Accepted
Transfer
Enqueued
Done
Abort

Data Tables

Events

jobid	event	code	host	time_stamp	arrived	level
---BrI1BgbIqkwtszqGfmA	0	17	atlfarm008.mi.infn.it	2004-09-17 16:17:48	2004-09-17 16:17:49	8
---BrI1BgbIqkwtszqGfmA	1	1	atlfarm008.mi.infn.it	2004-09-17 16:17:48	2004-09-17 16:17:49	8
---BrI1BgbIqkwtszqGfmA	2	2	lxb0728.cern.ch	2004-09-17 16:17:53	2004-09-17 16:17:53	8
---BrI1BgbIqkwtszqGfmA	3	4	lxb0728.cern.ch	2004-09-17 16:18:00	2004-09-17 16:18:01	8
---BrI1BgbIqkwtszqGfmA	4	1	atlfarm008.mi.infn.it	2004-09-17 16:18:00	2004-09-17 16:18:01	8
---BrI1BgbIqkwtszqGfmA	5	5	lxb0728.cern.ch	2004-09-17 16:18:01	2004-09-17 16:18:01	8

Short Fields

0	JOBTYPE	SIMPLE
0	NS	lxb0728.cern.ch:7772
0	NSUBJOBS	0
0	SEED	uLUOBArrdV98041PLThJ5Q
0	SEQCODE	UI=000001:NS=0000000000:WM=000000:BH=0000000000:JSS=000000:LM=000000:LRMS=000000:APP=000000
0	SRC_INSTANCE	
1	DESTINATION	NetworkServer
1	DEST_HOST	lxb0728.cern.ch
1	DEST_INSTANCE	lxb0728.cern.ch:7772
1	DEST_JOBID	
1	REASON	
1	RESULT	START
1	SEQCODE	UI=000002:NS=0000000000:WM=000000:BH=0000000000:JSS=000000:LM=000000:LRMS=000000:APP=000000
1	SRC_INSTANCE	
2	FROM	UserInterface
2	FROM_HOST	lxb0728.cern.ch
2	FROM_INSTANCE	
2	LOCAL_JOBID	
2	SEQCODE	UI=000003:NS=0000000001:WM=000000:BH=0000000000:JSS=000000:LM=000000:LRMS=000000:APP=000000
2	SRC_INSTANCE	7772
3	QUEUE	/var/edgwl/workload_manager/input.fl
3	REASON	
3	RESULT	OK
3	SEQCODE	UI=000003:NS=0000000003:WM=000000:BH=0000000000:JSS=000000:LM=000000:LRMS=000000:APP=000000
3	SRC_INSTANCE	

