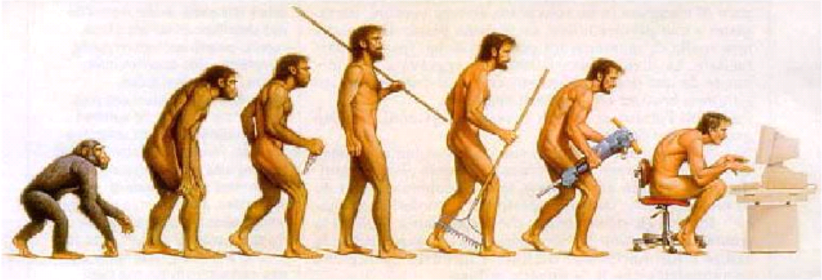
Module Master Recherche Apprentissage
Autonomic Computing – Analyse Exploratoire

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

14 Janvier 2008

Autonomic Computing



Considering current technologies, we expect that the total number of device administrators will exceed 220 millions by 2010.

Gartner 6/2001

in Autonomic Computing Wshop, ECML / PKDD 2006

Irina Rish & Gerry Tesauro.

Autonomic Computing

The need

- Main bottleneck of the deployment of complex systems: shortage of skilled administrators

Vision

- Computing systems take care of the mundane elements of management by themselves.
- Inspiration: central nervous system (regulating temperature, breathing, and heart rate without conscious thought)

Goal

Computing systems that manage themselves in accordance with high-level objectives from humans

Kephart & Chess, IEEE Computer 2003

Autonomic Computing

Activity: A growing field

- IBM Manifesto for Autonomic Computing 2001
  http://www.research.ibm.com/autonomic
- ECML/PKDD Wshop on Autonomic Computing 2006
  http://www.ecmlpkdd2006.org/workshops.html
- JIC. on Measurement and Performance of Systems 2006
  http://www.cs.wm.edu/sigm06/
- NIPS Wshop on Machine Learning for Systems 2007
  http://radlab.cs.berkeley.edu/MLSys/
- Networked System Design and Implementation 2008
  http://www.usenix.org/events/nsdi08/

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

ML & DM for Systems

Some applications

- Cohen et al., OSDI 2004, Performance management

  detailed next
- Palatin-Wolf-Schuster, KDD06. Find misconfigured CPUs in a grid system

  find outliers
- Xiao et al. AAAI05, Active learning for game player modeling

  situations where it's too easy
- Zheng et al. NIPS03-ICML06, Use traces to identify bugs

  put probes, suggest causes for failures
- Baskiotis et al., IJCAI07, ILP07, Statistical Structural Software Testing

  construct test cases for software testing

Performance management

The goal

Ensure that the system complies with performance level objectives

The problem: System Modelling

Large-scale system complex behavior depends on:

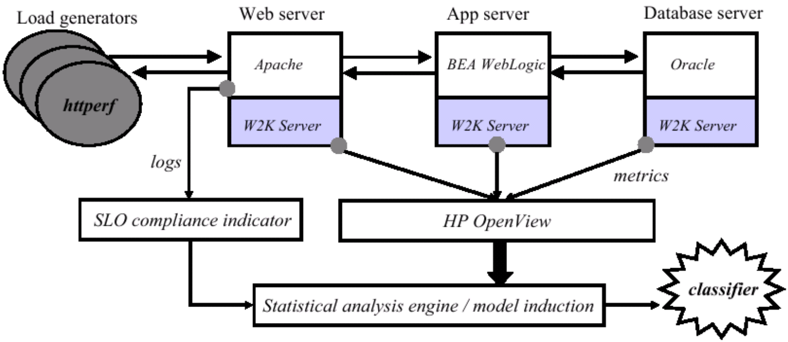
- Workload
- Software structure
- Hardware
- Traffic
- System goals

The approaches

- Prior knowledge — set of (event - condition - action) rules
- Statistical learning

exploiting pervasive instrumentation / query facilities

Example: a 3-tier Web application with a Java middleware component, backed by a DB



Correlating instrumentation data to system states: A building block for automated diagnosis and control, Cohen et al. OSDI 2004

Supervised Learning, Notations

Training set, set of examples, data base

(iid sample $\sim P(\mathbf{x}, y)$)

$$\mathcal{E} = \{(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1 \ldots N\}$$

- $\mathcal{X}$: Instance space
  - propositional (examples described after D attributes) $\mathbb{R}^D$

    $$\mathbf{x} = (X_1(\mathbf{x}), \ldots X_D(\mathbf{x}))$$

  - relational (examples described after objects in relation, e.g. events - see later on)
- $\mathcal{Y}$: Label space
  - Discrete: classification (compliant, not-compliant)
  - Continuous: regression (average response time)

Example

Instance space, set of attributes

Metric	Description
mean_AS_CPU_1_USERTIME	CPU time spent in user mode on the application server.
var_AS_CPU_1_USERTIME	Variance of user CPU time on the application server.
mean_AS_DISK_1_PHYSREAD	Number of physical disk reads for disk 1 on the application server, includes file system reads, raw I/O and virtual memory I/O.
mean_AS_DISK_1_BUSYTIME	Time in seconds that disk 1 was busy with pending I/O on the application server.
var_AS_DISK_1_BUSYTIME	Variance of time that disk 1 was busy with pending I/O on the application server.
mean_DB_DISK_1_PHYSWRITEBYTE	Number of kilobytes written to disk 1 on the database server, includes file system reads, raw I/O and virtual memory I/O.
var_DB_GBL_SWAPSPACEUSED	Variance of swap space allocated on the database server.
var_DB_NETIF_2_INPACKET	Variance of the number of successful (no errors or collisions) physical packets received through network interface #2 on the database server.
mean_DB_GBL_SWAPSPACEUSED	Amount of swap space, in MB, allocated on the database server.
mean_DB_GBL_RUNQUEUE	Approximate average queue length for CPU on the database server.
var_DB_NETIF_2_INBYTE	Variance of the number of KBs received from the network via network interface #2 on the database server. Only bytes in packets that carry data are included.
var_DB_DISK_1_PHYSREAD	Variance of physical disk reads for disk 1 on the database server.
var_AS_GBL_MEMUTIL	Variance of the percentage of physical memory in use on the application server, including system memory (occupied by the kernel), buffer cache, and user memory.
numReqs	Number of requests the system has served.
var_DB_DISK_1_PHYSWRITE	Variance of the number of writes to disk 1 on the database server.
var_DB_NETIF_2_OUTPACKET	Variance of the number of successful (no errors or collisions) physical packets sent through network interface #2 on the database server.

Label space

Compliance with Service Level Objectives (SLO) YES / NO

Learning a model

Desiderata

Efficient	few prediction errors
Compact	fast to use on further cases
Easy/Fast to train	no expertise needed to use
Interpretable	guide design/improvement

Learning – Hypothesis search space

Learning = finding h with good quality

$$h \in \mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$$

Loss function

$$\ell(y, y') = \text{ Cost of predicting } y' \text{ instead of } y$$

- $\ell(y, y') = 1_{[y=y']}$ classification
- $\ell(y, y') = (y - y')^2$ regression

Learning – Hypothesis search space, 2

Learning criterion

- Generalization error (ideal, alas $P(x, y)$ is unknown)

$$Err_{gen}(h) = E[\ell(y, h(\mathbf{x}))] = \int \ell(y, h(\mathbf{x}))dP(\mathbf{x}, y)$$

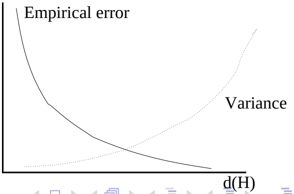
- Empirical error (known)

$$Err_{emp}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(x_i))$$

The bias/variance tradeoff

$d(\mathcal{H})$: dimension of Vapnik Cervonenkis

$$Err_{gen}(h) \leq Err_{emp}(h) + \mathcal{F}(n, d(\mathcal{H}))$$



Bayesian Learning

Bayes theorem

$$
\begin{aligned}
P(Y = y|X = \mathbf{x}) &= P(X = \mathbf{x}|Y = y).P(Y = y) \;/\; P(X = \mathbf{x}) \\
&\propto P(X = \mathbf{x}|Y = y).P(Y = y)
\end{aligned}
$$

Let $\mathbf{x} = (X_1(\mathbf{x}), \ldots, X_D(\mathbf{x})) \in \mathbb{R}^D$.
Assuming attributes are independent,

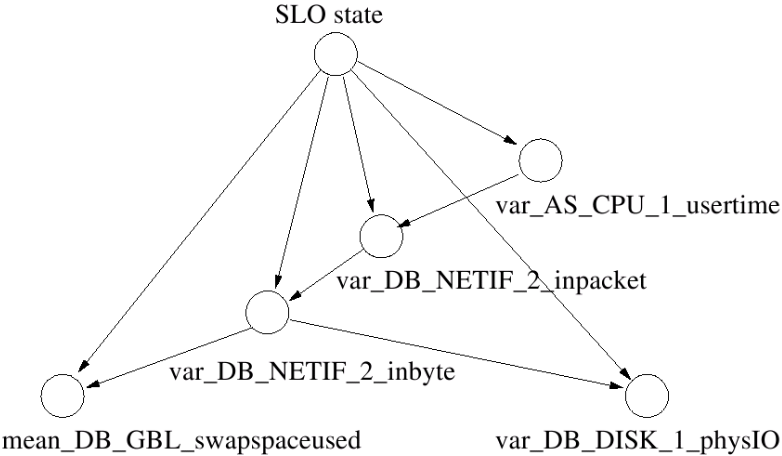
$$
P(X = \mathbf{x}|Y = y) = \prod_{i=1}^{d} P(X_i = X_i(\mathbf{x})|Y = y)
$$

Prediction: select class that maximizes the probability of $\mathbf{x}$

$$
\hat{y}(\mathbf{x}) = argmax\{\prod_{i=1}^{d} P(X_i = X_i(\mathbf{x})|Y = y_j).P(Y = y_j), y_j \in \mathcal{Y}\}
$$

Tree-Augmented Naive Bayes

Learn probability of attribute X_i conditionally to
* label Y;
* at most one other attribute X_j.



Tree-Augmented Naive Bayes, 2

Friedman, Geiger, Goldszmidt, MLJ 1997

Algorithm

- For each pair of attributes (X_i, X_j), compute $I(X_i, X_j) =$

$$\sum_{v_i, v_j, y} P(X_i = v_i, X_j = v_j, Y = y) \ln \frac{P(X_i = v_i, X_j = v_j | Y = y)}{P(X_i = v_i | Y = y)\, P(X_j = v_j | Y = y)}$$

- Define the complete graph $\mathcal{G}$ with $I(X_i, X_j)$ on edge (X_i, X_j)
- Define the maximum weight spanning tree from $\mathcal{G}$

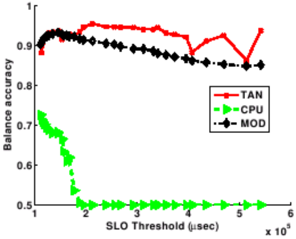
Complexity

D : number of attributes
N : number of examples
Complexity: $\mathcal{O}(D^2 N)$

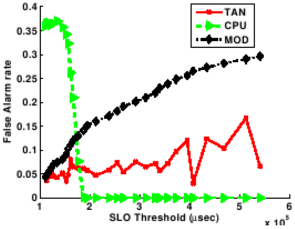
Results: 1. Accuracy

Balanced accuracy $= \frac{1}{2}$(True Pos. rate + True Neg rate).
Measured by 10 fold CV

Depending on performance threshold



Balanced accuracy



False alarm rate

- CPU: baseline predictor, use the CPU level only
- MOD: TAN trained with highest performance threshold
- TAN: TAN trained for each performance threshold

Results: 2. Using the model

Forecasting the failures

$$\ln \frac{P(X_{i,t+1} = v | X_{i,t} = v', Y = 0) P(Y = 0)}{P(X_{i,t+1} = v | X_{i,t} = v', Y = 1) P(Y = 1)} > 0$$

Interpreting the causes of failures

- Direct interpretation might be hindered by limited description.
- Learning would select an effect for a (missing) cause.
- Example: *minute-average-load* used as *disk queue* is missing.

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

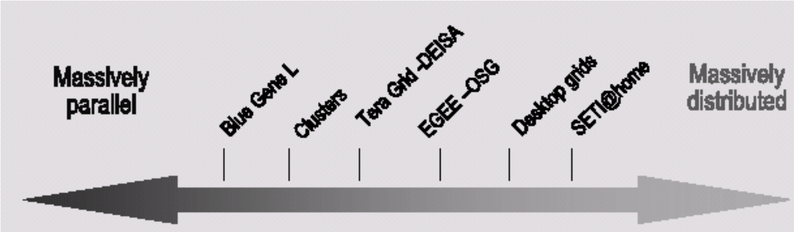
Modelling Jobs

- 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

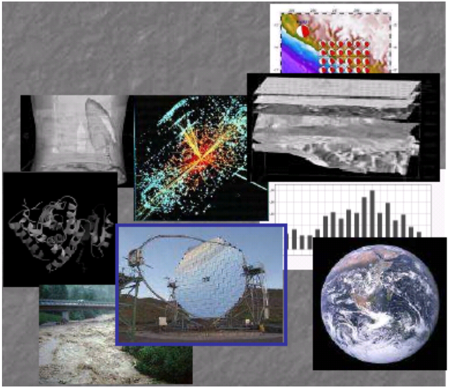
EGEE, 2

- 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
- 24×7, 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 — here
  for EGEE services
  - redundant
  - little expert help