Data Tables

Long Fields (4Gb)

```
+------------------------+-------+------
| jobid                  | event | name | value
+------------------------+-------+------
| ---BrI1BgbIqkwtszqGfmA |     0 | JDL  |[ requirements = ( ( ( ( Member("VO-atlas-lcg-release
-0.0.2",other.GlueHostApplicationSoftwareRunTimeEnvironment) ) && Member("VO-atlas-release
-8.0.5",other.GlueHostApplicationSoftwareRunTimeEnvironment) ) && ( other.GlueCEPolicyMaxCPUTime >= ( Member("LCG
-2\_1_0",other.GlueHostApplicationSoftwareRunTimeEnvironment) ? ( 36000000 / 60 ) : 36000000 ) / other.GlueHostBenchmarkSI00 ) ) &&
(other.GlueHostNetworkAdapterOutboundIP == true ) ) && ( other.GlueHostMainMemoryRAMSize >= 512 ); RetryCount = 0; edg_jobid =
 "https://lxb0728.cern.ch:9000/---BrI1BgbIqkwtszqGfmA"; Arguments = "dc2.003048.evgen.H4_170_WW._00002.pool.root
dc2.003048.simul.H4_170_WW._00208.pool.root.2 -6 6 50 350 208"; Environment = {
"LEXOR_WRAPPER_LOG=lexor_wrapper.log","LEXOR_STAGEOUT_MAXATTEMPT=5","LEXOR_STAGEOUT_INTERVAL=60",
"LEXOR_LCG_GFAL_INFOSYS=lxb2011.cern.ch:2170","LEXOR_T_RELEASE=8.0.5",
"LEXOR_T_PACKAGE=8.0.5.6/JobTransforms","LEXOR_T_BASEDIR=JobTransforms-08-00-05-06",
"LEXOR_TRANSFORMATION=share/
dc2.g4sim.trf","LEXOR_STAGEIN_LOG=dq_233387_stagein.log","LEXOR_STAGEIN_SCRIPT=dq_233387_stagein.sh",
"LEXOR_STAGEOUT_LOG=dq_233387_stageout.log","LEXOR_STAGEOUT_SCRIPT=dq_233387_stageout.sh" };
 MyProxyServer = "lxb0727.cern.ch"; JobType = "normal"; Executable =
"lexor_wrap.sh"; StdOutput = "dc2.003048.simul.H4_170_WW._00208.job.log.2"; OutputSandbox = {
"metadata.xml","lexor_wrapper.log","dq_233387_stagein.log","dq_233387_stageout.log",
"dc2.003048.simul.H4_170_WW._00208.job.log.2" }; VirtualOrganisation = "atlas";
 rank = ( other.GlueCEStateEstimatedResponseTime > 999 ) ?  -( other.GlueCEStateEstimatedResponseTime ) :  -(
other.GlueCEStateRunningJobs ); Type = "job"; StdError = "dc2.003048.simul.H4_170_WW._00208.job.log.2";
DefaultRank =  -other.GlueCEStateEstimatedResponseTime;
InputSandbox = {
"/home/negri/windmill-0.9.15/lexor/inputsandbox/lexor_wrap.sh",
"/home/negri/windmill-0.9.15/lexor/inputsandbox/dqlcg.py",
"/home/negri/windmill-0.9.15/lexor/inputsandbox/edgrmpi.sh",
"/home/negri/windmill-0.9.15/lexor/inputsandbox/dqrep.pl",
"/home/negri/windmill-0.9.15/lexor/inputsandbox/run_dqlcg.sh","/tmp/lexor/negri/dq_233387_stagein.sh",
"/tmp/lexor/negri/dq_233387_stageout.sh" } ]
+------------------------+-------+------
```

Preparation of the data

1. Functional dependencies
2. Dimensionality reduction curse of dimensionality
   - Principal Component Analysis
   - Random Projection
   - Non linear Dimensionality Reduction
3. Propositionalization

Functional dependency

Definition

Given attributes X and X', X' depends on X on $\mathcal{E}$ ($X' \prec X$) iff

$$\exists f : dom(X') \mapsto dom(X) \; s.t. \; \forall i = 1 \ldots N, X(\mathbf{x}_i) = f(X'(\mathbf{x}_i))$$

Examples

- X' = City code, X = City name
- X' = Machine name, X = IP
- X' = Job ID, X = User ID

Why removing FD ?

- Curse of dimensionality
- Biased distance

Functional dependency, 2

Trivial cases

$$\#dom(X) = \#dom(X') = N \text{ number of examples}$$

Algorithm

- Size:
$$(X' \prec X) \Rightarrow \#dom(X) \leq \#dom(X')$$
- Sample

  Repeat

  Select $v \in dom(X')$

  $\mathcal{E}_v$ = select $\mathbf{x}_i$ where $X'(\mathbf{x}_i) = v$

  Define $X(\mathcal{E}_v) = \{w \in dom(X), \exists x \in \mathcal{E}_v \; / \; X(x) = w\}$

  If $(\#X(\mathcal{E}_v) > 1)$ return false

  Until stop

  return true

Dimensionality Reduction – Intuition

Degrees of freedom

- Image: 4096 pixels; but not independent
- Robotics: (# camera pixels + # infra-red) $\times$ time; but not independent

Goal

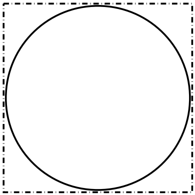
Find the (low-dimensional) structure of the data:

- Images
- Robotics
- Genes

Dimensionality Reduction

In high dimensions

- Everybody lives in the corners of the space
  Volume of Sphere $V_n = \frac{2\pi r^2}{n} V_{n-2}$
- All points are far from each other



Approaches

- Linear dimensionality reduction
  - Principal Component Analysis
  - Random Projection
- Non-linear dimensionality reduction

Criteria

- Complexity/Size
- Prior knowledge e.g., relevant distance

Linear Dimensionality Reduction

Training set *unsupervised*

$$\mathcal{E} = \{(\mathbf{x}_k), \mathbf{x}_k \in \mathbb{R}^D, k = 1 \ldots N\}$$

Projection from $\mathbb{R}^D$ onto $\mathbb{R}^d$

$$\mathbf{x} \in \mathbb{R}^D \rightarrow \quad h(\mathbf{x}) \in \mathbb{R}^d, \; d << D$$

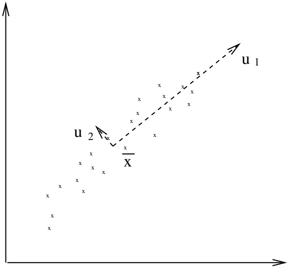
$$h(\mathbf{x}) = A\mathbf{x}$$

$$s.t. \text{ minimize} \quad \sum_{k=1}^{N} ||\mathbf{x}_k - h(\mathbf{x}_k)||^2$$

Principal Component Analysis

Covariance matrix S

Mean $\qquad \mu_i = \frac{1}{N}\sum_{k=1}^{N} X_i(\mathbf{x}_k)$



$$S_{ij} = \frac{1}{N}\sum_{k=1}^{N}(X_i(\mathbf{x}_k) - \mu_i)(X_j(\mathbf{x}_k) - \mu_j)$$

symmetric $\Rightarrow$ can be diagonalized

$$S = U\Delta U' \qquad \Delta = Diag(\lambda_1, \ldots \lambda_D)$$

Thm: Optimal projection in dimension d

projection on the first d eigenvectors of S

Let u_i the eigenvector associated to eigenvalue λ_i $\qquad \lambda_i > \lambda_{i+1}$

$$h : \mathbb{R}^D \mapsto \mathbb{R}^d, h(\mathbf{x}) = < \mathbf{x}, u_1 > u_1 + \ldots + < \mathbf{x}, u_d > u_d$$

Sketch of the proof

1. Maximize the variance of $h(\mathbf{x}) = A\mathbf{x}$

$\sum_k ||\mathbf{x}_k - h(\mathbf{x}_k)||^2 = \sum_k ||\mathbf{x}_k||^2 - \sum_k ||h(\mathbf{x}_k)||^2$

$$\text{Minimize } \sum_k ||\mathbf{x}_k - h(\mathbf{x}_k)||^2 \Rightarrow \text{ Maximize } \sum_k ||h(\mathbf{x}_k)||^2$$

$$Var(h(\mathbf{x})) = \frac{1}{N}\left(\sum_k ||h(\mathbf{x}_k)||^2 - ||\sum_k h(\mathbf{x}_k)||^2\right)$$

As

$$||\sum_k h(\mathbf{x}_k)||^2 = ||A\sum_k \mathbf{x}_k||^2 = N^2||A\mu||^2$$

where $\mu = (\mu_1, \ldots .\mu_D)$.
Assuming that $\mathbf{x}_k$ are centered ($\mu_i = 0$) gives the result.

Sketch of the proof, 2

2. Projection on eigenvectors u_i of S

Assume $h(\mathbf{x}) = A\mathbf{x} = \sum_{i=1}^{d} < \mathbf{x}, v_i > v_i$ and show $v_i = u_i$.

$$Var(AX) = (AX)(AX)' = A(XX')A' = ASA' = A(U\Delta U')A'$$

Consider $d = 1,\ v_1 = \sum w_i u_i$ $\qquad \sum w_i^2 = 1$

remind $\lambda_i > \lambda_{i+1}$

$$Var(AX) = \sum \lambda_i w_i^2$$

maximized for $w_1 = 1, w_2 = \ldots = w_N = 0$
that is, $v_1 = u_i$.