Data preprocessing

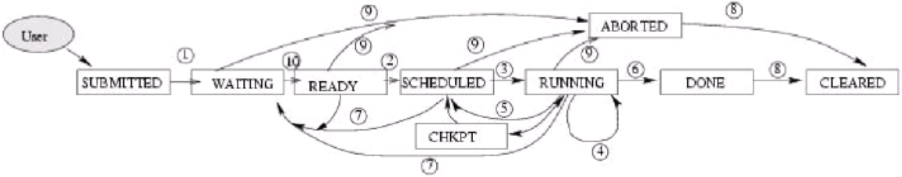
- 80% of the human cost
- Governs the quality of the output

The grid system and the data

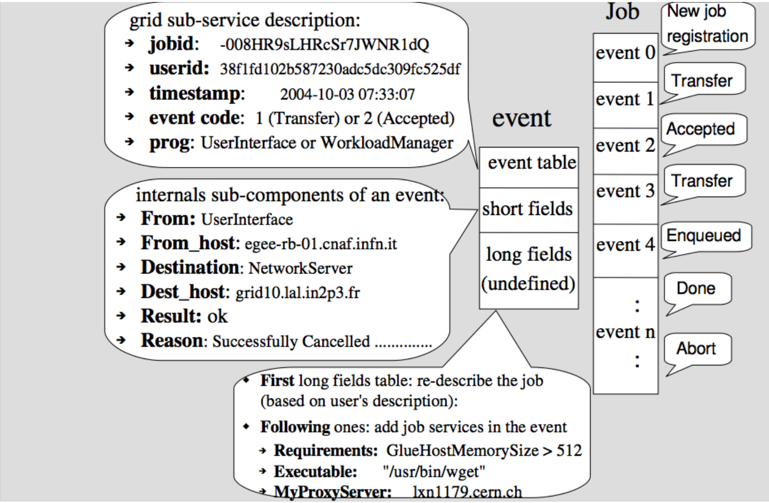
The Workload Management System

- 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

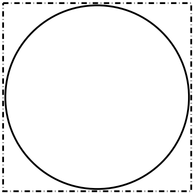
- Image: 4096 pixels; but not independent
- Robotics: (# camera pixels + # infra-red) $\times$ time; but not independent

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 $\mathbb{R}^d$

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

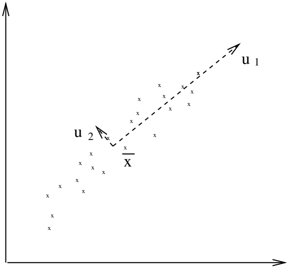
$$h(\mathbf{x}) = A\mathbf{x}$$

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

Principal Component Analysis

Covariance matrix S

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



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

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 $d = 1,\ v_1 = \sum w_i u_i$

$\sum w_i^2 = 1$
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, $v_1 = u_i$.

More :
`http://mplab.ucsd.edu/wordpress/tutorials/pca.pdf`

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

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

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

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

Random projection, 2

Johnson Lindenstrauss Lemma

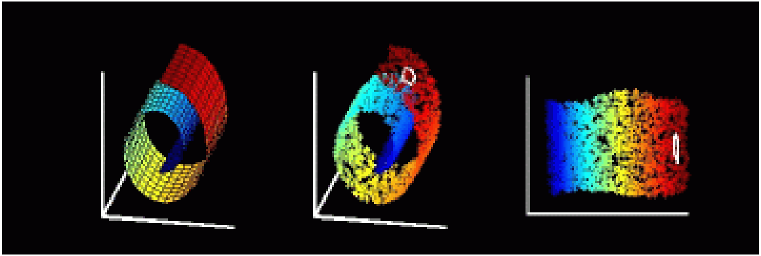
For $d > \frac{9 \ln D}{\varepsilon^2 - \varepsilon^3}$, with high probability

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

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

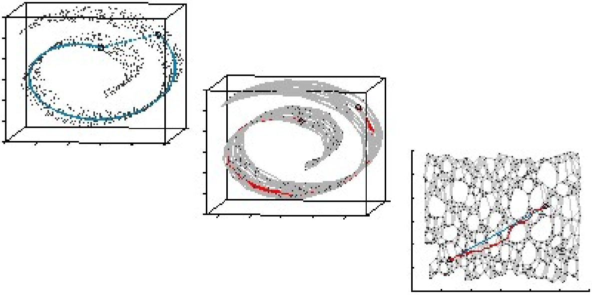
only *local* distances are meaningful

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)