Principal Component Analysis, Practicalities

Data preparation

- Mean centering the dataset

$$
\begin{aligned}
\mu_i &= \tfrac{1}{N}\sum_{k=1}^{N} X_i(\mathbf{x}_k)\\
\sigma_i &= \sqrt{\tfrac{1}{N}\sum_{k=1}^{N} X_i(\mathbf{x}_k)^2 - \mu_i^2}\\
z_k &= \left(\tfrac{1}{\sigma_i}(X_i(\mathbf{x}_k) - \mu_i)\right)_{i=1}^{D}
\end{aligned}
$$

Matrix operations

- Computing the covariance matrix

$$
S_{ij} = \frac{1}{N}\sum_{k=1}^{N} X_i(z_k)X_j(z_k)
$$

- Diagonalizing $S = U'\Delta U$ might be not affordable... Complexity $\mathcal{O}(D^3)$

Random projection

Random matrix

$$A : \mathbb{R}^D \mapsto \mathbb{R}^d \quad A[d, D] \quad A_{i,j} \sim \mathcal{N}(0, 1)$$

define

$$h(\mathbf{x}) = \frac{1}{\sqrt{d}} A\mathbf{x}$$

Property: h preserves the norm in expectation

$$E[||h(\mathbf{x})||^2] = ||\mathbf{x}||^2$$

With high probability $1 - 2exp\{-(\varepsilon^2 - \varepsilon^3)\frac{d}{4}\}$

$$(1 - \varepsilon)||\mathbf{x}||^2 \leq ||h(\mathbf{x})||^2 \leq (1 + \varepsilon)||\mathbf{x}||^2$$

Random projection

Proof

$$h(\mathbf{x}) = \frac{1}{\sqrt{d}} A\mathbf{x}$$

$$\begin{aligned}
E(||h(\mathbf{x})||^2) &= \frac{1}{d} E\left[\sum_{i=1}^{d} \left(\sum_{j=1}^{D} A_{i,j} X_j(\mathbf{x})\right)^2\right] \\
&= \frac{1}{d} \sum_{i=1}^{d} E\left[\left(\sum_{j=1}^{D} A_{i,j} X_j(\mathbf{x})\right)^2\right] \\
&= \frac{1}{d} \sum_{i=1}^{d} \sum_{j=1}^{D} E[A_{i,j}^2] E[X_j(\mathbf{x})^2] \\
&= \frac{1}{d} \sum_{i=1}^{d} \sum_{j=1}^{D} \frac{||\mathbf{x}||^2}{D} \\
&= ||\mathbf{x}||^2
\end{aligned}$$

Random projection, 2

Johnson Lindenstrauss Lemma

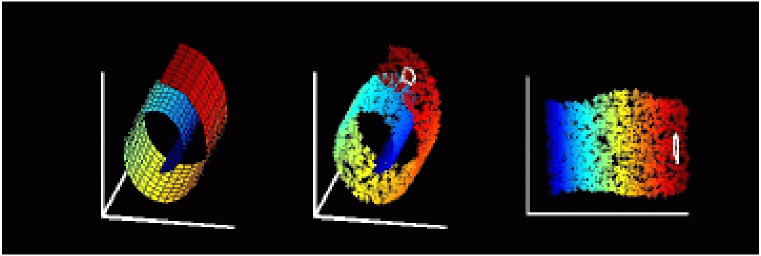
For $d > \frac{9 \ln N}{\varepsilon^2 - \varepsilon^3}$, with high probability

$$(1-\varepsilon)||\mathbf{x}_i - \mathbf{x}_j||^2 \leq ||h(\mathbf{x}_i) - h(\mathbf{x}_j)||^2 \leq (1+\varepsilon)||\mathbf{x}_i - \mathbf{x}_j||^2$$

More:

`http://www.cs.yale.edu/clique/resources/RandomProjectionMethod.pdf`

Non-Linear Dimensionality Reduction



Conjecture

Examples live in a manifold of dimension $d << D$

Goal: consistent projection of the dataset onto $\mathbb{R}^d$

Consistency:

- Preserve the structure of the data
- e.g. preserve the distances between points

Multi-Dimensional Scaling

Position of the problem

- Given $\{\mathbf{x}_1, \ldots, \mathbf{x}_N, \ \mathbf{x}_i \in \mathbb{R}^D\}$
- Given $sim(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}^+$
- Find projection Φ onto $\mathbb{R}^d$

$$\begin{array}{rl} x \in \mathbb{R}^D \rightarrow & \Phi(x) \in \mathbb{R}^d \\ sim(\mathbf{x}_i, \mathbf{x}_j) \sim & sim(\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)) \end{array}$$

Optimisation

Define $X, \ X_{i,j} = sim(\mathbf{x}_i, \mathbf{x}_j); \ X^\Phi, \ X^\Phi_{i,j} = sim(\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j))$

Find Φ minimizing $||X - X'||$

Rq : Linear Φ = Principal Component Analysis

But linear MDS does not work: preserves all distances, while

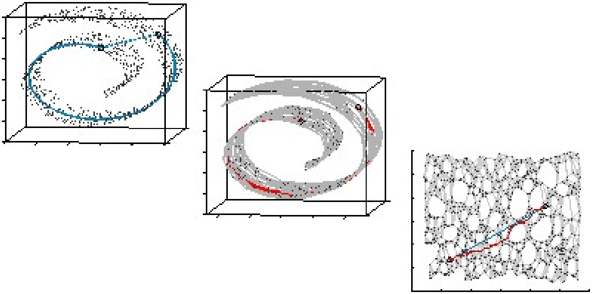
only *local* distances are meaningful

Non-linear projections

Approaches

- Reconstruct global structures from local ones and find global projection — Isomap
- Only consider local structures — LLE

Intuition: locally, points live in $\mathbb{R}^d$



Isomap

Tenenbaum, da Silva, Langford 2000
http://isomap.stanford.edu

Estimate $d(x_i, x_j)$

- Known if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close
- Otherwise, compute the shortest path between $\mathbf{x}_i$ and $\mathbf{x}_j$
  geodesic distance (dynamic programming)

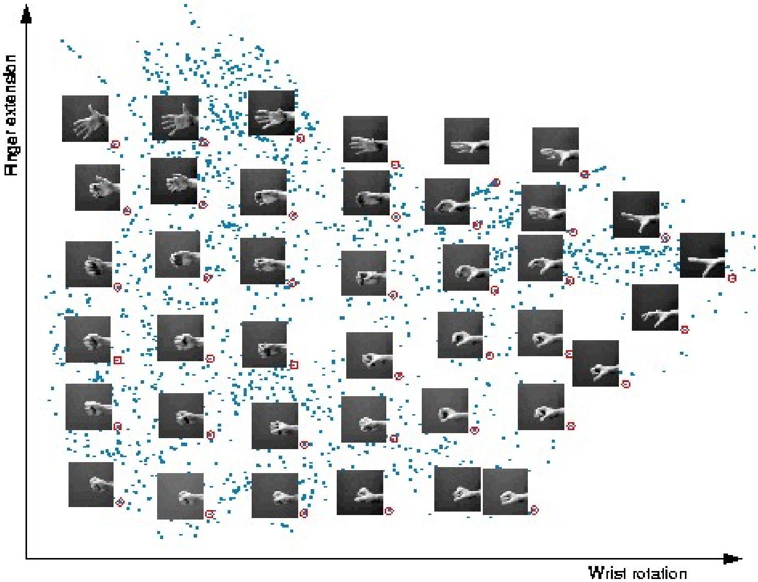
Requisite

If data points sampled in a convex subset of $\mathbb{R}^d$,
then geodesic distance $\sim$ Euclidean distance on $\mathbb{R}^d$.

General case

- Given $d(\mathbf{x}_i, \mathbf{x}_j)$, estimate $< \mathbf{x}_i, \mathbf{x}_j >$
- Project points in $\mathbb{R}^d$

Isomap, 2



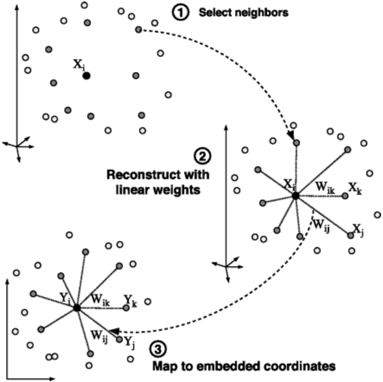
Locally Linear Embedding

Roweiss and Saul, 2000

http://www.cs.toronto.edu/∼roweis/lle/

Principle

- Find local description for each point: depending on its neighbors



Local Linear Embedding, 2

Find neighbors

For each $\mathbf{x}_i$, find its nearest neighbors $\mathcal{N}(i)$

Parameter: number of neighbors

Change of representation

Goal Characterize $\mathbf{x}_i$ wrt its neighbors:

$$\mathbf{x}_i = \sum_{j \in \mathcal{N}(i)} w_{i,j} \mathbf{x}_j \quad \text{with} \sum_{j \in \mathcal{N}(i)} w_{ij} = 1$$