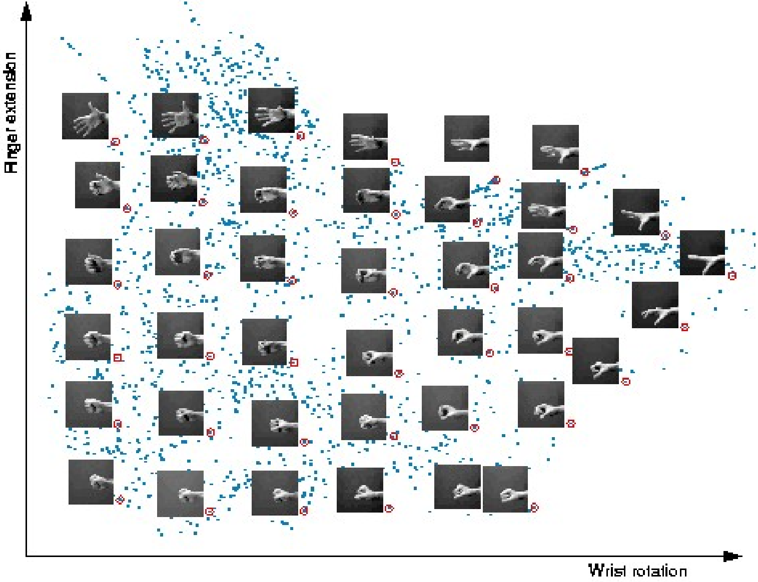
Requisite

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



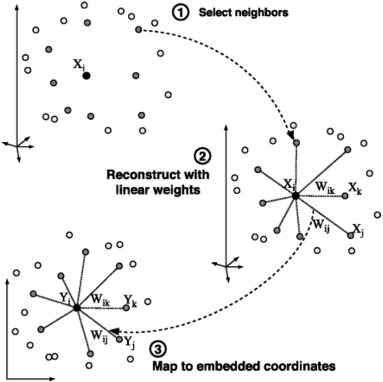
Locally Linear Embedding

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

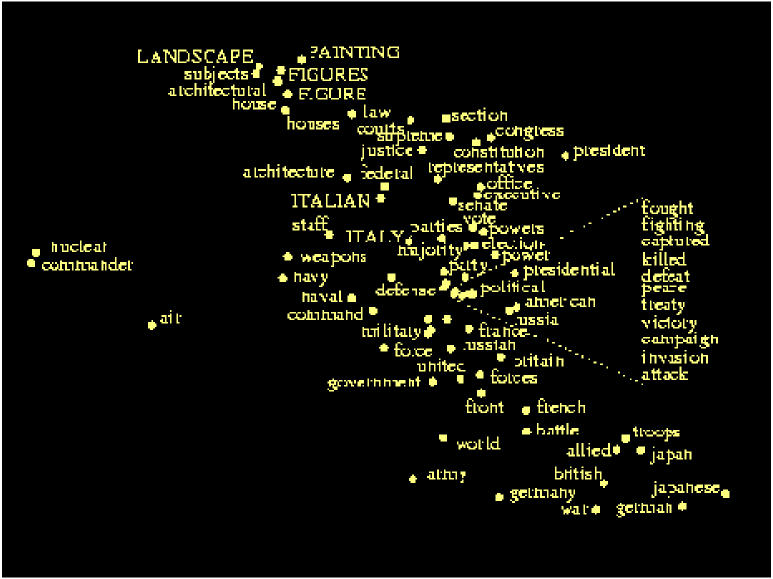
$$\sum_{i=1}^{N} ||z_i - \sum_j w_{i,j} z_j||^2$$

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

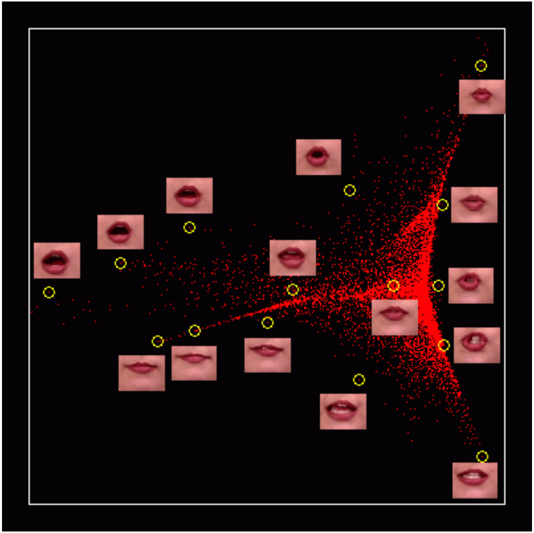
Solutions: vectors z_i are eigenvectors of $(I-W)'(I-W)$

- Keeping the d eigenvectors with lowest eigenvalues > 0

Example, Texts



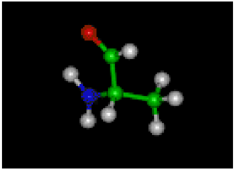
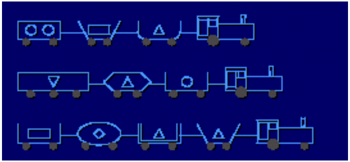
Example, Images



LLE

Propositionalization

Relational domains

Relational learning

PROS — Inductive Logic Programming

- Use domain knowledge

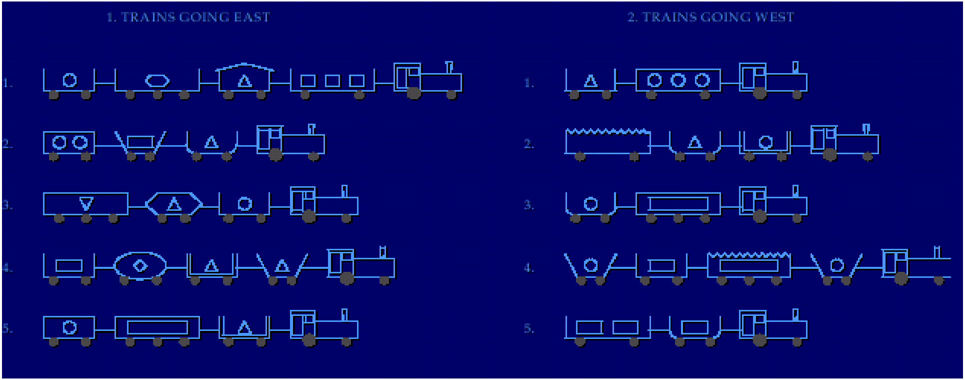
CONS — Data Mining

- 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

$$\begin{array}{ll} West(a) \leftarrow & Engine(a,b), first\_wagon(a,c), roof(c), load(c, square, 3)\ldots \\ West(a') \leftarrow & Engine(a',b'), first\_wagon(a',c'), load(c', circle, 1)\ldots \end{array}$$

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

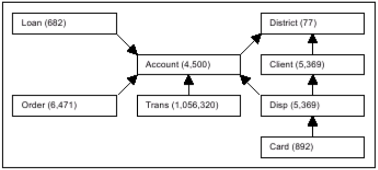
SINUS – RDS

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

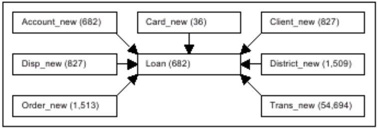
`http://labe.felk.cvut.cz/~zelezny/rsd`

- 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

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

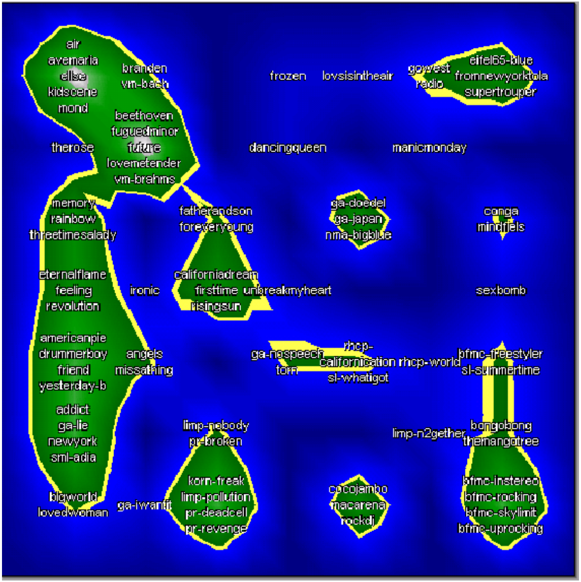
- Exploratory Analysis and Clustering
- Standard approaches, stability, affinity propagation

Part 3: Clustering

- Approaches
  - K-Means
  - EM
  - Selecting the number of clusters
- Clustering the EGEE jobs
  - Dealing with heterogeneous data
  - Assessing the results

Clustering

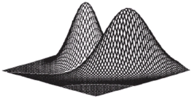
http://www.ofai.at/ elias.pampalk/music/



Clustering Questions

Hard or soft ?