Property: invariance by translation, rotation, homothety

How Compute the local covariance matrix:

$$C_{j,k} = < x_j - x_i, x_k - x_i >$$

Find vector w_i s.t. $Cw_i = 1$

Local Linear Embedding, 3

Algorithm

Local description: Matrix W such that $\sum_j w_{i,j} = 1$

$$W = argmin\{\sum_{i=1}^{N} ||\mathbf{x}_i - \sum_j w_{i,j}\mathbf{x}_j||^2\}$$

Projection: Find $\{z_1, \ldots, z_n\}$ in $\mathbb{R}^d$ minimizing

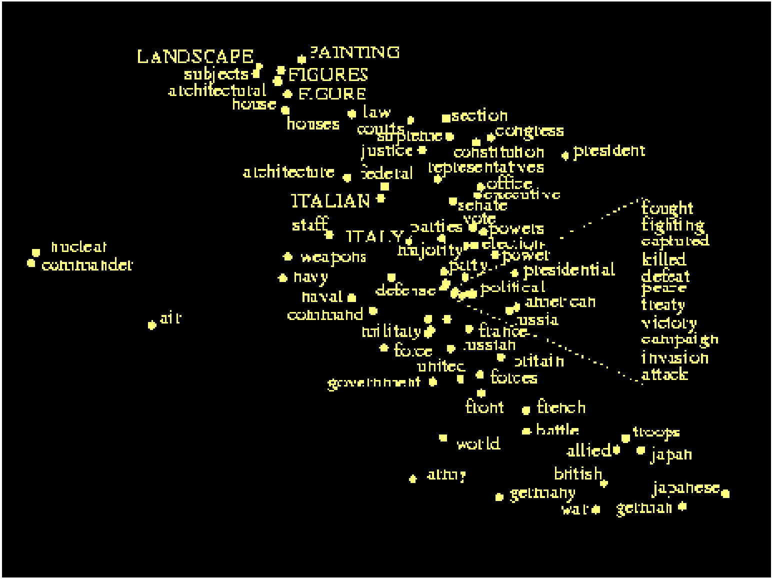
$$\sum_{i=1}^{N} ||z_i - \sum_j w_{i,j} z_j||^2$$

$$Minimize \ ((I - W)Z)'((I - W)Z) = Z'(I - W)'(I - W)Z$$

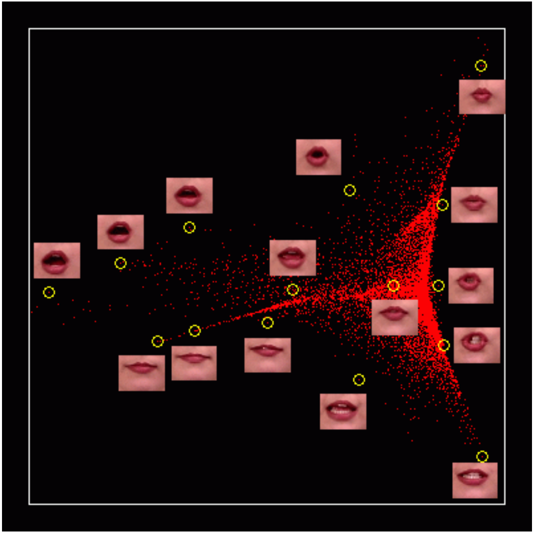
Solutions: vectors z_i are eigenvectors of $(I - W)'(I - W)$

- Keeping the d eigenvectors with lowest eigenvalues > 0

Example, Texts



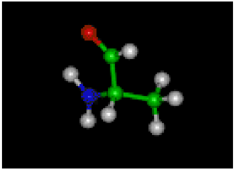
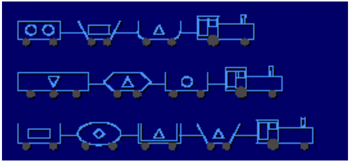
Example, Images