- Hard: find a partition of the data
- Soft: estimate the distribution of the data as a mixture of components.



Parametric *vs* non Parametric ?

- Parametric: number K of clusters is known
- Non-Parametric: find K
  (wrapping a parametric clustering algorithm)

Caveat:

- Complexity
- Outliers
- Validation

Formal Background

Notations

$\mathcal{E}$	$\{\mathbf{x}_1, \ldots \mathbf{x}_N\}$ dataset	
N	number of data points	
K	number of clusters	given or optimized
C_k	k-th cluster	Hard clustering
$\tau(i)$	index of cluster containing $\mathbf{x}_i$	
f_k	k-th model	Soft clustering
$\gamma_k(i)$	$Pr(\mathbf{x}_i \mid f_k)$	

Solution

Hard Clustering	Partition $\Delta = (C_1, \ldots C_k)$
Soft Clustering	$\forall i \; \sum_k \gamma_k(i) = 1$

Formal Background, 2

Quality / Cost function

Measures how well the clusters characterize the data

- (log)likelihood — soft clustering
- dispersion — hard clustering

$$\sum_{k=1}^{K} \frac{1}{|C_k|^2} \sum_{\mathbf{x}_i, \mathbf{x}_j \ in \ C_k} d(\mathbf{x}_i, \mathbf{x}_j)^2$$

Tradeoff

Quality increases with K $\Rightarrow$ Regularization needed

to avoid one cluster per data point

Clustering vs Classification

Marina Meila
http://videolectures.net/

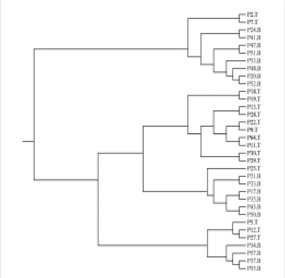
	Classification	Clustering
K	# classes (given)	# clusters (unknown)
Quality	Generalization error	many cost functions
Focus on	Test set	Training set
Goal	Prediction	Interpretation
Analysis	discriminant	exploratory
Field	mature	new

Non-Parametric Clustering

Hierarchical Clustering

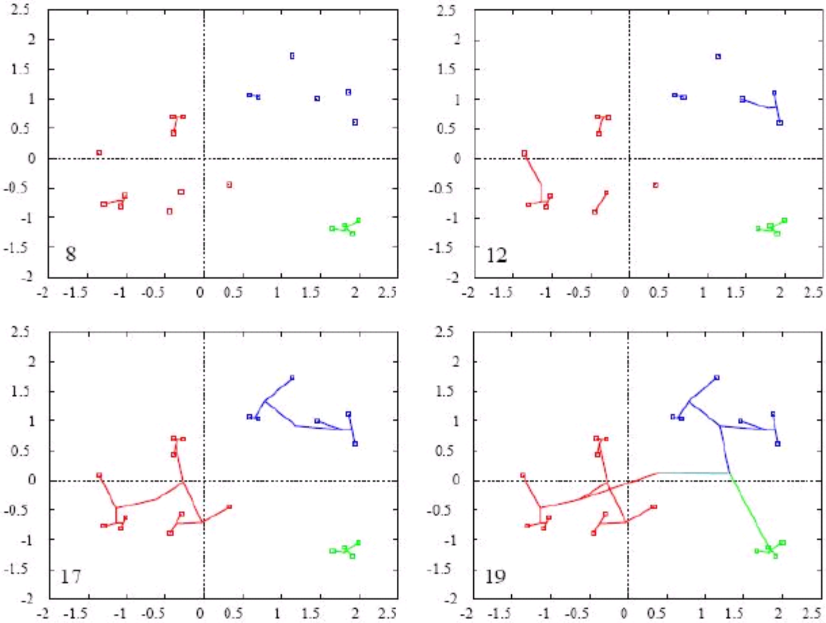
Principle

- agglomerative (join nearest clusters)
- divisive (split most dispersed cluster)

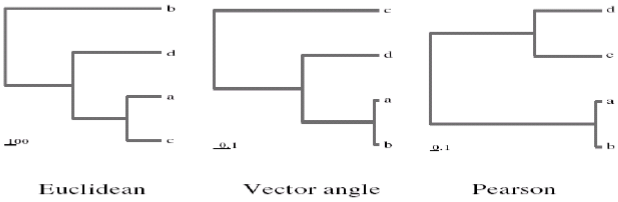


CONS: Complexity $\mathcal{O}(N^3)$

Hierarchical Clustering, example



Influence of distance/similarity



$$d(x,x') = \begin{cases} \sqrt{\sum_i (x_i - x'_i)^2} & \text{Euclidean distance} \\ 1 - \frac{\sum_i x_i x'_i}{||x||.||x'||} & \text{Cosine angle} \\ 1 - \frac{\sum_i (x_i - \bar{x})(x'_i - \bar{x}')}{||x-\bar{x}||.||x'-\bar{x}'||} & \text{Pearson} \end{cases}$$

Parametric Clustering

K is known

Algorithms based on distances

- K-means
- graph / cut

Algorithms based on models

- Mixture of models: EM algorithm

K-Means

Algorithm

1. Init:
   Uniformly draw K points $\mathbf{x}_{i_j}$ in $\mathcal{E}$
   Set $C_j = \{\mathbf{x}_{i_j}\}$
2. Repeat
3. Draw without replacement $\mathbf{x}_i$ from $\mathcal{E}$
4. $\tau(i) = argmin_{k=1\ldots K}\{d(\mathbf{x}_i, C_k)\}$ find best cluster for $\mathbf{x}_i$
5. $C_{\tau(i)} = C_{\tau(i)} \bigcup \mathbf{x}_i$ add $\mathbf{x}_i$ to $C_{\tau(i)}$
6. Until all points have been drawn
7. If partition $C_1 \ldots C_K$ has changed Stabilize
   Define $\mathbf{x}_{i_k} =$ best point in C_k, $C_k = \{x_{i_k}\}$, goto 2.

Algorithm terminates

K-Means, Knobs

Knob 1 : define $d(\mathbf{x}_i, C_k)$	favors
▶ $min\{d(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_j \in C_k\}$	long clusters
* $average\{d(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_j \in C_k\}$	compact clusters
▶ $max\{d(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_j \in C_k\}$	spheric clusters

Knob 2 : define "best" in C_k

- ▶ Medoid — $argmin_i\{\sum_{\mathbf{x}_j \in C_k} d(\mathbf{x}_i, \mathbf{x}_j)\}$
- * Average — $\frac{1}{|C_k|} \sum_{\mathbf{x}_j \in C_k} \mathbf{x}_j$

  (does not belong to $\mathcal{E}$)

No single best choice

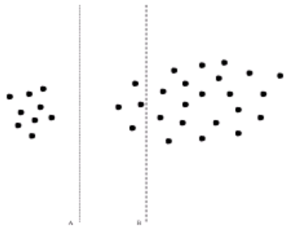


FIG. 1. Optimizing the diameter produces B while A is clearly more desirable.

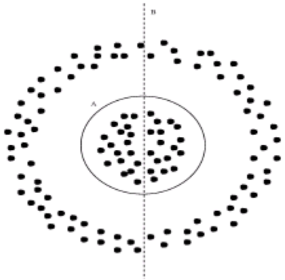


FIG. 2. The inferior clustering B is found by optimizing the 2-median measure.

K-Means, Discussion

PROS

- Complexity $\mathcal{O}(K \times N)$
- Can incorporate prior knowledge initialization

CONS

- Sensitive to initialization
- Sensitive to outliers
- Sensitive to irrelevant attributes

K-Means, Convergence

- For cost function

$$\mathcal{L}(\Delta) = \sum_{k} \sum_{i,j \ / \ \tau(i)=\tau(j)=k} d(\mathbf{x}_i, \mathbf{x}_j)$$

- for $d(\mathbf{x}_i, C_k) =$ average $\{d(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_j \in C_k\}$
- for "best" in C_k = average of $\mathbf{x}_j \in C_k$

K-means converges toward a (local) minimum of $\mathcal{L}$.

K-Means, Practicalities

Initialization

- Uniform sampling
- Average of $\mathcal{E}$ + random perturbations
- Average of $\mathcal{E}$ + orthogonal perturbations
- Extreme points: select $\mathbf{x}_{i_1}$ uniformly in $\mathcal{E}$, then

$$\text{Select } x_{i_j} = argmax\{\sum_{k=1}^{j} d(\mathbf{x}_i, x_{i_k})\}$$

Pre-processing

- Mean-centering the dataset

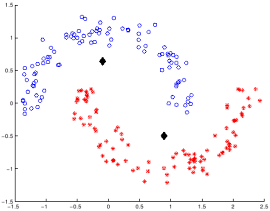
Model-based clustering

Mixture of components

- Density $f = \sum_{k=1}^{K} \pi_k f_k$
- f_k: the k-th component of the mixture
- $\gamma_k(i) = \frac{\pi_k f_k(x)}{f(x)}$
- induces $C_k = \{\mathbf{x}_j \ / \ k = argmax\{\gamma_k(j)\}\}$

Nature of components: prior knowledge

- Most often Gaussian: $f_k = (\mu_k, \Sigma_k)$
- Beware: clusters are not always Gaussian...



Model-based clustering, 2

Search space

- Solution : $(\pi_k, \mu_k, \Sigma_k)_{k=1}^K = \theta$

Criterion: log-likelihood of dataset

$$\ell(\theta) = \log(Pr(\mathcal{E})) = \sum_{i=1}^N \log Pr(\mathbf{x}_i) \propto \sum_{i=1}^N \sum_{k=1}^K \log(\pi_k f_k(\mathbf{x}_i))$$

to be maximized.

Model-based clustering with EM

Formalization

- Define $z_{i,k} = 1$ iff $\mathbf{x}_i$ belongs to C_k.
- $E[z_{i,k}] = \gamma_k(i)$ prob. $\mathbf{x}_i$ generated by $\pi_k f_k$
- Expectation of log likelihood

$$
\begin{aligned}
E[\ell(\theta)] &\propto \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_i(k) \log(\pi_k f_k(\mathbf{x}_i)) \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_i(k) \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_i(k) \log f_k(\mathbf{x}_i)
\end{aligned}
$$

EM optimization

E step Given θ, compute

$$\gamma_k(i) = \frac{\pi_k f_k(\mathbf{x}_i)}{f(x)}$$

M step Given $\gamma_k(i)$, compute

$$\theta^* = (\pi_k, \mu_k, \Sigma_k)^* = argmin E[\ell(\theta)]$$

Maximization step

π_k: Fraction of points in C_k

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \gamma_k(i)$$

μ_k: Mean of C_k

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_k(i) \mathbf{x}_i}{\sum_{i=1}^{N} \gamma_k(i)}$$

Σ_k: Covariance

$$\Sigma_k = \frac{\sum_{i=1}^{N} \gamma_k(i)(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)'}{\sum_{i=1}^{N} \gamma_k(i)}$$

Choosing the number of clusters

K-means constructs a partition whatever the K value is.

Selection of K

- Bayesian approaches
  Tradeoff between accuracy / richness of the model
- Stability
  Varying the data should not change the result
- Gap statistics
  Compare with null hypothesis: all data in same cluster.

Bayesian approaches

Bayesian Information Criterion

$$BIC(\theta) = \ell(\theta) - \frac{\#\theta}{2} \log N$$

Select $K = \text{argmax } BIC(\theta)$
where $\#\theta$ = number of free parameters in θ:

- if all components have same scalar variance σ

$$\#\theta = K - 1 + 1 + Kd$$

- if each component has a scalar variance σ_k

$$\#\theta = K - 1 + K(d + 1)$$

- if each component has a full covariance matrix Σ_k

$$\#\theta = K - 1 + K(d + d(d - 1)/2)$$

Gap statistics

Principle: hypothesis testing

1. Consider hypothesis H_0: there is no cluster in the data. $\mathcal{E}$ is generated from a no-cluster distribution π.
2. Estimate the distribution $f_{0,K}$ of $\mathcal{L}(C_1, \dots C_K)$ for data generated after π.   Analytically if π is simple
   Use Monte-Carlo methods otherwise
3. Reject H_0 with confidence α if the probability of generating the true value $\mathcal{L}(C_1, \dots C_K)$ under $f_{0,K}$ is less than α.

Beware: the test is done for all K values...

Gap statistics, 2

Algorithm

Assume $\mathcal{E}$ extracted from a no-cluster distribution, e.g. a single Gaussian.

1. Sample $\mathcal{E}$ according to this distribution
2. Apply K-means on this sample
3. Measure the associated loss function

Repeat : compute the average $\bar{\mathcal{L}}_0(K)$ and variance $\sigma_0(K)$
Define the gap:

$$Gap(K) = \bar{\mathcal{L}}_0(K) - \mathcal{L}(C_1, \ldots C_K)$$

Rule Select min K s.t.

$$Gap(K) \geq Gap(K+1) - \sigma_0(K+1)$$

What is nice: also tells if there are no clusters in the data...