LLE

Propositionalization

Relational domains

Relational learning

PROS — Inductive Logic Programming

- Use domain knowledge

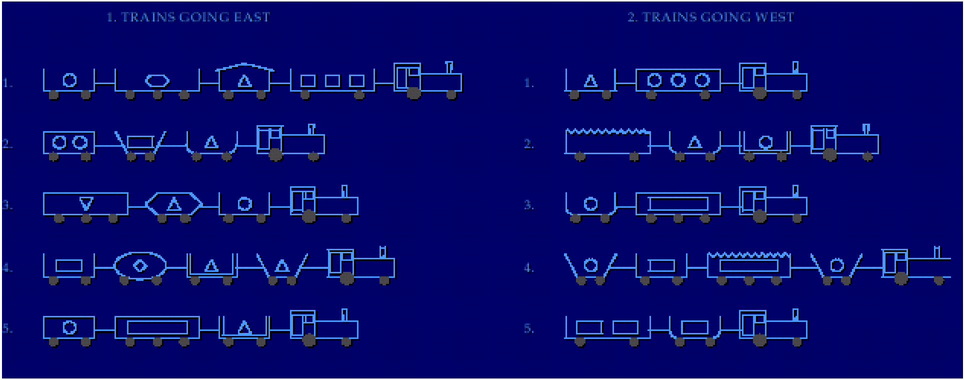
CONS — Data Mining

- Covering test $\equiv$ subgraph matching — exponential complexity

Getting back to propositional representation: propositionalization

West - East trains

Michalski 1983



Propositionalization

Linus (ancestor)

Lavrac et al, 94

$$West(a) \leftarrow \quad Engine(a,b), first\_wagon(a,c), roof(c), load(c, square, 3)\ldots$$
$$West(a') \leftarrow \quad Engine(a',b'), first\_wagon(a',c'), load(c', circle, 1)\ldots$$

West	Engine(X)	First Wagon(X,Y)	Roof(Y)	$Load_1$ (Y)	$Load_2$ (Y)
a	b	c	yes	square	3
a'	b'	c'	no	circle	1

Each column: a role predicate, where the predicate is determinate linked to former predicates (left columns) with a single instantiation in every example

Propositionalization

Stochastic propositionalization

Kramer, 98

Construct random formulas $\equiv$ boolean features

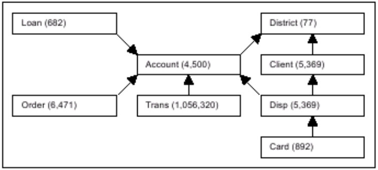
SINUS – RDS

http://www.cs.bris.ac.uk/home/rawles/sinus

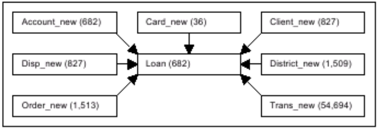
http://labe.felk.cvut.cz/∼zelezny/rsd

- Use modes (user-declared) `modeb(2,hasCar(+train,-car))`
- Thresholds on number of variables, depth of predicates...
- Pre-processing (feature selection)

Propositionalization



DB Schema



Propositionalization

RELAGGS

Database aggregates

- average, min, max, of numerical attributes
- number of values of categorical attributes

Going ubiquitous in Data Preparation

Principles: same as usual

- Act locally
- Think globally

The local level

- An ideal feature $\equiv$ a good hypothesis
- What is a promising hypothesis ?
  - Behaves well on (part of) the data
  - Is not trivial

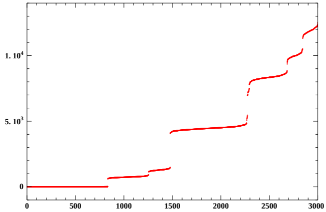
Going ubiquitous in Data Preparation, 2

What is a good behaviour?

- Showing regularities
- Locally constant

How to test triviality?

- Syntactical analysis:
  $xy - yx = 0$
- Statistical triviality:
  - Test on random data
  - Test on permutations of the data



Going ubiquitous in Data Preparation, 3

Internally: an optimization problem

- Define bins
- Compute histogram, associated quantity of information
- Compare histograms on real data / on random data

Externally: an optimization problem

- Upon receiving a new feature
- Check whether this is relevant to your data
- Check whether this brings new information