Stability

Principle

- Consider $\mathcal{E}'$ perturbed from $\mathcal{E}$
- Construct $C'_1, \ldots C'_K$ from $\mathcal{E}'$
- Evaluate the "distance" between $(C_1, \ldots C_K)$ and $(C'_1, \ldots C'_K)$
- If small distance (stability), K is OK

Distortion $D(\Delta)$

$$\begin{array}{llll} \text{Define} & S & S_{ij} = & < \mathbf{x}_i, \mathbf{x}_j > \\ & & (\lambda_i, v_i) & \text{i-th (eigenvalue, eigenvector) of } S \\ & X & X_{i,j} = & 1 \text{ iff } \mathbf{x}_i \in C_j \end{array}$$

$$D(\Delta) = \sum_i ||\mathbf{x}_i - \mu_{\tau(i)}||^2 = tr(S) - tr(X'SX)$$

Minimal distortion $D^* = tr(S) - \sum_{k=1}^{K-1} \lambda_k$

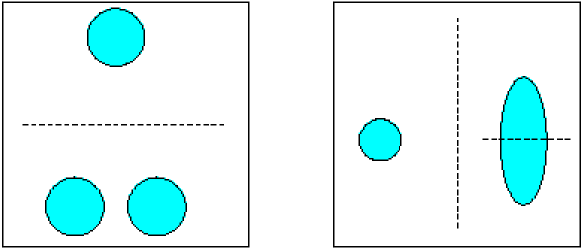
Stability, 2

Results

- Δ has low distortion $\Rightarrow$ $(\mu_1, \ldots \mu_K)$ close to space $(v_1, \ldots v_K)$.
- Δ_1, and Δ_2 have low distortion $\Rightarrow$ "close"
- (and close to "optimal" clustering)

Meila ICML 06

Counter-example



Overview

Autonomic Computing

- A booming field of applications
- Machine Learning and Data Mining for Systems

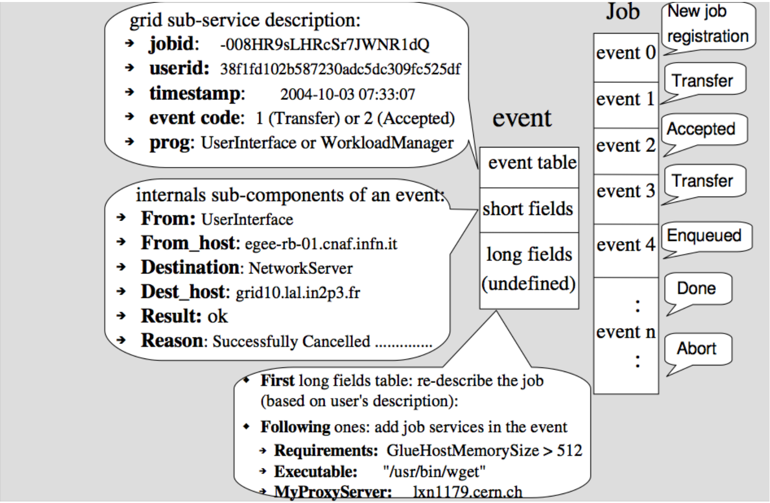
Autonomic Grid

- 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
- Clustering the jobs

Job representation



Xiangliang Zhang et al., ICDM wshop on Data streams, 2007

Job representation

Challenges

- Sparse representation, e.g. “user id”
- No natural distance

Prior knowledge

- Coarse job classification: succeeds (SUC) or fails (FAIL)
- Many failure types: Not Available Resources (NAR); User Aborted (ABU); Generic and non-Generic Error (GNG).
- Jobs are heterogeneous
  - Due to users (advanced or naive)
  - Due to virtual organizations (jobs in physics $\neq$ jobs in biology)
  - Due to time: grid load depends on the community activity

Feature extraction

Slicing data

to get rid of heterogeneity

- Split jobs per user: $U_i = \{$ jobs of i-th user $\}$
- Split jobs per week: $W_j = \{$ jobs launched in j-th week $\}$

Building features

- Each data slice: a supervised learning problem (discriminating *SUCC* from *FAIL*)

$$h : \mathcal{X} \mapsto \mathbb{R}$$

- Supervised Learning Algorithms:
  - Support Vector Machine — SVMLight
  - Optimization of AUC — ROGER

Feature Extraction, 2

New features

Define

$h_{u,i}$ hypothesis learned from data slice U_i

$U : \mathcal{X} \mapsto \mathbb{R}^{\#u}$

$$U(\mathbf{x}) = (h_{u,1}(\mathbf{x}), \ldots h_{u,\#u}(\mathbf{x}))$$

Symmetrically $h_{w,i}$ hypothesis learned from data slice W_i

$W : \mathcal{X} \mapsto \mathbb{R}^{\#w}$

$$W(\mathbf{x}) = (h_{w,1}(\mathbf{x}), \ldots h_{w,\#w}(\mathbf{x}))$$

Change of representation

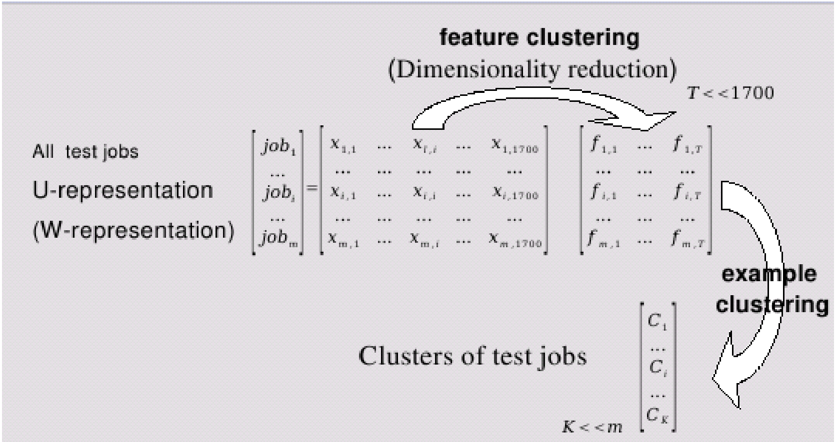
$$\begin{array}{ccl} \mathcal{E} & \rightarrow & \mathcal{E}_U = \{(U(\mathbf{x}_i), y_i), i = 1 \ldots N\} \\ & \rightarrow & \mathcal{E}_W = \{(W(\mathbf{x}_i), y_i), i = 1 \ldots N\} \end{array}$$

Discussion

- Natural distance on $\mathbb{R}^d$
- But new attributes $h_{u,i}$ likely to be redundant

Feature Extraction: Double clustering

Slonim & Tishby, 2000



Experimental setting

The datasets

- Training set $\mathcal{E}$: 222,500 jobs   36% SUCC, 74% FAIL
- Test set $\mathcal{T}$: 21,512 jobs

Hypothesis construction

- SVM: one hypothesis per slice:
  $$U : \mathcal{X} \mapsto \mathbb{R}^{34}$$
  $$W : \mathcal{X} \mapsto \mathbb{R}^{45}$$
- ROGER: 50 hypotheses per slice
  $$U : \mathcal{X} \mapsto \mathbb{R}^{1700}$$
  $$W : \mathcal{X} \mapsto \mathbb{R}^{2250}$$

Clustering

Foreach $K = 5 \ldots 30$, Apply K-means to $\mathcal{T}$

- Considering new representations U and W
- Learned after SVM and Roger.

Goal of Experiments

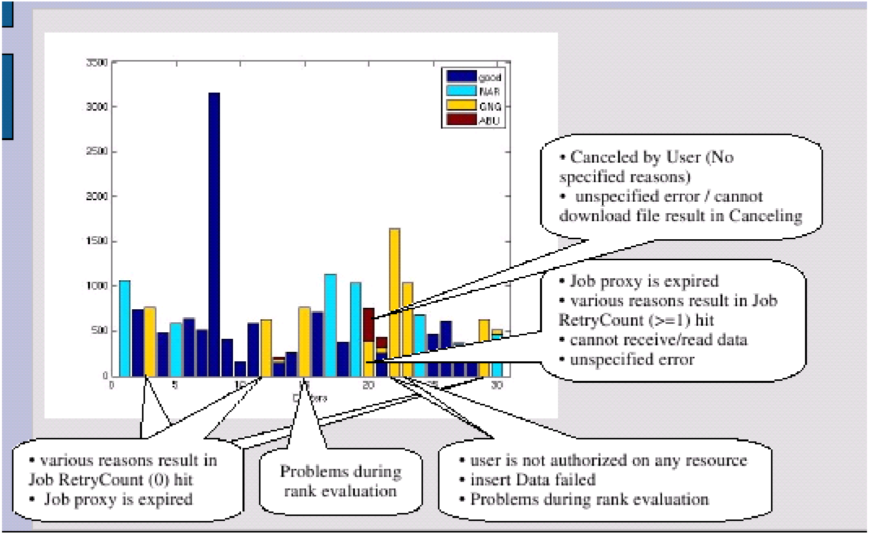
Interpretation

Examine the clusters

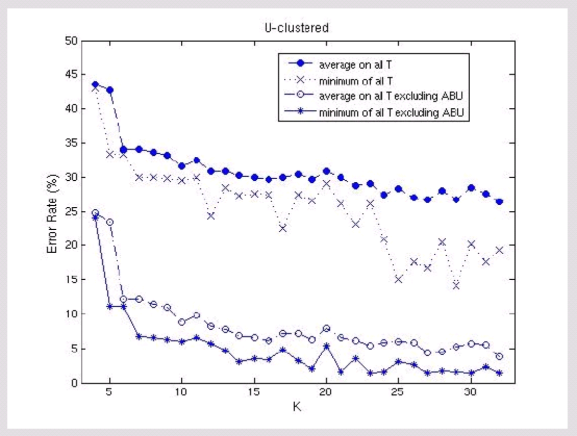
Stability

- Compare Δ_K and $\Delta_{K'}$
- Compare $\Delta_{K,U}$ and $\Delta_{K,W}$

Interpretation



Interpretation, 2



Interpretation, 3

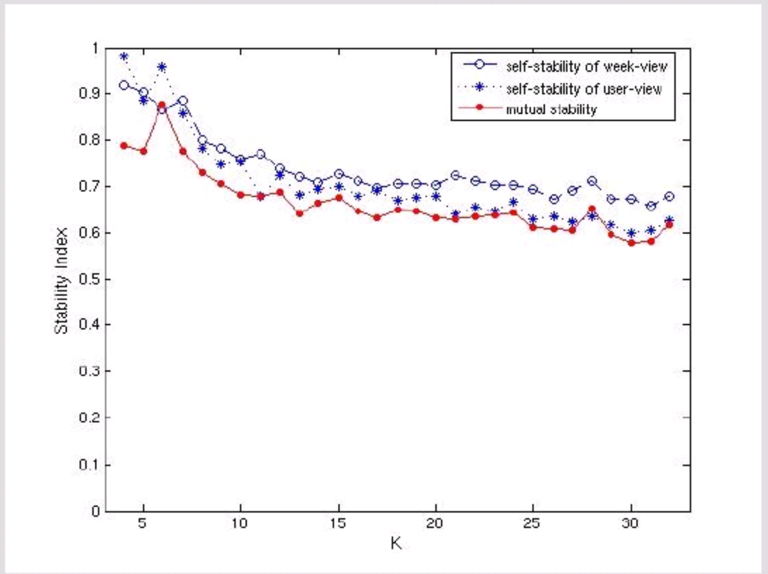
Pure clusters

- Most clusters are pure wrt sub-classes NAR, GNG which were unknown from the algorithm
- Finer-grained classes are discovered: Problem during rank evaluation; job proxy expired; insert Data failed
- ABU class (1.2%) is not properly identified: many reasons why job might be *Aborted by User*

Usage

Use prediction for user-friendly service
Anticipate job failures

Stability



self-stability of week-view
self-stability of user-view
mutual stability
Stability Index
K
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30

Stability, 2

- Stability wrt initialization, for both W and U representations
- Stability of clusters based on W and U-based representations
- Decreases gracefully with K
  (optimal value $= 1$)

Grid Modelling, wrap-up

Conclusion

- Importance of representation as usual
- Clustering: stable wrt K and representation change
  re-discovers types of failures
  discovers finer-grained failures

Future work

- Cluster users (= sets of jobs)
- Cluster weeks (= sets of jobs)
- Find scenarios
  naive users gaining expertise;
  grid load & temporal regularities
- Identify communities of users.
- Use scenarios to test/optimize grid services (e.g. scheduler)

Autonomic Computing, wrap-up

Huge needs

- Modelling systems

  Black box to calibrate, train, optimize services
- Understanding systems — Hints to repair, re-design systems

Dealing with Complex Systems

- Findings often challenge conventional wisdow
- Theoretical *vs* Empirical models
- Complex systems are counter-intuitive — sometimes

Autonomic Computing, wrap-up, 2

Good practice

- No Magic !
  I don't see anything, I'll use ML or DM
- Use all of your prior knowledge
  If you can measure/model it, don't guess it!
- Have conjectures
- Test them!

Beware: False Discovery Rate

Thanks to

- Cécile Germain-Renaud
- Xiangliang Zhang
- Cal Loomis
- Nicolas Baskiotis
- Moises Goldszmidt
- The PASCAL Network of Excellence

http://www.pascal-network.org